PSYCHOMETRIKA
https://doi.org/10.1007/s11336-022-09874-6

REGULARIZED VARIATIONAL ESTIMATION FOR EXPLORATORY ITEM FACTOR ANALYSIS

APRIL E. CHO

UNIVERSITY OF MICHIGAN

JIAYING XIAO

UNIVERSITY OF WASHINGTON

CHUN WANG

UNIVERSITY OF WASHINGTON

GONGJUN XU

UNIVERSITY OF MICHIGAN

Item factor analysis (IFA), also known as Multidimensional Item Response Theory (MIRT), is a general framework for specifying the functional relationship between respondents' multiple latent traits and their responses to assessment items. The key element in MIRT is the relationship between the items and the latent traits, so-called item factor loading structure. The correct specification of this loading structure is crucial for accurate calibration of item parameters and recovery of individual latent traits. This paper proposes a regularized Gaussian Variational Expectation Maximization (GVEM) algorithm to efficiently infer item factor loading structure directly from data. The main idea is to impose an adaptive L_1-type penalty to the variational lower bound of the likelihood to shrink certain loadings to 0. This new algorithm takes advantage of the computational efficiency of GVEM algorithm and is suitable for high-dimensional MIRT applications. Simulation studies show that the proposed method accurately recovers the loading structure and is computationally efficient. The new method is also illustrated using the National Education Longitudinal Study of 1988 (NELS:88) mathematics and science assessment data.

Key words: latent variable selection, multidimensional item response theory, variational inference, expectation-maximization, lasso, adaptive lasso.

1. Introduction

Full Information Item factor analysis (IFA), known as factor analysis of ordered categorical (such as binary) item-level data, has been a useful tool to explore the latent structure underlying educational and psychological tests (Bock, Gibbons, & Muraki, 1988). IFA provides a wealth of information regarding the characteristics of the items and tests, which are important to ensure reliability and validity of a measure. As IFA deals with item-level responses, it is also considered as multidimensional item response theory (MIRT) (Embretson & Reise, 2000; Reckase, 2009)

Cho and Xiao contributed equally to this work. This research is partially supported by Institute of Education Sciences grant R305D200015 and National Science Foundation grants SES-1846747, SES-1659328, and DMS-1712717.

Correspondence should be made to Chun Wang, 312E Miller Hall, 2012 Skagit Ln, Seattle, WA98105, USA. Email: wang4066@uw.edu

Correspondence should be made to Gongjun Xu, 456 West Hall, 1085 South University, Ann Arbor, MI48109, USA. Email: gongjun@umich.edu

Published online: 13 July 2022

The widely used multidimensional 2-parameter logistic (M2PL) model assumes item response function of the ith individual to the jth item as

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i) = \frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}, \tag{1}$$

where there are N subjects who respond to J items independently with binary response variables Y_{ij}, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$. $\boldsymbol{\alpha}_j$ denotes a K-dimensional vector of item discrimination parameters for the jth item and b_j denotes the corresponding item difficulty parameter. $\boldsymbol{\theta}_i$ denotes the K-dimensional vector of latent ability for student i. $\boldsymbol{\alpha}_j$ may contain structural 0's implying that item j does not measure (hence not load on) certain factors. When both $\boldsymbol{\alpha}_j$ and $\boldsymbol{\theta}_i$ are unidimensional, the 2PL model and one-factor categorical factor analysis model are mathematically equivalent (Takane & De Leeuw, 1987; Wirth & Edwards, 2007). Another popular MIRT model that is often suitable for multiple-choice binary response items is the multidimensional 3-parameter logistic (M3PL) model. It includes an additional parameter c_j to quantify guessing probability of the jth item. Hence, the item response function is expressed as

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i) = c_j + (1 - c_j)\frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}. \tag{2}$$

Although the inclusion of the guessing parameter makes the model more flexible, it no longer belongs to the exponential family and its estimation becomes much more challenging (Thissen & Wainer, 1982; Yen, 1987).

In an exploratory IFA, the item factor loading structure that is reflected by the systematic 0's in $\boldsymbol{\alpha}_j$ is unknown. Identifying the loading structure, which is equivalent to the sparsity structure of $\boldsymbol{\alpha}_j$, is crucial not only for accurate calibration of item parameters and recovery of individual latent traits, but also for understanding the construct validity of a measure. Traditional approaches for identifying item factor loading structure proceed in two steps: (1) allowing all item factor loadings to be freely estimated, subject to identifiability constraints; and (2) conducting a post-hoc rotation (Browne, 2001a). Most software packages use varimax (Kaiser, 1958) for orthogonal rotation or promax (Hendrickson & White,1966) for oblique rotation by default. Other popular methods include, for instance, the CF-Quartimax rotation (Browne, 2001a). While these rotation methods intend to produce a near-simple structure, an arbitrary cutoff for the rotated factor loadings is often needed. Rotation methods that encourage sparse solutions have also been developed in Jennrich (2004; 2006) using the component loss functions for orthogonal and oblique rotations.

To avoid setting subjective cutoffs, Sun, Chen, Liu, Ying, and Xin (2016) recently proposed to formulate the problem of estimating the loading structure in MIRT as a latent variable selection problem. Specifically, for each item, a set of latent traits influencing the distribution of the responses are selected by the L_1-regularized regression. The L_1-regularized regression, also known as the constrained least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), has received much attention for solving variable selection problems for both linear and generalized linear models (Friedman, Hastie, & Tibshirani, 2010). The principle idea is to penalize the factor loadings toward zero if the corresponding latent traits are not associated with an item. This leads to correctly estimating an optimal nonzero factor loading structure, instead of setting subjective cutoffs. This approach also has the advantage over the information criterion-based model selection methods in terms of the computational cost because it simultaneously estimates both loading structure and model parameters. Despite its appeal, the computation is still quite challenging in MIRT model due to its intractable marginal likelihood function that involves

high-dimensional integration. For parameter estimation, Sun et al. (2016) used direct numerical approximation of the likelihood in the iterative expectation-maximization (EM) procedure, which can be computationally inefficient especially in higher dimensions. Specifically, they showed that the computation time for the latent variable selection with dimension $K = 3$ is about 30 minutes for the first penalization tuning parameter λ and additional 10 minutes for the subsequent λs. Considering that multiple λs have to be used for the latent variable selection via regularization, it can take a few hours to estimate a test structure for a single dataset with high dimensions.

Indeed, developing efficient estimation algorithms for MIRT parameter estimation has always been a productive research topic. A number of methods have been proposed to deal with the computational challenge (Rabe-Hesketh, Skrondal, & Pickles, 2005; von Davier & Sinharay, 2010). The first one is the adaptive Gaussian quadrature method. Although the number of quadrature points per dimension could be small, the total number of quadrature points still increases exponentially with the number of dimensions. Moreover, an extra step is needed to compute the posterior mode and variance of latent factors in each iteration, which adds additional computation costs (Pinheiro & Bates, 1995). The second one is the Monte Carlo techniques. This family of methods include, for instance, the Monte Carlo EM algorithm (McCulloch, 1997; C. Wang & Xu, 2015), stochastic EM algorithm (von Davier & Sinharay, 2010; S. Zhang, Chen, & Liu, 2020), or Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010b; 2010a). These methods circumvent intractable integrations by sampling from the posterior distributions; however, they may be still computationally intensive for complicated high-dimensional models, as a large Monte Carlo sample size is typically needed and the posterior distributions usually do not have a closed form. Fully Bayesian estimation methods, such as Markov chain Monte Carlo (MCMC) (Albert, 1992; Patz & Junker, 1999), is equally computationally intensive, even though it is preferable with smaller sample sizes. It usually needs a long chain to converge for complex models. In addition, Chen, Li, and Zhang (2019) and H. Zhang, Chen, and Li (2020) studied the joint maximum likelihood estimation by treating the latent abilities as fixed effect parameters instead of random variables; though computationally efficient, such joint likelihood based estimation approaches may be less statistically efficient than the marginal likelihood estimation (e.g., Cho, Wang, Zhang, and Xu (2021)).

Most recently, a variational approximation approach to the marginal likelihood was proposed, namely the Gaussian Variational EM (GVEM) algorithm (Cho et al., 2021). GVEM adopts a variational lower bound of the intractable likelihood within the EM framework. The carefully constructed variational lower bound allows one to derive closed-form updates for all model parameters in the iterative EM steps, making the algorithm computationally efficient. Cho et al. (2021) also proposed a stochastic version of GVEM to further improve its computational efficiency when both the number of subjects, N, and the number of test items, J, are large. The idea is to stochastically optimize the variational approximation in the E step, i.e., subsample data to form noisy estimate of the variational lower bound and iteratively update the estimate with a decreasing step size (Hoffman, Blei, Wang, & Paisley, 2013). The combined advantage of having simple closed-form updates and stochastic optimization makes the GVEM algorithm appealing to high-dimensional MIRT models. Additionally, it was shown that GVEM works well in complex M3PL models compared to the existing methods.

In this paper, we propose to extend the GVEM algorithm by adding a regularization penalty to simultaneously estimate item factor loading structure and model parameters. Our study differs from Sun et al. (2016)'s in the following aspects: (1) we use GVEM as the estimation algorithm instead of the quadrature-based EM algorithm, hence the new method is more suitable to tackle high-dimensional challenge; (2) we consider both Lasso and adaptive Lasso (Zou, 2006), the latter of which produces more accurate loading structure recovery; (3) we apply the new method to both the M2PL and M3PL models.

The rest of the paper is organized as follows. Section 2 briefly introduces the GVEM algorithm for the MIRT models. Section 3 presents the general regularized variational algorithm. Sections 4 and 5 illustrate the performance of the proposed methods with simulation studies and real data analysis, respectively. Section 6 discusses potential future studies, and the supplementary material includes the derivations of the estimation procedures and additional data analysis results.

2. Variational Estimation for MIRT

In this section, we will briefly present the key idea of variational approximation discussed in Cho et al. (2021). The exposition will be based on the M3PL model, but it can be easily simplified to the M2PL model. For conciseness, let us denote the model parameters for the MIRT models by $\boldsymbol{A} = \{\boldsymbol{\alpha}_j, j = 1, \ldots, J\}$, $\boldsymbol{B} = \{b_j, j = 1, \ldots, J\}$, and $\boldsymbol{C} = \{c_j, j = 1, \ldots, J\}$. Also, denote the responses $\boldsymbol{Y} = \{Y_i, i = 1, \ldots, N\}$ where $Y_i = \{Y_{ij}, j = 1, \ldots, J\}$ is the ith subject's response vector. Due to the typical local independence assumption in IRT, the log-marginal likelihood of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ in M3PL model given the responses $\mathbf{Y}$ is

$$l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}; \mathbf{Y}) = \sum_{i=1}^{N} \log P(Y_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = \sum_{i=1}^{N} \log \int \prod_{j=1}^{J} P(Y_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \phi(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \quad (3)$$

where N is the total number of respondents and J is the total number of items in the test. Similarly this holds for the M2PL model with model parameters $\boldsymbol{A}$ and $\boldsymbol{B}$. Here, ϕ denotes the K-dimensional Gaussian distribution of $\boldsymbol{\theta}$ with mean $\mathbf{0}$ and covariance $\Sigma_{\boldsymbol{\theta}}$. The maximum likelihood estimators of the model parameters are then obtained from maximizing the marginal likelihood function, which is often intractable under MIRT.

