GAUSSIAN VARIATIONAL ESTIMATION FOR MULTIDIMENSIONAL ITEM RESPONSE THEORY

April E. Cho

department of statistics, university of michigan,
ann arbor, mi, usa
email : aprilcho@umich.edu

Chun Wang

college of education, university of washington,
seattle, wa, usa.
email : wang4066@uw.edu

Xue Zhang

china institute of rural education development,
northeast normal university, changchun, china.
email : zhangx815@nenu.edu.cn

Gongjun Xu

department of statistics, university of michigan,
ann arbor, mi, usa
email : gongjun@umich.edu

Corresponding Author: Gongjun Xu, 456 West Hall, 1085 South University, Ann Arbor, MI, USA, 48109; Email gongjun@umich.edu.

Data Availability Statement: The data that support the findings of this study are available from Institute of Education Sciences. Restrictions apply to the availability of these data, which were used under license for this study. Data are available with the permission of Institute of Education Sciences.

Gaussian Variational Estimation for Multidimensional Item Response Theory

Abstract

Multidimensional Item Response Theory (MIRT) is widely used in assessment and evaluation of educational and psychological tests. It models the individual response patterns by specifying functional relationship between individuals' multiple latent traits and their responses to test items. One major challenge in parameter estimation in MIRT is that the likelihood involves intractable multidimensional integrals due to latent variable structure. Various methods have been proposed that either involve direct numerical approximations to the integrals or Monte Carlo simulations. However, these methods are known to be computationally demanding in high dimensions and rely on sampling data points from a posterior distribution. We propose a new Gaussian Variational EM (GVEM) algorithm which adopts a variational inference to approximate the intractable marginal likelihood by a computationally feasible lower bound. In addition, the proposed algorithm can be applied to assess the dimensionality of the latent traits in an exploratory analysis. Simulation studies are conducted to demonstrate the computational efficiency and estimation precision of the new GVEM algorithm in comparison to the popular alternative Metropolis-Hastings Robbins-Monro (MHRM) algorithm. In addition, theoretical results are also presented to establish the consistency of the estimator from the new GVEM algorithm.

Keywords: Multidimensional IRT, Variational Inference, EM algorithm

1 Introduction

The increasing availability of rich educational survey data and the emerging needs of assessing competencies in education pose great challenges to existing techniques used to handle and analyze the data, in particular when the data are collected from heterogeneous populations. Different forms of multilevel, multidimensional item response theory (MIRT) models have been proposed in the past decades to extract meaningful information from complex education data. The advancement of computational and statistical techniques, such as the adaptive Gaussian quadrature methods, the Metropolis-Hastings Robbins-Monro algorithm, the stochastic expectation maximization algorithm, or the fully Bayesian estimation methods, also help promote the usage of the MIRT models. However, even with these state-of-the-art algorithms, the computation can still be time-consuming, especially when the number of factors is large. The main aim of this paper is to propose a new Gaussian variational expectation maximization (GVEM) algorithm for high-dimensional MIRT models.

As summarized in Reckase (2009), the MIRT models contain two or more parameters to describe the interaction between the latent traits and the responses to test items. In this paper, we focus on the logistic model with dichotomous responses. Specifically for the multidimensional 2-Parameter Logistic (M2PL) model, there are N individuals who respond to J items independently with binary response variables Y_{ij}, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Then the item response function of the ith individual to the jth item is modeled by

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

where $\boldsymbol{\alpha}_j$ denotes a K-dimensional vector of item discrimination parameters for the jth item

and b_j specifies the corresponding difficulty level with item difficulty parameter as $b_j/\|\boldsymbol{\alpha}_j\|_2$. $\boldsymbol{\theta}_i$ denotes the K-dimensional vector of latent ability for student i.

For the multidimensional 3-Parameter Logistic (M3PL) model, there is an additional parameter c_j, which denotes the guessing probability of the jth test item. The item response function is expressed as

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

For both the M2PL and M3PL models, denote all model parameters as M_p. Then given the typical local independence assumption in IRT, the marginal log-likelihood of M_p given the responses $\mathbf{Y}$ is

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

where $Y_i = (Y_{ij}, j = 1, \ldots, J)$ is the ith subject's response vector and J is the total number of items in the test. The ϕ denotes the K-dimensional Gaussian distribution of $\boldsymbol{\theta}$ with mean 0 and covariance $\Sigma_{\boldsymbol{\theta}}$. The maximum likelihood estimators of the model parameters are then obtained from maximizing the log-likelihood function. However, due to the latent variable structure, maximizing the log-likelihood function involves a K dimensional integrals that are usually intractable. Direct numerical approximation to the integrals have been proposed in the literature, such as the Gauss–Hermite quadrature (Bock & Aitkin, 1981) and the Laplace approximation (Lindstrom & Bates, 1988; Tierney & Kadane, 1986; Wolfinger & O'connell, 1993). However, the Gauss–Hermite quadrature approximation is known to become com-