From here onwards, M_p is used to denote all model parameters for simplicity. Following Cho et al. (2021), the variational approximation of (3) can be derived as follows. First, for any arbitrary probability density function $q_i(\cdot)$, we can rewrite the log-marginal likelihood in Eq. 3 as

$$\begin{aligned} l(M_p; \mathbf{Y}) &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log P(Y_i \mid M_p) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log \frac{P(Y_i, \boldsymbol{\theta}_i \mid M_p)}{P(\boldsymbol{\theta}_i \mid Y_i, M_p)} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log \frac{P(Y_i, \boldsymbol{\theta}_i \mid M_p)}{q_i(\boldsymbol{\theta}_i)} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + KL\{q_i(\boldsymbol{\theta}_i) \| P(\boldsymbol{\theta}_i \mid Y_i, M_p)\}, \end{aligned}$$

where $KL\{q_i(\boldsymbol{\theta}_i) \| P(\boldsymbol{\theta}_i \mid Y_i, M_p)\} = \int_{\boldsymbol{\theta}_i} \log \frac{q_i(\boldsymbol{\theta}_i)}{P(\boldsymbol{\theta}_i \mid Y_i, M_p)} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$ denotes the Kullback–Leibler (KL) distance between the distributions $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$. Then, since $KL\{q_i(\boldsymbol{\theta}_i) \| P(\boldsymbol{\theta}_i \mid Y_i, M_p)\} \geq 0$, we have a lower bound of the marginal likelihood as

$$l(M_p; \mathbf{Y}) \geq \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log P(Y_i, \boldsymbol{\theta}_i \mid M_p) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i - \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log q_i(\boldsymbol{\theta}_i) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \quad (4)$$

Note that the equality in (4) holds if and only if $q_i(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ for $i = 1, \ldots, N$. Thus, to use the lower bound in (4) to approximate the marginal likelihood $l(M_p; \mathbf{Y})$, the posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ gives the best choice of the variational distribution function $q_i(\boldsymbol{\theta}_i)$. However, such a choice of $q_i(\boldsymbol{\theta}_i)$ is not practically applicable as the posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ is unknown. Alternatively, we could choose $q_i(\boldsymbol{\theta}_i)$ as a tractable approximation of $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$. One example is the EM algorithm, which can be viewed as choosing $q_i(\boldsymbol{\theta}_i)$ as the estimated posterior $P(\boldsymbol{\theta}_i \mid Y_i, \hat{M}_p)$ with $\hat{M}_p$ from a previous EM step estimate. However, in the MIRT model, it is known that the expectation in E-step with respect to the posterior distribution of $\boldsymbol{\theta}_i$, i.e., the first term in (4) with $q_i(\boldsymbol{\theta}_i)$ being the estimated posterior $P(\boldsymbol{\theta}_i \mid Y_i, \hat{M}_p)$, does not have an explicit form and often is challenging to compute.

Different from the EM algorithm, the variational inference method uses alternative choices of the $q_i(\boldsymbol{\theta}_i)$'s to have a computationally more efficient estimation of the lower bound in (4). Since the posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ for the MIRT model can be well approximated by a Gaussian distribution as the number of items J increases, following (Cho et al. 2021), we choose $q_i(\boldsymbol{\theta}_i)$ from a family of Gaussian distributions and estimate the model parameters by the GVEM algorithm. In particular, in the E-step, q_i is estimated within the Gaussian family to minimize the KL distance between $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$, and we then evaluate the expectation of the likelihood lower bound with respect to the estimated $q_i(\boldsymbol{\theta}_i)$. In the M-step, the expectation is maximized to update all model parameters. Carefully chosen q_i yield closed-form updates for all model parameters (Cho et al., 2021), making the algorithm computationally efficient.

3. Regularized Estimation of Loading Structure

In this paper, our main interest is to estimate a sparse loading structure, denoted as $Q_A = (q_{jk})$ where $q_{jk} = I(\boldsymbol{\alpha}_{jk} \neq 0)$. Similar to Sun et al. (2016), we cast the problem of sparsity estimation as a latent variable selection problem and solve it using the regularized regression via L_1–type penalization. One main contribution is to apply variational approach to avoid directly calculating intractable marginal likelihood while solving the regularization problem.

Although Lasso regularization is a popular technique for simultaneous model estimation and efficient variable selection, there has been some arguments against the Lasso oracle statement. For instance, Zou (2006) argued that there exist nontrivial conditions for the Lasso variable selection to be consistent and thus Lasso rarely enjoys oracle properties. Although the computational efficiency of Lasso is appealing for the estimation problems in high-dimensional MIRT models, the bias of the Lasso may prevent consistent variable selection and model estimation. On the other hand, adaptive Lasso is shown to enjoy oracle properties if the regularization parameters are chosen to be data-dependent (Zou, 2006). Since it is a convex optimization problem, its global optimizer can be efficiently solved. Additionally, adaptive Lasso is a simple extension of Lasso, which makes it easy to implement with the existing algorithm for the Lasso and is computationally efficient as well. Hence, adaptive Lasso is a good candidate as a penalization method for identifying item factor loading structure in MIRT. Specifically for parameter estimation, we solve the following optimization problem;

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\mathbf{C}}_\lambda) = \text{argmax}_{\mathbf{A},\mathbf{B},\mathbf{C}} l(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{Y}) - P_\lambda(\mathbf{A}) \tag{5}$$

where

$$P_\lambda(\mathbf{A}) = \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{w}_{jk} |\alpha_{jk}|$$

with $\hat{w}_{jk} = 1/|\hat{\alpha}_{jk}^{(0)}|^{\gamma}$, $\hat{\alpha}_{jk}^{(0)}$ an initial estimator of α_{jk} without the regularization penalty, and $\gamma > 0$ and $\lambda > 0$ the tuning parameters. In the adaptive Lasso penalization, we use adaptive penalization weights for each parameter α_{jk}, instead of a constant penalization parameter λ as in Lasso. The penalization weight for α_{jk} is $\lambda\hat{w}_{jk} = \lambda/|\hat{\alpha}_{jk}^{(0)}|^{\gamma}$. Thus, $\hat{\alpha}_{jk}^{(0)} < 1$ will get penalized more than the bigger values such as $\hat{\alpha}_{jk}^{(0)} > 1$. The weight is chosen to be dependent on data to satisfy the regulatory conditions discussed in Zou (2006). Particularly, Zou (2006) recommended three values, 0.5, 1, and 2, for the γ parameter, and the selection of the λ parameter will be discussed in Sect. 3.2.

To ensure identifiability, we impose certain constraints on the a $K \times K$ sub-matrix of Q_A. For the remaining part of the A matrix, we do not assume any pre-specified zero structure but instead, the appropriate penalization is imposed to shrink $\boldsymbol{\alpha}_{jk}$'s to recover the true zero structure, Q_A^*. Below are two different constraints on the A matrix. Note that the second constraint is more flexible; hence, it is more challenging estimation wise. Except for adding constraints on Q_A, we also fix the diagonals of $\Sigma_{\boldsymbol{\theta}}$ at 1. Similar to Sun et al. (2016), we will compare the performance of these two constraint settings in the simulation study.

Constraint 1 To ensure identifiability, we designate one item for each latent factor and this item is associated with only that factor. That is, we set a $K \times K$ sub-matrix of Q_A to be an identity matrix, I_K. Together with the constraints on the variance of $\Sigma_{\boldsymbol{\theta}}$, we have K^2 constraints in total.

Constraint 2 Instead of setting all off-diagonals of a $K \times K$ sub-matrix of Q_A to be zero, we keep the sub-matrix of Q_A to be a triangular matrix with the diagonal being ones. That is, there are test items associated with each factor for sure and they may be associated with other factors as well. Nonzero entries except for the diagonal entries of Q_A are penalized during the estimation procedure. Although this constraint is much weaker than the Constraint 1, it still ensures empirical identifiability when proper regularized likelihood such as (5) is used for the model estimation (Sun et al., 2016).

3.1. Additional Penalty for M3PL

The parameter estimation for M3PL in practice often gets more challenging due to the inclusion of guessing parameters. To tackle this challenge and improve the accuracy of the parameter estimation in M3PL, we propose to impose additional constraints on the model parameters, $\mathbf{B} = \{b_j;\, j = 1, \ldots, J\}$ and $\mathbf{C} = \{c_j;\, j = 1, \ldots, J\}$ in addition to the parameter matrix $\mathbf{A}$. Specifically for parameter estimation, we solve the following optimization problem where $P(\cdot)$ denotes a penalty function on model parameters:

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\mathbf{C}}_\lambda) = \text{argmax}_{\mathbf{A},\mathbf{B},\mathbf{C}} \quad l(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{Y}) - P_\lambda(\mathbf{A}) + P(\mathbf{B}) + P(\mathbf{C}) \tag{6}$$

where $P(\mathbf{B}) = \sum_{j=1}^{J} \log N(b_j|\mu_b, \sigma_b^2)$, and $P(\mathbf{C}) = \sum_{j=1}^{J} \log Beta(c_j|\alpha_c, \beta_c)$ for some distribution parameters μ_b, σ_b^2, α_c, and β_c. These penalty functions are chosen to satisfy the ranges of values on which the parameters are defined. For instance, since the guessing parameters $\mathbf{C}$ naturally satisfy the constraint $\{0 < c_j < 1;\, j = 1, \ldots, J\}$, we can assume a "prior" distribution of $c_j \sim Beta(\alpha_c, \beta_c)$. Similarly, we can assume a "prior" distribution of $b_j \sim N(\mu_b, \sigma_b^2)$. The penalty on b_j and c_j are essentially a L_2-type and Laplace penalization, respectively. By imposing these additional penalties on model parameters $\mathbf{B}$ and $\mathbf{C}$, the parameter estimation becomes more stable and robust.

The approach of imposing additional penalty on model parameters $\mathbf{B}$ and $\mathbf{C}$ with the chosen distributions is similar to the Bayes modal estimation presented by Tierney and Kadane (1986). That is, an augmented optimization objective is employed that includes the likelihood and some prior beliefs on the item parameters. These priors can be used to prevent deviant parameter

estimates and help the algorithm to produce more accurate estimation in complex M3PL models. Essentially, Bayes modal estimation can be seen as a regularization on maximum likelihood estimation where maximum likelihood estimation is a special case of Bayes model estimation that assumes uniform prior distributions.

The amount of penalization can be flexibly controlled using the distribution parameters. For instance, one can use non-informative priors on $\mathbf{C}$ such as $Beta(1, 1)$, which is equivalent to flat uniform distribution on $[0, 1]$. Additionally, one can similarly choose non-informative normal prior with high variance σ_b for $\mathbf{B}$. This suggests that although additional penalization functions are added, the algorithm also allows the flexible estimation with essentially no penalty with the choice of non-informative distributions. The advantage of this is that practitioners can adjust the amount of prior knowledge they would like to impose on the model. The less prior knowledge one uses, the more flexible the estimation is and the results will be based more on the observed data. With these prior-like penalties, our algorithm yields more precise parameter estimates for the M3PL model.

3.2. Computation via GVEM

This section introduces the main estimation algorithm to obtain the estimate $(\hat{A}_\lambda, \hat{B}_\lambda, \hat{C}_\lambda)$ via (6) using GVEM algorithm. As introduced in Sect. 2, we will use a variational lower bound to approximate the intractable marginal log-likelihood $l(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{Y})$ in (6).

To derive a lower bound for easy estimation of the M3PL parameters, instead of directly working with (4), we employ an equivalent representation of the M3PL model with auxiliary latent variable Z_{ij}, which is an indicator function of whether the ith individual answers the jth item based on the latent ability or guesses it correctly (von Davier, 2009). Specifically $Z_{ij} = 1$ if the ith individual solves item j based on his/her ability, and $Z_{ij} = 0$ if he/she guesses item j correctly. The distribution of Y_{ij} given the latent variables $\boldsymbol{\theta}_i$ and Z_{ij} is then

$$P(Y_{ij}|Z_{ij}, \boldsymbol{\theta}_i) = \left\{ \left[\frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} \right]^{Y_{ij}} \left[\frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} \right]^{1-Y_{ij}} \right\}^{Z_{ij}} I(Y_{ij} = 1)^{1-Z_{ij}},$$

where we define $0^0 = 1$, and it can be seen that this new model with auxiliary variable Z is equivalent to the M3PL model (von Davier, 2009, Cho et al., 2021). Denote $\boldsymbol{Z}_i = \{Z_{i1}, Z_{i2}, \ldots, Z_{iJ}\}$ and its distribution as $p(\boldsymbol{Z}_i) = \prod_{j=1}^{J} p(Z_{ij})$. Then the complete data likelihood of the ith subject can be written as

$$\begin{aligned}
&\log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \\
&= \log P(Y_i \mid \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) + \log \phi(\boldsymbol{\theta}_i) + \log p(\boldsymbol{Z}_i) \\
&= \sum_{j=1}^{J} \left\{ Y_{ij} Z_{ij} (\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + Z_{ij} \log \frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} + (1 - Z_{ij}) \log I(Y_{ij} = 1) \right\} \\
&\quad + \log \phi(\boldsymbol{\theta}_i) + \log p(\boldsymbol{Z}_i), \qquad (7)
\end{aligned}$$

where ϕ denotes the normal probability density function for latent variable $\boldsymbol{\theta}$. Here, without loss of generality, we focus on the ith subject's likelihood function due to the independence of different subjects.

With the above representation, for any variational distribution functions q_i and r_{ij} (to be estimated later) of the latent variables $\boldsymbol{\theta}_i$ and Z_{ij}, similar to the derivation in Sect. 2, we have the

following variational lower bound, which generalizes (4),

$$\log P(Y_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \geq \int_{\boldsymbol{\theta}_i} \sum_{\boldsymbol{Z}_i} \log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \times q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) d\boldsymbol{\theta}_i \tag{8}$$

$$- \int_{\boldsymbol{\theta}_i} \sum_{\boldsymbol{Z}_i} \log \big(q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i)\big) \times q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) d\boldsymbol{\theta}_i, \tag{9}$$

where $r_i(\boldsymbol{Z}_i) = \prod_{j=1}^{J} r_{ij}(Z_{ij})$. Since (9) doesn't depend on parameters $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, we focus on (8) for the derivation of the lower bound. For (8), note that $\log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ takes the form of (7). To obtain a closed form lower bound expression for (8), we further use a local variational method (Bishop, 2006; Jordan, Ghahramani, Jaakkola, & Saul, 1999). Particularly, define $\xi_{i,j}$ as a variational parameter indexed by i and j, and let $\eta(\xi_{i,j}) = (2\xi_{i,j})^{-1}[e^{\xi_{i,j}}/(1+e^{\xi_{i,j}}) - 1/2]$. Let $\boldsymbol{\xi}_i = (\xi_{i,j}, j = 1, \cdots, J)$ denote the ith subject's variational parameters for the J items. Then following the local variational method (Bishop, 2006), we have

$$\log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \geq l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}_i; Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i),$$

where $l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}_i; Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i)$ is defined as

$$\begin{aligned}
&l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}_i; Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i) \\
&= \sum_{j=1}^{J} Z_{ij} \log \frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + \sum_{j=1}^{J} Z_{ij} Y_{ij} (\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + \sum_{j=1}^{J} \frac{1}{2} Z_{ij} (b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \xi_{i,j}) \\
&\quad - \sum_{j=1}^{J} Z_{ij} \eta(\xi_{i,j}) \{(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \xi_{i,j}^2\} \\
&\quad + \sum_{j=1}^{J} \big\{(1 - Z_{ij}) \log I(Y_{ij} = 1)\big\} + \log \phi(\boldsymbol{\theta}_i) + \log p(\boldsymbol{Z}_i),
\end{aligned} \tag{10}$$

and it gives a lower bound of $\log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ in (8). We then have the following expression for the variational lower bound of the marginal likelihood of all observed responses in (6),

$$l(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{Y}) = \sum_{i=1}^{N} \log P(Y_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \geq E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}),$$

with the lower bound $E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi})$ defined as

$$E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}) = \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \left[\sum_{\boldsymbol{Z}_i} l(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}_i; Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i) \times r_i(\boldsymbol{Z}_i) \right] \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \tag{11}$$

Appropriate choices of the variational distributions will lead to a closed form expression of the lower bound in (11). Particularly, following the derivations in Cho et al. (2021), the above

likelihood function implies that an optimal choice of q_i is $q_i(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \mu_i, \Sigma_i)$ where the mean and covariance are

$$\mu_i = \Sigma_i \times \sum_{j=1}^{J} \left\{ 2\eta(\xi_{i,j}) b_j + Y_{ij} - \frac{1}{2} \right\} (1 - Y_{ij} + E_r(Z_{ij}) Y_{ij}) \boldsymbol{\alpha}_j^\top, \tag{12}$$

$$\Sigma_i^{-1} = \Sigma_{\boldsymbol{\theta}}^{-1} + 2 \sum_{j=1}^{J} (1 - Y_{ij} + E_r(Z_{ij}) Y_{ij}) \eta(\xi_{i,j}) \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^\top, \tag{13}$$

and the variational distributions $r_{ij}(Z_{ij})$ are $r_{ij}(Z_{ij}) \sim Bernoulli(s_{ij})$, where $s_{ij} = 1$ if $Y_{ij} = 0$, and otherwise

$$s_{ij}^{-1} = 1 + \frac{c_j}{1 - c_j} \frac{1 + e^{\xi_{i,j}}}{e^{\xi_{i,j}}} \exp \Big\{ - Y_{ij} (\boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i] - b_j) + \frac{1}{2} (b_j - \boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i] - \xi_{i,j}) - \eta(\xi_{i,j}) \{ E_{q_i}[(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2] - \xi_{i,j}^2 \} \Big\}. \tag{14}$$

With the above chosen q_i's and r_{ij}'s, we aim to estimate model parameters $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, together with the introduced local variational parameters $\boldsymbol{\xi}$, by maximizing the variational lower bound of the marginal likelihood, $E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi})$ in (11), with the proposed penalties in (6), that is,

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\mathbf{C}}_\lambda, \hat{\boldsymbol{\xi}}) = \text{argmax}_{\mathbf{A},\mathbf{B},\mathbf{C},\boldsymbol{\xi}} \; E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}) - P_\lambda(\mathbf{A}) + P(\mathbf{B}) + P(\mathbf{C}) \tag{15}$$

The corresponding solution $(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\mathbf{C}}_\lambda)$ gives the our GVEM estimators for the penalized likelihood in (6).

To estimate $(\boldsymbol{A}, \mathbf{B}, \mathbf{C})$, we use the coordinate descent algorithm (Friedman, Hastie, Höfling, & Tibshirani, 2007; Friedman et al., 2010), which solves the target optimization problem by successively minimizing along each coordinate direction of $(\boldsymbol{A}, \mathbf{B}, \mathbf{C})$. For each item j, there are one difficulty parameter b_j, one guessing parameter c_j, and K discrimination parameters $\boldsymbol{\alpha}_j$. The coordinate descent algorithm updates each of the $K + 2$ variables according to the following updating rule. (Please see Appendix for a detailed derivation of the updating rule.) Note that the derivation of the below soft-thresholding update rule of a_{jk} can be viewed as from the proximal gradient descent algorithm (Beck & Teboulle, 2009). Define a function S to be a soft threshold operator such that

$$S(\delta, \lambda) = \text{sign}(\delta)(|\delta| - \lambda)_+, \tag{16}$$

where for any real number x, $\text{sign}(x)$ denotes the sign of x and x_+ denotes $\max\{0, x\}$. The model parameters, $\boldsymbol{\alpha}_j$'s, b_j and c_j are updated using Equations (17), (18), and (19), respectively,

$$\alpha_{jk} = \left[\sum_{i=1}^{N} (1 - Y_{ij} + s_{ij} Y_{ij}) \left(2\eta(\xi_{i,j}) [\Sigma_i + (\mu_i)(\mu_i)^\top]_{k,k} \right) \right]^{-1} \times S\left(\sum_{i=1}^{N} (1 - Y_{ij} + s_{ij} Y_{ij}) \left\{ (Y_{ij} - \frac{1}{2}) \mu_{i,k} + 2 b_j \eta(\xi_{i,j}) \mu_{i,k} \right.\right.$$

$$-2\eta(\xi_{i,j})\sum_{l\neq k}\alpha_{jl}[\Sigma_i+(\mu_i)(\mu_i)^\top]_{l,k}\Big\},\ \frac{\lambda}{|\hat{\alpha}_{jk}^{(0)}|^\gamma}\Big) \tag{17}$$

$$b_j=\frac{\sum_{i=1}^N(1-Y_{ij}+s_{ij}Y_{ij})\big[\frac{1}{2}-Y_{ij}+2\eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top\mu_i\big]+\frac{\mu_b}{\sigma_b^2}}{2\sum_{i=1}^N(1-Y_{ij}+s_{ij}Y_{ij})\eta(\xi_{i,j})+\frac{1}{\sigma_b^2}}, \tag{18}$$

$$c_j=\frac{\sum_{i=1}^N Y_{ij}(1-s_{ij})+\alpha-1}{N+\alpha+\beta-2}. \tag{19}$$

where $\hat{\alpha}_{jk}^{(0)}$ is the initial estimator of α_{jk} by the GVEM algorithm without including the penalty terms in (15). Additionally, the variational parameter ξ's are updated as

$$\xi_{i,j}^2=b_j^2-2b_j\boldsymbol{\alpha}_j^\top\mu_i+\boldsymbol{\alpha}_j^\top[\Sigma_i+\mu_i\mu_i^\top]\boldsymbol{\alpha}_j, \tag{20}$$

and the covariance can be updated as

$$\Sigma_{\boldsymbol{\theta}}=\frac{1}{N}\sum_{i=1}^N[\Sigma_i+\mu_i\mu_i^\top]. \tag{21}$$

To choose the constant sparsity parameter λ, we can apply popular information criteria, such as Akaike Information Criterion (AIC), Bayesian information criterion (BIC) and generalized information criterion (GIC) (Nishii, 1984; Y. Fan & Tang, 2013). We estimate the information criteria by substituting the log-likelihood with the variational lower bound from the GVEM algorithm. The sparsity parameter that minimizes these information criteria will be considered optimal. Our pilot study shows that the GIC method proposed for high-dimensional model selection in Y. Fan and Tang (2013) performs better than AIC and BIC, and hence GIC is used throughout the study.

The detailed algorithm of the regularized estimation of the loading structure via adaptive Lasso penalization is illustrated in Algorithm 1.

Algorithm 1
Regularization with Adaptive Lasso Penalization (M2PL as an example)

1: Set a range of λ. Choose $\gamma>0$.
2: Use GVEM algorithm to conduct EFA+rotation assuming all items load on all factors to initialize model parameters A_0, B_0, Σ_0 and obtain $\hat{A}_w:=[\hat{\alpha}_{jk}^{(0)}]_{J\times K}$
3: **for** each λ starting from smallest **do**
4: Update $\boldsymbol{A}$, $\boldsymbol{B}$, according to (17), (18), respectively. Update ξ and $\Sigma_{\boldsymbol{\theta}}$ as in (20) and (21). Iterate until convergence.
5: Estimate $G\hat{I}C$ with recent updates.
6: Set $\hat{A}_\lambda$,$\hat{B}_\lambda$ as the initial values for next step.
7: **end for**
8: Find λ^* that minimizes the information criteria. Calculate the evaluation criteria and save $\hat{Q}_\lambda$.
9: Re-estimate $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\Sigma_{\boldsymbol{\theta}}$ according to confirmatory factor analysis with $\hat{Q}_\lambda$ as the factor loading matrix.

Algorithm 2
Regularization with Lasso Penalization (M3PL as an example)

Set a range of λ.
Use GVEM algorithm to conduct EFA+rotation assuming all items load on all factors to initialize model parameters A_0, B_0, C_0, Σ_0.
for each λ starting from smallest **do**
  Update $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ according to (17) with $\lambda/|\hat{\alpha}_{jk}^{(0)}|^{\gamma}$ in (17) replaced by λ. Update $\boldsymbol{B}$ and $\boldsymbol{C}$ according to (18) and (19). Update ξ and $\Sigma_{\boldsymbol{\theta}}$ as in (20) and (21). Iterate until convergence.
  Re-estimate $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\Sigma_{\boldsymbol{\theta}}$ according to confirmatory factor analysis with most recent updates (i.e. $\hat{Q}_{\lambda}$) as the factor loading matrix.
  With re-estimated A, B, C, and $\Sigma_{\boldsymbol{\theta}}$, estimate $G\hat{I}C$.
  Set $\hat{A}_{\lambda}$,$\hat{B}_{\lambda}$, $\hat{C}_{\lambda}$ as the initial values for next step.
end for
Find λ^* that minimizes the information criteria. Calculate the evaluation criteria.

Remark 1. In addition to our choice of adaptive Lasso for $P_{\lambda}(\mathbf{A})$ in (6), there are generally other methods of penalization. For instance, J. Fan and Li (2001) showed that the Lasso penalization problem is suboptimal to their proposed method called smoothly clipped absolute deviation (SCAD) penalty as Lasso produces biased estimates for the large coefficients. They showed that the SCAD penalization enjoys asymptotic normality and oracle properties with proper choice of regularization parameters. Due to its solid theoretical properties, SCAD has been widely applied in variable selection problems (T. Wang, Xu, & Zhu 2012; Liu, Yao, & Li, 2016; Breheny & Huang 2011). Additionally, Minimax Concave Penalty (MCP) has been presented as a fast, continuous and nearly unbiased method of penalization and hence claimed to be a good alternative to Lasso (C. H. Zhang 2010). Truncated Lasso is also another popular penalization method (Shen, Pan, & Zhu, 2012; Xu & Shang, 2018); however, penalty function for these methods is non-convex and it makes local solutions to be nonunique in general, which is computationally challenging to solve as well. On the other hand, adaptive Lasso uses a convex penalty and it is computationally efficient, which makes it a good candidate for regularization problem under complex MIRT models. Hence, we choose adaptive Lasso for solving our regularized problem.

4. Simulation Study

4.1. Design

A simulation study was conducted to evaluate the performance of the regularized GVEM algorithm in identifying true item factor loading structure with both M2PL and M3PL models. Three manipulated factors were considered: (1) the number of dimensions was fixed at 3 and 5 (i.e., $K = 3, 5$); (2) the correlations among factors were fixed at either 0.1, 0.3, or 0.7; and (3) both between-item and within-item multidimensional structures were considered. The sample size was fixed at 2000 (i.e., $N = 2000$) and 100 replications were run.[1]

For the between-item MIRT model, the test length was 45, with 15 items loaded onto each factor. The true item parameters were selected from the 2013 NAEP item bank (combined national and state assessments) for grade 8. For the within-item MIRT, the true item discrimination parameters were simulated from $Unif(0.75, 2)$, and the difficulty parameters were drawn from the

[1] In our pilot study, we varied the sample size (i.e., N =2000 or 3000) and the number of replications (i.e., 100 or 500 replications) and noted that results were stabilized with 100 replications, and relative performance of different methods under different conditions was the same between two sample size settings. Hence, all results reported in the paper were based on $N = 2000$ and 100 replications.

standard normal distribution. Additionally in M3PL, the guessing parameters were fixed at 0.2. The generated item parameters resemble the item parameters in Table 6.1 of (Reckase, 2009) closely. When the dimension was 3, about 60% of the items were loaded onto one factor, about 25% were loaded onto two factors, and the rest were loaded onto all three factors, whereas for the 5-dimension conditions, about 60% , 20%, 20% of the items were loaded onto one, two, and three factors, respectively. In all cases, the latent traits $\boldsymbol{\theta}$ were simulated from $MVN(0, \Sigma_{\boldsymbol{\theta}})$ with variance 1, where $r = 0.1, 0.3$ or 0.7.

Six methods were compared in the study, and they are (1) traditional exploratory item factor analysis followed by the CF-Quartimax rotation. This method is denoted as "Rotation" in all results. For this method, during estimation, we did not assume any constraint on the item discrimination parameter but fixed the population covariance matrix to an identity matrix, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}$. The GVEM algorithm was used for model estimation. The final discrimination parameters were transformed to standardized factor loadings, the value of which was compared to 0.3 (Henson & Roberts, 2006; Costello & Osborne, 2005). We used $\boldsymbol{\psi}_j = \boldsymbol{U}^{-1}\boldsymbol{\alpha}_j$, where $\boldsymbol{U}^\top\boldsymbol{U} = \left(\boldsymbol{I} + (\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j)\hat{\boldsymbol{\Sigma}}_\theta\right)$ to obtain the standardized factor loadings. If $|\psi_{jk}|$ exceeds 0.3, the item is assumed to load on the corresponding factor. This transformation function worked for all simulated conditions except for the within-item structure, $r = 0.7$, $K = 3$, M2PL and M3PL. In these two conditions, we transformed the true discrimination parameters to standardized factor loadings, and found some values were smaller than 0.3. Under these two conditions, we set the cutoff values as 0.75 instead, as the true values were generated from $Unif(0.75, 2)$. Setting a different cutoff will certainly affect the results, and this, to some extent, implies the subjectivity in the traditional EFA rotation method. (2) Exploratory item factor analysis with fixed anchors, and it is denoted as "Fixed Anchors" in all results. For this method, we imposed constraint 1 on the Q_A such that post-hoc rotation is no longer needed. We used the same transformation formula to calculate standardized factor loadings. This method was considered to ensure a direct and fair comparison to the regularization methods. (3) Lasso with constraint 1 and 2; and (4) adaptive Lasso with constraint 1 and 2. For the regularization methods, the tuning parameter λ was chosen by GIC. The GIC was computed as follows:

$$GIC = \log(\log(N)) \times \log(N) \times k - 2 \times E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}),$$

where N refers to the sample size, k refers to the number of parameters estimated by the model, $E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi})$ refers to the lower bound.

In additional to the two constraints for the model ability, we truncated $\hat{\boldsymbol{\alpha}}_{jk}$ to 0 if $|\hat{\boldsymbol{\alpha}}_{jk}| < 0.001$. As to γ in adaptive Lasso, Zou (2006) recommended three values, 0.5, 1, and 2. A few pilot trials were conducted to decide on the optimal γ, and $\gamma = 2$ was used for all conditions except a few conditions in which case $\gamma = 1$ was used. These conditions are within-item M2PL, $r = 0.7, k = 5$, constraint 1 and 2, as well as between-item M2PL, $r = 0.7, k = 5$, constraint 2 only.

As the main objective of this section is to estimate relationship between test items and latent traits, we used the correct estimation rate of A matrix (eq. (22)). It measures how well the sparsity of the A matrix is estimated by the regularized estimation. Notice that we only calculated correct rate for entries excluding the first K by K sub-matrix since we fixed this part to have identity matrix as a zero structure to ensure identifiability.

$$CR = \frac{1}{K \times J} \sum_{1 \le j < J, 1 \le k \le K} I(\hat{Q}_{jk} = Q_{jk}^{true}) \tag{22}$$

We also compared the performance of Lasso and adaptive Lasso penalization using two measures: sensitivity and specificity. In our context, sensitivity is the probability of correctly identifying nonzero entries among true nonzero entries. Specificity is the probability of correctly identifying zero entries among true zero entries. In other words, sensitivity measures the true negative rate, while specificity illustrates the true positive rate. Naturally, a test with both high sensitivity and high specificity is desired, although there is always a trade-off.

Other criteria include the average relative bias and root mean squared error (RMSE). The parameter recovery for $\Sigma_{\boldsymbol{\theta}}$ is calculated by taking differences between each freely estimated entries of the true $\Sigma_{\boldsymbol{\theta}}$ and estimated $\hat{\Sigma}_{\boldsymbol{\theta}}$. Relative bias and RMSE were obtained for each nonzero model parameter across all items within a condition first and then averaged over 100 replications.

4.2. Simulation Results

In this section, we first present the simulation results under various settings in M2PL and M3PL with boxplots to show the distribution of correct estimation rates, sensitivities, and specificities. Among the three information criteria, GIC showed the best performance at selecting the optimal result as it favors the models that penalizes more on the number of parameters; thus, we present the simulation results with GIC selection criteria in figures in this section.

Figures 1 and 2 show the recovery of item factor loading structure in terms of correct rates, sensitivity and specificity under M2PL and M3PL, respectively. All six methods were presented in the same order under each manipulated condition in Figures 1, 2, 3, 4, 5, 6. For M2PL, the adaptive Lasso method is consistently the best-performing method under all conditions, except when $r = 0.1$, $K = 5$, item structure is within-item M2PL, and when $r = 0.3$, $K = 5$, item structure is between-item M2PL. Under these two conditions, EFA rotation method performs slightly better than the adaptive Lasso method. The EFA with fixed anchors and Lasso regularization methods, on the other hand, performs a lot worse. When $K = 3$, and within-item M2PL is used, EFA rotation method performs considerably worse than EFA with fixed anchors and adaptive Lasso methods. Between the two constraint settings, constraint 2 yields more free parameters and hence it is harder to handle than constraint 1. Therefore, it is not surprising that adaptive Lasso with constraint 1 performs slightly better than with constraint 2 in more challenging scenarios (i.e., higher correlation, larger K, and within-item multidimensionality), whereas the difference between the two types is almost negligible in simpler scenarios.

When M3PL is the data generating model, the recovery of item factor loading structure is generally worse than that from M2PL, with a decrement of correct rate, sensitivity, and specificity in the range of 5% to 20%. The general trend of the manipulated factors on the results stay the same as compared to M2PL. That is, increasing factor correlation or allowing item cross-loadings makes the recovery of factor structure harder, although adaptive Lasso still performs the best among the six methods in all conditions except when $r = 0.7$, $K = 5$ and test exhibits between-item multidimensional structure. In this case, EFA rotation method tends to excel.

Figure 3 presents the relative bias of model parameters under M2PL. When the test has 5 latent factors and $r = 0.1$ or 0.3, although relative bias vary slightly differently for different parameters, the results from the six methods are almost indistinguishable. In a between-item structure with 3 factors, the relative bias for b has more variability across replications. It is because the true parameters of some items are close to 0. The relative bias vary more for the within-item condition in general. In a within-item structure, the two regularization methods appear to produce less bias than the EFA rotation method especially for Σ_θ. Under $K = 3$, $r = 0.1$ or 0.7, within-item M2PL conditions, the relative bias values for Σ_θ estimated by EFA rotation fall outside of the range. Figure 4 presents the RMSE of model parameters under M2PL. Again, all six methods produce comparable RMSE when $r = 0.1$ under between-item condition. When the factor correlation increases, the EFA rotation method generates larger RMSE for α and Σ_θ. The same trend holds

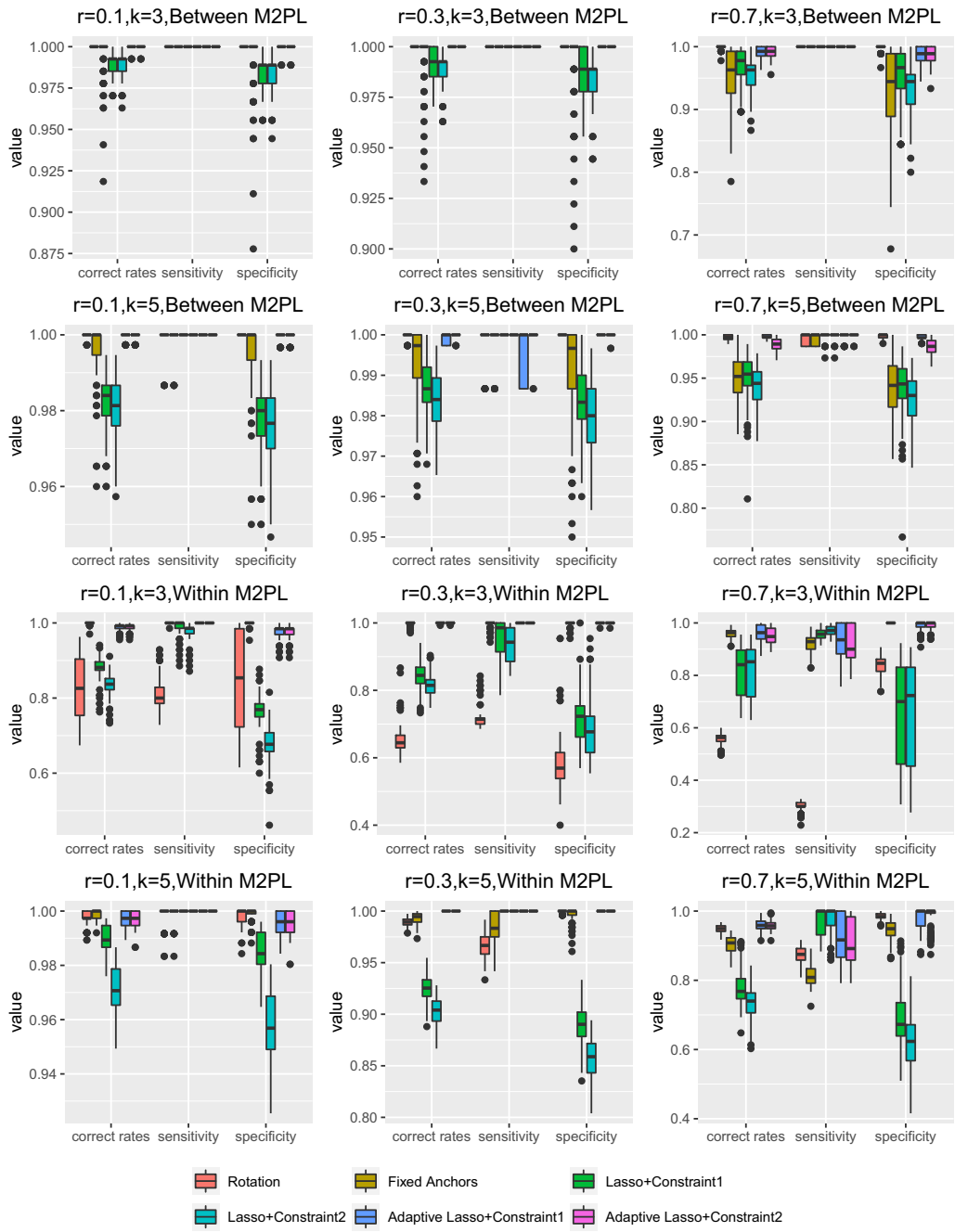

FIGURE 1.
Correct estimation rates of item factor loading structure under M2PL.

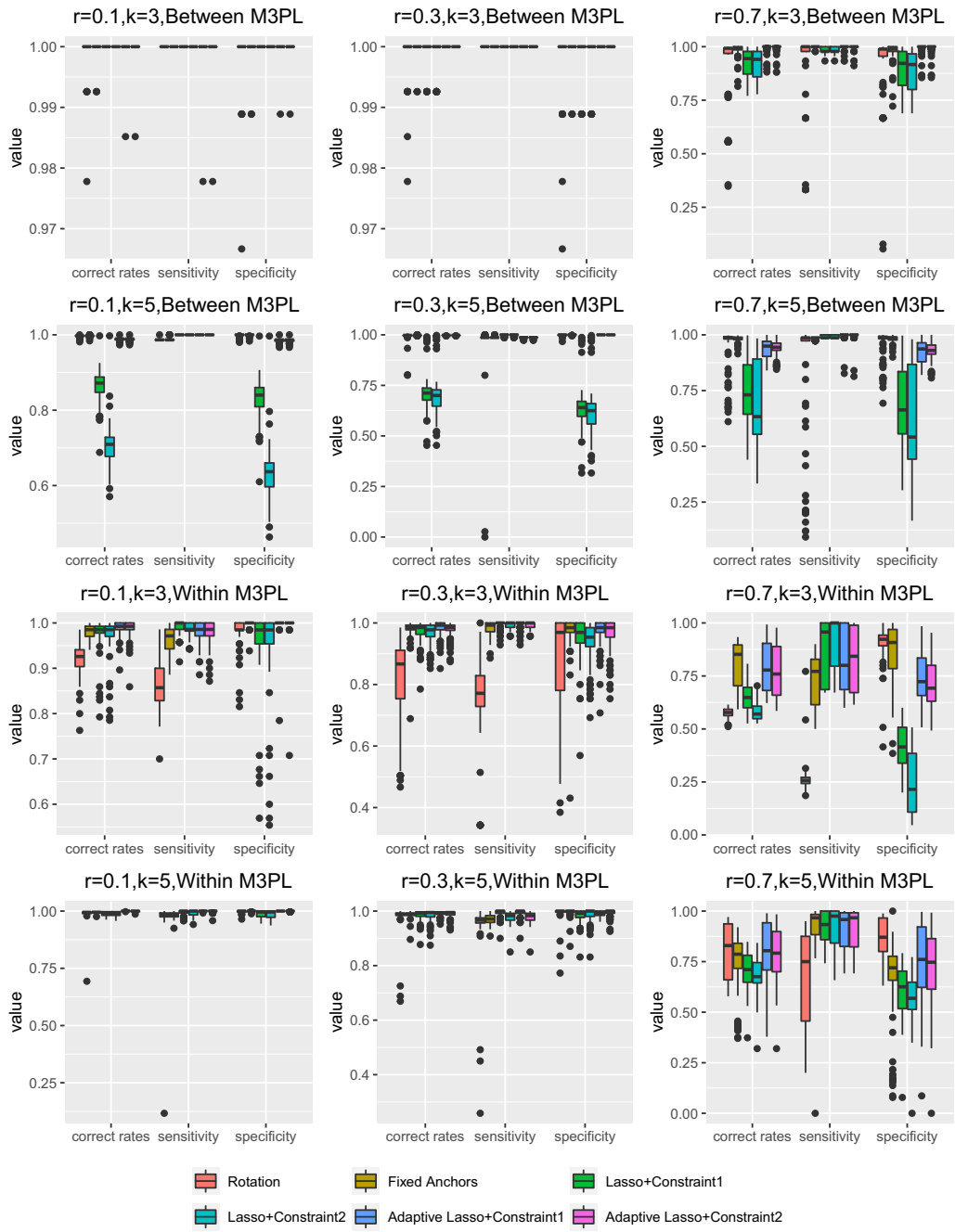

FIGURE 2.
Correct estimation rates of item factor loading structure under M3PL.

under the within-item M2PL conditions, although adaptive Lasso method seems to generate large RMSE for some conditions. Under the most difficult condition of $r = 0.7$, $K = 5$, one can see a lot of variability of RMSE across replications. In this case, the two Lasso methods seem to produce smaller median RMSE for majority of the parameters than EFA rotation method and EFA with fixed anchors method. The better performance of Lasso methods compared to adaptive Lasso may be because of two reasons: (1) we computed bias and RMSE only on those parameters whose true values were nonzero. Hence, even if the Lasso method fails to shrink some true zero loadings to zero, they will not count toward bias or RMSE. (2) Initial values play an important role in adaptive Lasso to determine an adaptive penalty weight. We used the results from EFA rotation methods as initial values, and other better initial values could be explored in the future, such as the SVD method in H. Zhang et al. (2020).

Figures 5 and 6 show the relative bias and RMSE of model parameters under M3PL. The inclusion of the guessing parameter, unsurprisingly, makes the model parameter recovery much harder, as shown in larger bias and RMSE as well as more variability across replications. The overall pattern observed from M2PL results continued to hold. That is, increasing factor correlation and using within-item factor structure not only increase relative bias and RMSE, but also yield more instability across replications. The EFA rotation method produces the largest average absolute bias and mean RMSE in almost all conditions, followed by EFA with fixed anchors method, although results from regularization methods seem to have more variability when the factor correlation is high.

In summary, the adaptive Lasso method outperforms EFA rotation method under almost all conditions regarding item factor loading structure recovery. There are only 3 exceptions: when $r = 0.1$, $K = 5$, item structure is within-item M2PL, when $r = 0.3$, $K = 5$, item structure is between-item M2PL, and $r = 0.7$, $K = 5$, item structure is between-item M3PL. In these three conditions, EFA rotation method performs better than the adaptive Lasso method by a small margin. Under some simple scenarios (i.e., low-correlation or medium-correlation and $K = 3$, item structure is between-item M2PL), there is no appreciable difference between the EFA rotation method and the adaptive Lasso method with either type of constraints. As for item parameter recovery, the adaptive Lasso method outperforms EFA rotation method for all of the high-correlation scenarios in M2PL. For small-correlation, between-item M2PL conditions, the results of adaptive Lasso and EFA rotation method appear to be indistinguishable. In M3PL, the adaptive Lasso method produces more accurate results compared to EFA rotation method under all conditions. Only under between-item M3PL conditions, EFA rotation generates smaller RMSE values and relative bias with less variability for Σ_θ, but it produces larger RMSE and relative bias for other parameters.

5. Real Data Analysis

In this section, the proposed regularization method was applied to the National Education Longitudinal Study of 1988 (NELS:88) data, and results were compared with those from EFA rotation method. NELS:88 was collected from a nationally representative sample of students whose performance on different cognitive batteries were tracked from 8th to 12th grade (the first three studies) in years 1988, 1990, and 1992. In this study, we focused on the science and mathematics test data where the multidimensional factorial structure has been previously investigated (e.g, Kupermintz & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997). Table 1 shows an example of the content of the questions in science test. For the science subject, there are 25 items and four factors were found from the data collected in 1988: "Elementary science (ES)", "Chemistry knowledge (CK)", "Scientific reasoning (SR)" and "Reasoning with knowledge (RK)". For the math subject, there are 40 items in 1988 and two factors emerged. They are "Mathematical reason-

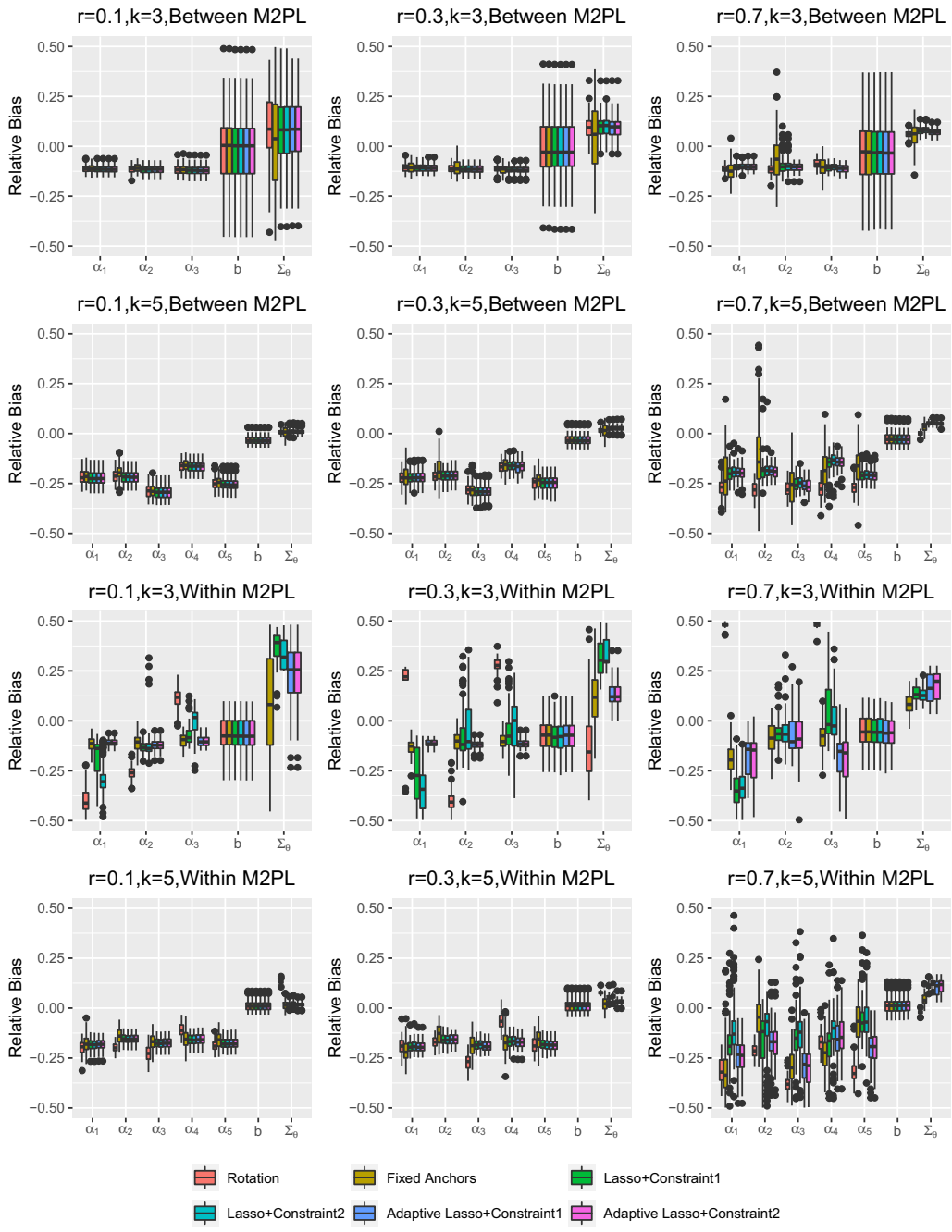

FIGURE 3.
Relative bias of model parameter estimates under M2PL.

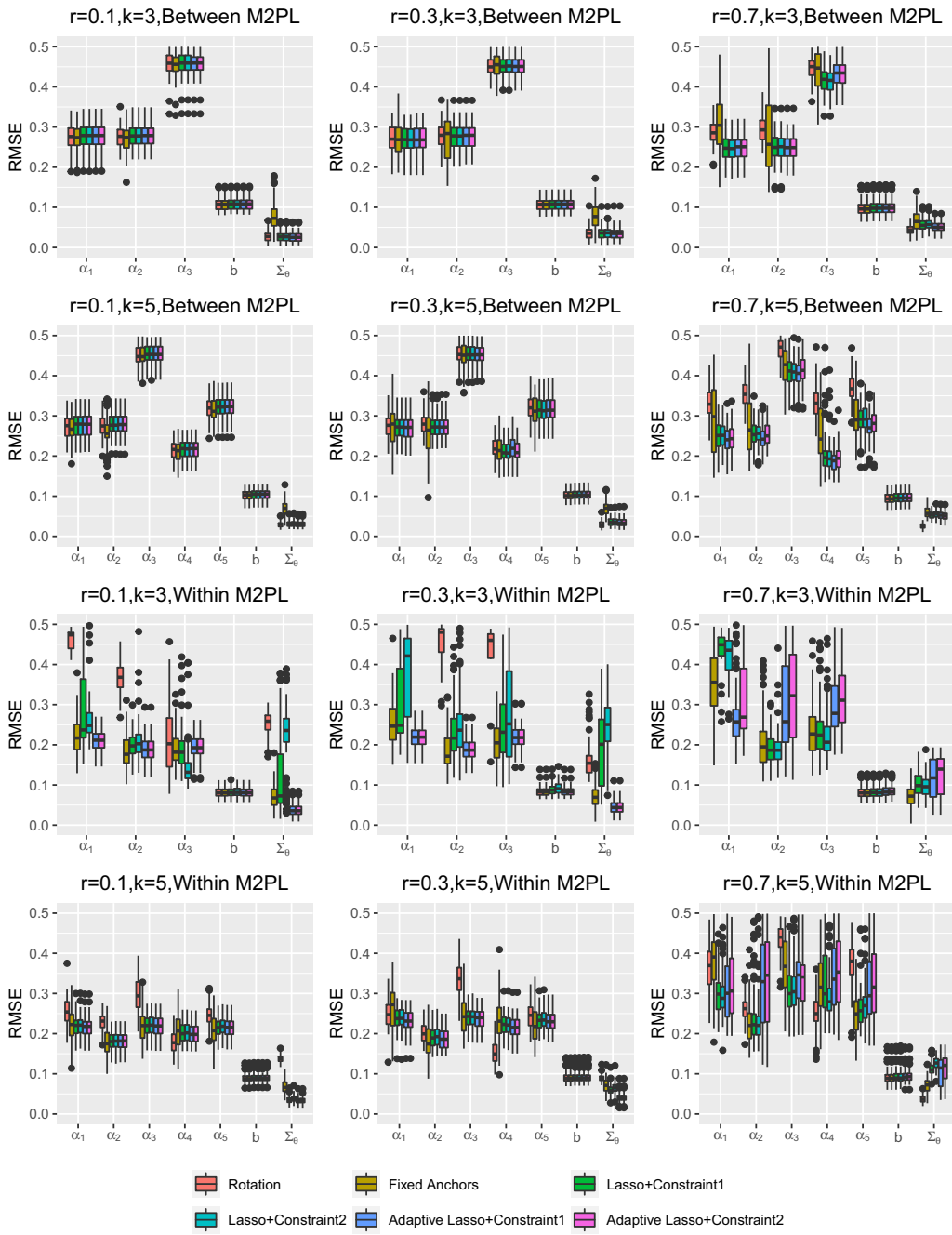

FIGURE 4.
RMSE of model parameter estimates under M2PL.

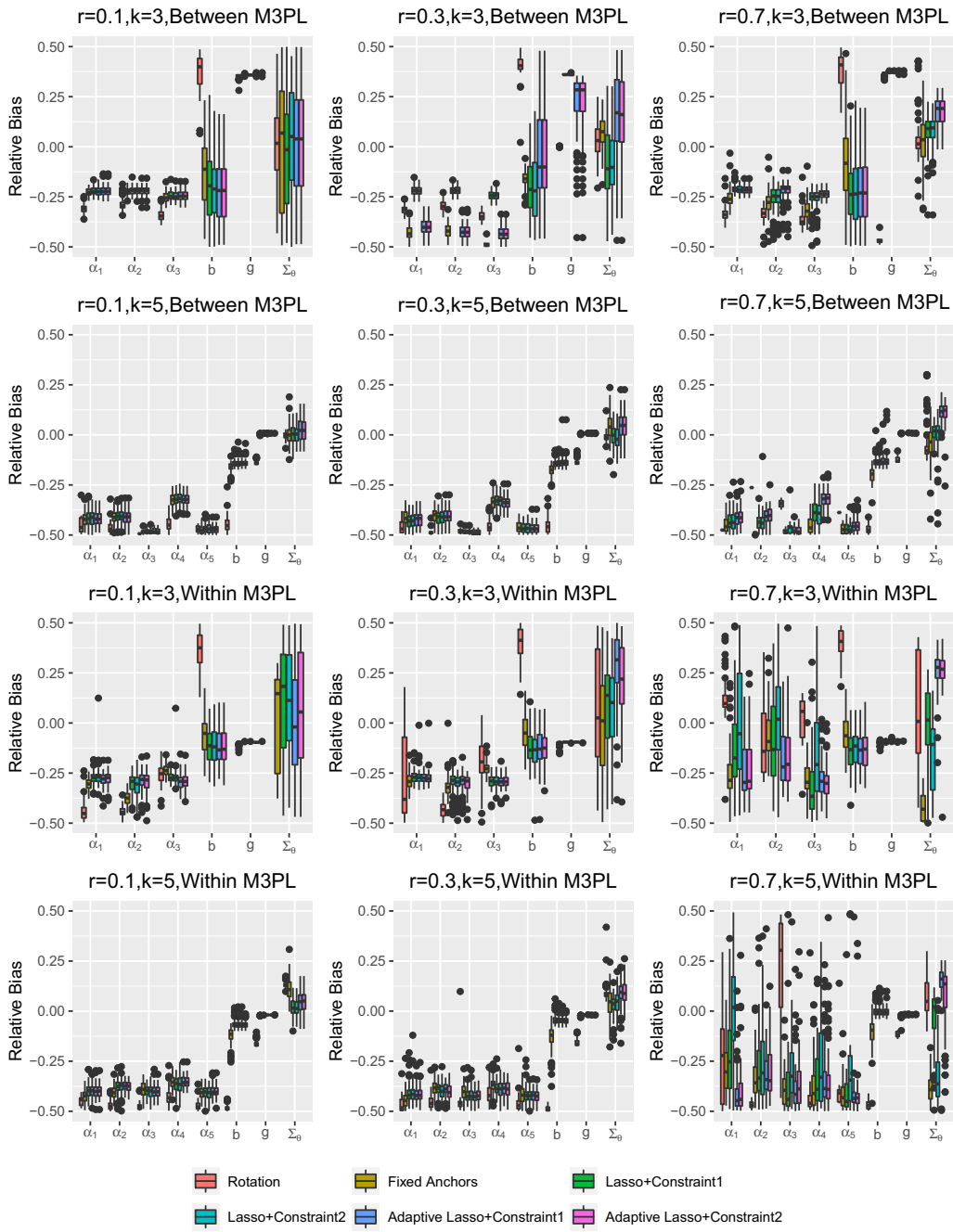

FIGURE 5.
Relative bias of model parameter estimates under M3PL.

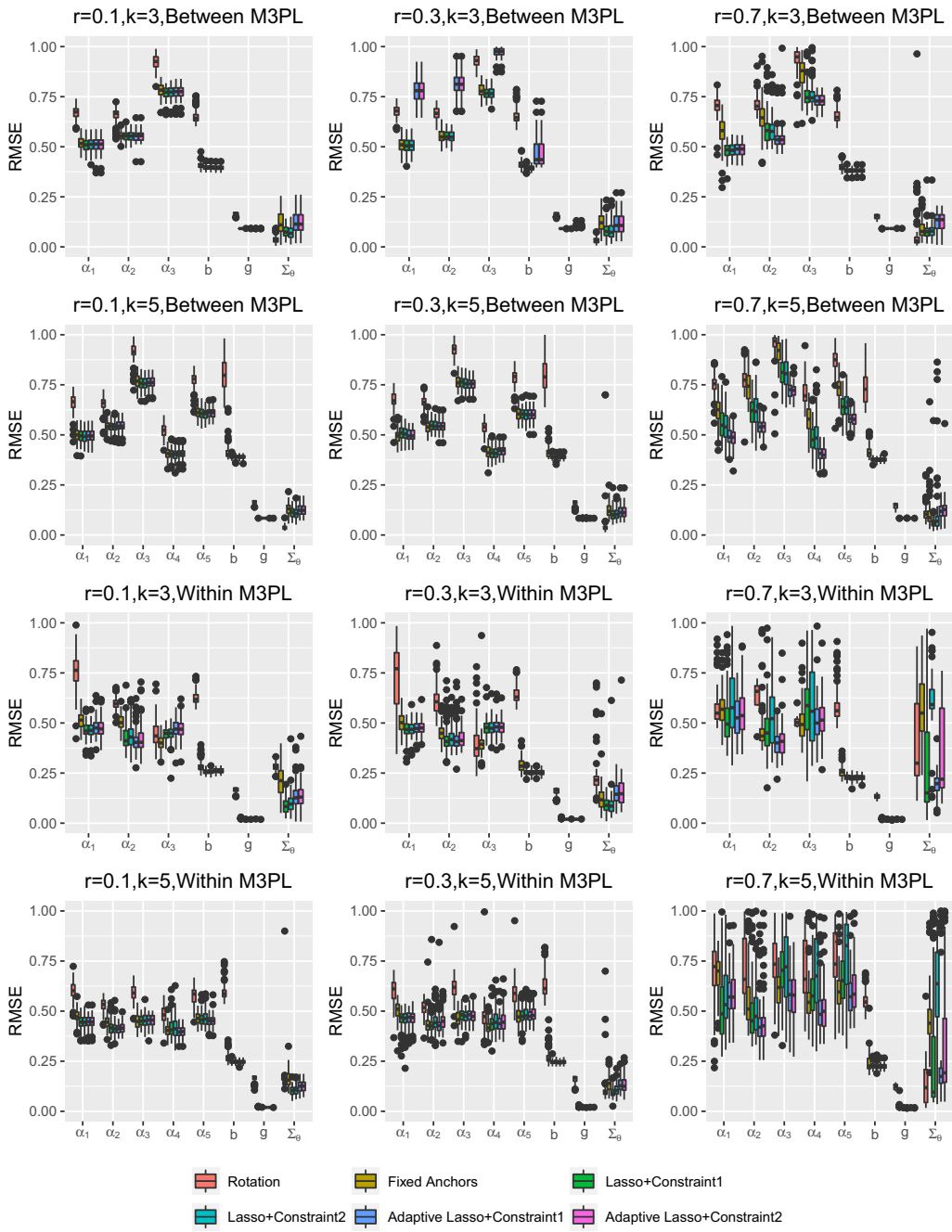

FIGURE 6.
RMSE of model parameter estimates under M3PL.

TABLE 1.
NELS:88 science items and descriptions were adopted from Rock et al. (1991).

Item	8th grade	10th grade	Description
S01	1		Infer geologic history from facts about limestone deposits
S02	2		Identify components of solar system
S03	3	2	Read a graph depicting solubility of chemicals
S04	4	3	Choose an improvement for an experiment on mice
S05	5	4	Choose a statement about source of moon's light
S06	6	5	Identify the example of a simple reflex
S07	7		Choose viable way of communicating on moon
S08	8		Select statement about position of sun, moon, earth in diagram
S09	9		Identify source of oxygen in ocean water
S10	10	1	Choose the property used to classify a list of substances
S11	11		Explain lower freezing temperature of ocean water
S12	12	6	Answer question about the earth's orbit
S13	13		Infer use of oxygen from description of condition of aquarium
S14	14	7	Estimate temperature of a mixture
S15	15	8	Select a statement about the process of respiration
S16	16	9	Read a graph depicting digestion of a protein by an enzyme
S17	17	10	Explain location of marine algae
S18	18	11	Choose best indication of an approaching storm
S19	19	12	Choose the alternative that is not a chemical change
S20	20	13	Infer statement from results of an experiment using a filter
S21	21	14	Explain reason for late afternoon breeze from the ocean
S22	22	15	Select basis for a statement about a food chain
S23	23	16	Interpret symbols describing a chemical reaction
S24	24	17	Differentiate statements based on a model or an observation
S25	25	18	Describe color of offspring from a guinea-pig cross
S26		19	Calculate a mass given density and dimensions
S27		20	Locate the balance point of a weighted lever
S28		21	Interpret a contour map
S29		22	Identify diagram depicting path of light through camera lens
S30		23	Calculate grams of a substance given its half life
S31		24	Read population graph; identify equilibrium point
S32		25	Identify cause of fire from overloaded circuit

S stands for Science items, and item descriptions were adopted from.

ing (MR)" and "Mathematical knowledge (MK)". We pooled together data from both domains, resulting in 65 items and a complete sample size of $N = 13,488$.

In the previous analysis of NELS:88 by Cho et al. (2020), the GVEM approach was used to empirically estimate the optimal number of latent traits from this data set. The result suggests there exists six latent traits measured by NELS:88. This finding is consistent with what the previous literature implies (e.g, Kupermintz & Snow, 1997; Nussbaum et al., 1997). Thus, we fix the dimension of latent factors as six for this analysis. Also, Kupermintz and Snow (1997) and Nussbaum et al. (1997) analyzed the latent traits required by each test item based on the content of the questions. Based on their findings, we chose 6 questions that only associate with each one of latent factors and performed our proposed regularized estimation under Constraint 1.

Both the EFA rotation (with the CF-Quartimax rotation) and adaptive Lasso methods with M2PL and M3PL were fitted to the data set. EFA rotation assumed all items load on all factors. For the adaptive Lasso method, we assumed all items load on all factors and hence penalty is added on

TABLE 2.
GIC comparison from two methods and two models (AL stands for adaptive Lasso).

M2PL		M3PL	
EFA	AL	EFA	AL
1.20×10^6	0.73×10^6	1.28×10^6	1.81×10^6

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS: 88), "Base Year Through Second Follow-up".

every element in the loading matrix except for the constraints. Note that we also consider another version of adaptive Lasso which assumes math items only load on the two math factors, and science items only load on the four science factors, hence there are structural 0's in the loading matrix that we neither estimate nor add penalty on. The results of this version are reported in the online supplementary materials to save space. Given the large sample size and test length (65 items in total), stochastic version of GVEM algorithm was used for M3PL. Specifically we used a stochastic sampling of 200 at each iteration and initially sampled 3000 for more stable convergence. For models with penalty, only adaptive Lasso was considered at it was shown to perform better than Lasso penalty under majority conditions in the simulation studies. The penalty parameter γ was fixed at 3 in adaptive Lasso. This is because the item factor loading structure is more complex (as compared to simulation study), hence, heavier penalization (i.e., higher γ) was used to produce a nicer sparse structure. Table 2 shows that the M2PL in general yields smaller GIC than M3PL. For the same model, adaptive Lasso produces the smaller GIC compared to the EFA rotation method. The fact that M2PL is preferred over M3PL implies that guessing may not play a big role on the performance in NELS:88 math and science assessments. Moreover, larger GIC from EFA rotation method implies that the factor loading structure obtained from it may not reflect the true item factor relationship as closely as the adaptive Lasso method.

Next, using M2PL with adaptive Lasso, Tables 3 and 4 illustrate the estimated sparse test structure from math and science test, respectively, and Tables 5 and 6 present the results from the EFA rotation method. Note the order of the latent traits for the EFA rotation method is arbitrary. As shown, for the EFA rotation method, before using 0.3 as the cutoff, we observe more cross-loadings, but after standardizing the factor loadings and using 0.3 as the cutoff, the EFA rotation method yields more spare and close to simple structure compared to the adaptive Lasso method. However, it appears that, in the EFA rotation method, both math and science items load dominantly on a single factor, which contracts with the findings that 2 and 4 best reflect the underlying factor structure. On the other hand, item factor loadings obtained from adaptive Lasso method, although they are less sparse and contain more cross loadings, appear to be more reasonable.

Tables 7 and 8 present the estimated factor correlations for two methods. Factor correlations obtained from EFA are in the range of 0.01 to 0.73, which are much lower than those from the regularization method. Adaptive Lasso estimated the correlations between latent factors in the range of 0.81 to 0.99. Such discrepancy could be explained from simulation findings. That is, the simulation results indicate EFA rotation method appears to underestimate the factor correlation especially when the true correlation is high and the number of factors is large. Also, GIC favors the regularization method which implies that high factor correlations are likely present in the data. The observed high correlation also appears to be consistent with decades of research on NELS data that treats math and science as unitary constructs.

TABLE 3.
Estimated sparse test structure for math test in NELS:88 (adaptive Lasso).

Factor	MR	MK	ES	SR	CK	RK
M1	0	0.768	0.923	0	0	0
M2	0	0.645	0.500	0	0	0
M3	0.899	0	0	0	0.940	0
M4	0.470	1.009	0	0	0	0
M5	0	1.484	0	0	0	0
M6	1.149	0	0	0	0.812	0
M7	1.016	0	0	0	0.263	0
M8	0	1.009	0	0	0.041	0
M9	1.373	0	0	0	1.255	0
M10	0	0	0	0	6.625	0
M11	1.182	0	0	0	1.361	0
M12	0	0	0	0	6.259	0
M13	1.466	0	0	0	0	0
M14	0	1.154	0	0	0	0
M15	3.535	0	1.345	0	0	0
M16	1.573	0	0	0	0	0
M17	0	0	0	0	5.784	0
M18	0.867	0	0	0	0	0
M19	1.459	0	0	0	0	0
M20	1.021	0	0	0	0	0
M21	1.521	0	0	0	0	0
M22	1.709	0	0.821	0	0	0
M23	0.449	0	0.496	0	0	0
M24	0.412	0	0.466	0	0	0
M25	1.456	0	0	0	0	0
M26	0.954	0	0.390	0	0	0
M27	0	0.463	0	0	0	0
M28	0.758	0	0.328	0	0	0
M29	0	0	2.903	0	0	0
M30	1.299	0	0	0	1.689	0
M31	0	0.855	0.715	0	0	0
M32	1.222	0	0	0	1.186	0
M33	0.423	0	0	0.099	0	0
M34	1.046	0	0	0	0.292	0
M35	0	0.706	0.147	0	0	0
M36	1.120	0	0	0	1.887	0
M37	0.727	0	1.049	0	0	0
M38	0	1.765	0	0	0	0
M39	0	0.475	1.460	0	0	0
M40	1.447	0	0	0	0.640	0

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS: 88), “Base Year Through Second Follow-up”.

TABLE 4.
Estimated sparse test structure for science test in NELS:88 (adaptive Lasso).

Factor	MR	MK	ES	SR	CK	RK
S1	0	0	0	0	0.992	0
S2	0	0	0.822	0	0	0
S3	0.175	0	0.629	0	0	0
S4	0	0	0.695	0	0	0
S5	0	0	1.483	0	0	0
S6	0	0	1.338	0	0	0
S7	0	0	0	0.966	0	0
S8	0	0	0	0.741	0	0
S9	0	0	2.469	0	0	0
S10	0.238	0	0	0	0.715	0
S11	0	0	0	0.542	0	0
S12	0	0.441	0	0.787	0	0
S13	0	0	0.900	0	0	0
S14	0	0.988	0	0	0.614	0
S15	0	0	0	0	0	0.65
S16	0.351	0	0	0	0.313	0
S17	0	0	0.899	0	0	0
S18	0	0	0	0.056	1.093	0
S19	0	0	0	0	1.491	0
S20	0.164	0	0	0	0.368	0
S21	0	0	0	0	0.535	0
S22	0	0	0.563	0	0	0
S23	0	0	0	0	1.620	0
S24	0	0	0	0	0.960	0
S25	0.154	0	0	0	0.404	0

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS: 88), "Base Year Through Second Follow-up".

6. Discussions

Exploratory factor analysis (EFA) is a popular statistical tool to gain insight into latent structures underlying the observed data (Gorsuch, 1988; Fabrigar & Wegener, 2011). Exploratory item factor analysis is a subset of EFA methods that deals with categorical observed data. In exploratory IFA, the relationship among observed item responses are explained by a few number of common factors. The naming of the common factors can be inferred from the content of the items that load on those factors, and hence, a simple structure with items loading exclusively on a single factor is usually preferred.

In this paper, a Gaussian variational regularization method is proposed for the estimation of the sparse item-trait relationship in M2PL and M3PL models. This computationally efficient method estimates both item factor loading structure and model parameters simultaneously. Both Lasso and adaptive Lasso penalties are considered, and simulation studies demonstrate that they perform well in correctly estimating the sparse item-trait structure for both M2PL and M3PL models. Adaptive Lasso penalization is preferred between the two. With adaptive Lasso penalization, GIC is used to choose the tuning parameter λ, whereas the tuning parameter γ takes one of the three suggested values by Zou (2006). Adaptive lasso also outperforms traditional EFA rotation method in most of the simulation conditions, and the two methods are almost indistinguishable for simpler

TABLE 5.
Estimated sparse test structure for math test in NELS:88 (EFA rotation Method).

	Estimated item discrimination parameters						Estimated standardized factor loadings					
Factor	F1	F2	F3	F4	F5	F6	F1	F2	F3	F4	F5	F6
M1	0.700	0.288	0.167	0.053	0.097	0.106	0.536	0	0	0	0	0
M2	0.671	0.252	0.091	0.187	0.087	0.147	0.528	0	0	0	0	0
M3	1.001	0.134	0.162	0	0.017	0.146	0.685	0	0	0	0	0
M4	1.454	0.135	0	0.168	0	0	0.816	0	0	0	0	0
M5	1.238	0.135	0.062	0.269	0.071	0.035	0.767	0	0	0	0	0
M6	1.162	0	0.114	0	0.015	0.044	0.751	0	0	0	0	0
M7	0.954	0.074	0.169	0.017	0.127	0.051	0.672	0	0	0	0	0
M8	0.783	0.163	0.084	0.07	0.088	0.036	0.603	0	0	0	0	0
M9	1.315	0	0.074	0	0.167	0.056	0.781	0	0	0	0	0
M10	0.392	0.127	0.919	0	0.174	0.137	0	0	0.679	0	0	0
M11	1.087	0	0.023	0	0.18	0	0.711	0	0	0	0	0
M12	0.610	0.053	0.917	0	0.163	0.111	0.301	0	0.693	0	0	0
M13	1.362	0	0.094	0	0	0.042	0.793	0	0	0	0	0
M14	1.107	0.024	0.082	0.221	0	0.022	0.734	0	0	0	0	0
M15	0.340	0.166	0	0	0.164	0.116	0	0	0	0	0	0
M16	1.733	0	0	0	0	0.002	0.860	0	0	0	0	0
M17	0.225	0.124	0.832	0.004	0.215	0.169	0	0	0.628	0	0	0
M18	0.710	0.019	0.01	0	0.220	0	0.562	0	0	0	0	0
M19	1.421	0	0	0.006	0	0	0.800	0	0	0	0	0
M20	1.071	0.017	0	0	0.039	0	0.728	0	0	0	0	0
M21	1.588	0	0	0	0.022	0.063	0.837	0	0	0	0	0
M22	0.847	0.257	0	0	0	0.206	0.602	0	0	0	0	0
M23	0.600	0.246	0.003	0	0	0.221	0.485	0	0	0	0	0
M24	0.513	0.307	0.043	0	0.018	0.112	0.431	0	0	0	0	0
M25	1.315	0.003	0.111	0	0.091	0	0.788	0	0	0	0	0
M26	0.999	0.201	0.082	0	0	0.208	0.678	0	0	0	0	0
M27	0.454	0.106	0.037	0.162	0	0	0.407	0	0	0	0	0
M28	0.882	0.147	0.029	0.030	0	0.103	0.651	0	0	0	0	0
M29	0.577	0.378	0.155	0	0	0.148	0.448	0.348	0	0	0	0
M30	1.460	0	0.214	0	0.115	0	0.802	0	0	0	0	0
M31	0.779	0.296	0.172	0.175	0.239	0.289	0.536	0	0	0	0	0
M32	1.244	0	0.173	0	0.096	0	0.759	0	0	0	0	0
M33	0.601	0.059	0	0.006	0.241	0	0.470	0	0	0	0	0
M34	1.053	0.050	0.117	0.093	0.207	0	0.704	0	0	0	0	0
M35	0.734	0.290	0	0.121	0	0	0.566	0	0	0	0	0
M36	1.355	0.079	0.316	0	0.062	0.229	0.751	0	0.304	0	0	0
M37	0.923	0.344	0.156	0.040	0.087	0.453	0.574	0	0	0	0	0.422
M38	1.542	0.132	0	0.205	0.011	0.142	0.827	0	0	0	0	0
M39	0.532	0.132	0.056	0.213	0	0.202	0.438	0	0	0	0	0
M40	1.539	0.147	0.182	0.085	0.161	0.186	0.803	0	0	0	0	0

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS: 88), “Base Year Through Second Follow-up”.

TABLE 6.
Estimated sparse test structure for science test in NELS:88 (Rotation Method).

	Estimated item discrimination parameters						Estimated standardized factor loadings					
Factor	F1	F2	F3	F4	F5	F6	F1	F2	F3	F4	F5	F6
S1	0.135	0.145	0.116	0.101	0.370	0.426	0	0	0	0	0.321	0.393
S2	0.092	0.145	0.096	0.093	0.235	0.399	0	0	0	0	0	0.371
S3	0.276	0.063	0.022	0	0.114	0.456	0	0	0	0	0	0.416
S4	0	0.245	0.103	0	0.164	0.322	0	0	0	0	0	0.307
S5	0.284	0.307	0.109	0.107	0.303	0.727	0	0	0	0	0	0.598
S6	0.280	0.271	0.153	0.060	0.287	0.613	0	0	0	0	0	0.529
S7	0	0.138	0	0.117	0.254	0.602	0	0	0	0	0	0.519
S8	0.014	0.144	0	0.073	0.336	0.388	0	0	0	0	0	0.363
S9	0.136	0.193	0	0.017	0.006	0.573	0	0	0	0	0	0.499
S10	0.310	0.132	0.094	0.043	0.377	0.215	0	0	0	0	0.347	0
S11	0.056	0.04	0	0.058	0.309	0.237	0	0	0	0	0	0
S12	0.426	0.168	0.027	0.198	0.301	0.420	0.325	0	0	0	0	0.391
S13	0.055	0.23	0.215	0.06	0.067	0.515	0	0	0	0	0	0.459
S14	1.007	0.159	0.091	0.303	0.307	0.170	0.656	0	0	0	0	0
S15	0.124	0	0.087	0.019	0	0.645	0	0	0	0	0	0.543
S16	0.273	0.248	0.040	0	0.332	0	0	0	0	0	0.315	0
S17	0.219	0	0	0.104	0.494	0.279	0	0	0	0	0.430	0
S18	0.278	0	0	0.193	0.507	0.311	0	0	0	0	0.436	0
S19	0.206	0.002	0.059	0.095	0.422	0.291	0	0	0	0	0.375	0
S20	0.196	0	0	0	0.392	0.114	0	0	0	0	0.363	0
S21	0.072	0.170	0	0.01	0.538	0	0	0	0	0	0.474	0
S22	0.045	0.099	0	0	0.158	0.358	0	0	0	0	0	0.337
S23	0	0.278	0.094	0	0.399	0	0	0	0	0	0.362	0
S24	0.160	0.061	0.101	0.082	0.290	0.523	0	0	0	0	0	0.465
S25	0.167	0.114	0.073	0	0.126	0.199	0	0	0	0	0	0

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS: 88), "Base Year Through Second Follow-up".

TABLE 7.
Estimated Correlation between latent factors (Adaptive Lasso).

	MR	MK	ES	SR	CK	RK
MR	1.0000	0.9808	0.8465	0.7740	0.8646	0.8328
MK	0.9808	1.0000	0.9242	0.8684	0.9364	0.9119
ES	0.8465	0.9242	1.0000	0.9901	0.9968	0.9931
SR	0.7740	0.8684	0.9901	1.0000	0.9822	0.9807
CK	0.8646	0.9364	0.9968	0.9822	1.0000	0.9873
RK	0.8328	0.9119	0.9931	0.9807	0.9873	1.0000

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS: 88), "Base Year Through Second Follow-up".

scenarios such as lower factor correlation and lower dimensions. Since a user specified cutoff is needed to decide "significant" factor loadings, future studies could consider sparsity-encouraging rotation (e.g., Jennrich, 2006) to avoid arbitrarily truncating the rotated factor loadings.

TABLE 8.
Estimated Correlation between latent factors (EFA Rotation).

	F1	F2	F3	F4	F5	F6
F1	1.0000	0.5784	0.7344	0.0848	0.6897	0.6780
F2	0.5784	1.0000	0.3224	0.3783	0.4007	0.7268
F3	0.7344	0.3224	1.0000	0.0067	0.4671	0.4186
F4	0.0848	0.3783	0.0067	1.0000	0.1829	0.2630
F5	0.6897	0.4007	0.4671	0.1829	1.0000	0.5051
F6	0.6780	0.7268	0.4186	0.2630	0.5051	1.0000

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS: 88), "Base Year Through Second Follow-up".

The current study can also be expanded in the following directions. First, we assume the total number of factors, K, is known in advance. However, this assumption may be freed by varying K and use GIC as the model selection criterion to select the optimal K. This approach is in contrast to the family of criteria based on eigenvalues of the sample tetrachoric or polychoric correlation matrix of the observed data. Examples of this latter approach include the scree test (Cattell, 1966), the parallel analysis (Horn, 1965), among others. Future studies on evaluating the relative performance of these two approaches are worth pursuing. Note that because K is usually defined as the minimum number of latent common factors that is needed to describe the statistical dependencies in data, challenges may arise when there are additional nuisance factors, such as a bi-factor structure.

Second, the proposed method may be generalized to other types of MIRT models, such as the non-compensatory models (e.g., C. Wang & Nydick, 2015), which is essentially a nonlinear item factor model. While the adaptive Lasso idea can be directly applied, more work is needed to derive a suitable variational lower bound to enable the GVEM algorithm.

Third, it is of interest to further study the theoretical properties of the estimation and the model selection consistency for the proposed method. As shown in Cho et al. (2021), the GVEM algorithm (without additional penalty) can consistently estimate the model parameters of the 2-parameter MIRT model under a global Frobenius norm evaluation and the asymptotic regime when both N and J increase to infinity. With the additional adaptive Lasso penalty, it is expected that a similar global consistency result would hold when the tuning parameter is properly chosen. For instance, with the tuning parameter $\lambda = 0$, the proposed estimator becomes that in Cho et al. (2021) and the consistency result then follows. Moreover, it would also be of interest to study the variable selection consistency as well as the oracle properties as in Zou (2006) under the MIRT setting. However, such a problem is much more challenging due to several reasons. First, additional work is still needed to derive the entry-wise consistency and convergence rate results under the double asymptotic regime with $N, J \to \infty$. In particular, to show the oracle properties, we would need a sharp characterization of the entry-wise convergence rate of the GVEM estimators, which, however, is a challenging problem in the high-dimensional MIRT model. Second, the theoretical analysis of adaptive Lasso (or other penalties) is more challenging under the high-dimensional latent variable models, such as MIRT, and the variational approximation further complicates the problem. In fact, the frequentist consistency properties of many variational approximation methods remain unaddressed in the current literature. For such reasons, we would leave this interesting problem for future study.

Finally, it is interesting to obtain standard errors of the proposed regularized estimators. For variational approximations, the commonly used de-biasing technique in high-dimensional statistics may not be directly applicable, due to the additional approximation bias induced by

the variational method. One way to reduce such a variational bias is to perform an importance sampling based reweighing after the variational estimation so that the likelihood function can be better approximated (Domke & Sheldon, 2018); then a de-biasing step for the regularized estimation could be used to obtain the standard error estimates. Another approach is to use the bootstrap method to obtain the standard errors. As the setting of variational estimation for MIRT differs from many of the existing works on de-biasing estimation or bootstrap, the theoretical consistency properties of these methods are challenging and remain open problems in the literature. We therefore leave this interesting problem for future study.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics, 17*(3), 251–269.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences, 2*(1), 182–202.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261–280.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithm for nonconvex penalized regression, with application to biological feature selection. *The Annals of Applied Statistics, 74,* 5232–253.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111–150.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75*(1), 33–57.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307–335.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276.

Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika, 84*(1), 124–146.

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 4,* 7452–85.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1–13), 7.

Domke, J. , & Sheldon, D. (2018). Importance weighting and variational inference. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 4475–4484).

Embretson, S. E., & Reise, S. P. (2000). *Psychometric methods: Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 2,* 961348–1360.

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B: Statistical Methodology, 7,* 531–552.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics, 1*(2), 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1.

Gorsuch, R. L. (1988). Exploratory factor analysis. In *Handbook of multivariate experimental psychology* (pp. 231–258). Springer.

Hendrickson, A. E., & White, P. O. (1966). A method for the rotation of higher-order factors. *British Journal of Mathematical and Statistical Psychology, 19*(1), 97–103.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393–416.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research, 14*(1), 1303–1347.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Jennrich, R. I. (2004). Rotation to simple loadings using component loss functions: The orthogonal case. *Psychometrika, 69*(2), 257–273.

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika, 71*(1), 173–191.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning, 37*(2), 183–233.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*(3), 187–200.

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III NELS: 8.8 mathematics achievement to 12th grade. *American Educational Research Journal, 34*(1), 124–150.

Liu, H., Yao, T., & Li, R. (2016). Global solutions to folded concave penalized nonconvex learning. *Annals of Statistics, 44*(2), 629.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association, 92*(437), 162–170.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics, 2,* 758–765.

Nussbaum, E. M., Hamilton, L. S., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: IV NELS: 88 science achievement to 12th grade. *American Educational Research Journal, 34*(1), 151–173.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146–178.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics, 4*(1), 12–35.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*(2), 301–323.

Reckase, M. D. (2009). *Multidimensional item response theory* (Vol. 150). Springer.

Rock, D. A., Pollack, J. M., Owings, J., & Hafner, A. (1991). *Psychometric report for the NELS: 88 base year test battery*. National Center for Education Statistics.

Shen, X., Pan, W., & Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association, 107*(497), 223–232.

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika, 81*(4), 921–939.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 2,* 47397–412.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81*(393), 82–86.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics, 35*(2), 174–193.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives., 7*(2), 110–114.

Wang, C., & Nydick, S. W. (2015). Comparing two algorithms for calibrating the restricted non-compensatory multidimensional irt model. *Applied Psychological Measurement, 39*(2), 119–134.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*(3), 456–477.

Wang, T., Xu, P. R., & Zhu, L. X. (2012). Non-convex penalized estimation in high-dimensional models with single-index structure. *Journal of Multivariate Analysis, 3,* 109221–235.

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*(1), 58.

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association, 113*(523), 1284–1295.

Yen, M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 4,* 52275–291.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics, 38*(2), 894–942.

Zhang, H., Chen, Y., & Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika, 5,* 85358–372.

Zhang, S., Chen, Y., & Liu, Y. (2020). An improved stochastic EM algorithm for large-scale full-information item factor analysis. *British Journal of Mathematical and Statistical Psychology, 73*(1), 44–71.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association, 7,* 1011418–1429.

Manuscript Received: 15 SEP 2020
Final Version Received: 9 MAY 2022
Accepted: 2 JUN 2022