putationally demanding in the high-dimensional setting, which happens in MIRT especially when the dimension of latent traits increases. The Laplace approximation, though computationally efficient, could become less accurate when the dimension increases or when the likelihood function is in skewed shape. Other numerical approximation methods based on Monte Carlo simulations have also been developed in the literature, such as the Monte Carlo expectation-maximization (McCulloch, 1997), stochastic expectation-maximization (von Davier & Sinharay, 2010), Metropolis-Hastings Robbins-Monro algorithms (Cai, 2010b, 2010a). These methods usually depends on sampling data points from a posterior distribution and would be computationally involving. Recently, Zhang, Chen, and Liu (2020) proposed to use the stochastic EM algorithm (Celeux & Diebolt, 1985) for the item factor analysis, where an adaptive-rejection-based Gibbs sampler is still needed for the stochastic E step. Moreover, Chen, Li, and Zhang (2019) studied the joint maximum likelihood estimation by treating the latent abilities as fixed effect parameters instead of random variables as in (3).

In this paper, we propose a computationally efficient method that is based on the variational approximation to the log-likelihood. Variational approximation methods are mainstream methodology in computer science and statistical learning, and they have been applied to diverse areas including speech recognition, genetic linkage analysis, and document retrieval (Blei & Jordan, 2004; Titterington, 2004). Recently, there is an emerging interest in developing and applying variational methods in statistics (Blei, Kucukelbir, & McAuliffe, 2017; Ormerod & Wand, 2010). In particular, Gaussian variational approximation methods were developed for standard generalized linear mixed effects models (GLMM) with nested random effects (Ormerod & Wand, 2012; Hall, Ormerod, & Wand, 2011). However, the variational

methods have only been slowly recognized in psychometrics and educational measurement, with the pioneer papers by Rijmen and Jeon (2013) as well as Jeon, Rijmen, and Rabe-Hesketh (2017).

In essence, variational approximations refer to a family of deterministic techniques for making approximate inference for parameters in complex statistical models (Ormerod & Wand, 2010). The key is to approximate the intractable integrals (e.g. Eq.(3)) with a computational feasible form, known as the variational lower bound to the original marginal likelihood. In psychometrics, Rijmen and Jeon (2013) first developed a variational algorithm for a high dimensional IRT model, but their algorithm was limited to only discrete latent variables. Recently, Jeon et al. (2017) proposed a variational maximization-maximization (VMM) algorithm for maximum likelihood estimation of GLMMs with crossed random effects. They showed that VMM outperformed Laplace approximation with small sample size. However, their study is limited in several respects: (i) They only considered the Rasch model. Although extending their algorithm to the 2PL model may be straightforward, its generalization to 3PL is unknown because 3PL does not belong to the GLMM family; (ii) The key component in their algorithm is the mean-field approximation (Parisi, 1988) that assumes independence of the latent variables given observed data. Even though it seems acceptable to assume independence of each random item effect, this independence assumption can no longer apply to the MIRT models when different dimensions are assumed to be correlated; (iii) In their first maximization step, the closed-form solution still contains a two-dimensional integration where adaptive quadrature is used; in the second maximization step, a Newton-Raphson algorithm is used. Therefore, both steps involve iterations, which may slow down the algorithm. Instead, our proposed GVEM algorithm has closed-form solutions

for all parameters in both the E and M steps, and it can deal with high-dimensional MIRT models when the multiple latent traits are correlated. Moreover, the GVEM algorithm is established for both the M2PL and M3PL models. Consistency theory of the estimators from our proposed algorithm is established, and the performance of the algorithm is thoroughly evaluated via simulation studies.

The rest of the paper is organized as follows. Section 2 introduces the general framework of the Gaussian Variational method and derivation of EM algorithm in MIRT models. Section 3 presents the GVEM algorithm for M2PL with the use of local variational approximation and presents the theoretical properties of the proposed algorithm. Section 4 extends the GVEM algorithm to M3PL and also presents the stochastically optimized algorithm to further improve its computational efficiency. Section 5 and section 6 illustrate the performance of the proposed GVEM method with simulation studies and on real data, respectively. The paper is concluded with Section 7, which discusses any future steps. The Supplementary Material includes the detailed mathematical derivations of the EM steps and the proofs of the theorem and proposition.

2 Gaussian Variational EM (GVEM)

From here onwards, for the MIRT models in (1) and (2), we denote the model parameters by $\boldsymbol{A} = \{\boldsymbol{\alpha}_j, j = 1, \ldots, J\}$, $\boldsymbol{B} = \{b_j, j = 1, \ldots, J\}$, and $\boldsymbol{C} = \{c_j, j = 1, \ldots, J\}$. As defined in Section 1, we use the notation $M_p = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$ in the 3PL model and $M_p = \{\boldsymbol{A}, \boldsymbol{B}\}$ in the 2PL model for simplicity. Latent traits $\boldsymbol{\theta}$ from different dimensions are correlated, resulting in a K by K covariance matrix $\Sigma_{\boldsymbol{\theta}}$. To fix the origin and units of measurement, it

is conventional to fix the mean and variance of all $\boldsymbol{\theta}$'s to be 0 or 1, respectively. To remove rotational indeterminacy in the exploratory analysis, (i.e. to ensure the model identifiability) researchers often assume $\Sigma_{\boldsymbol{\theta}} = I_K$ and $\boldsymbol{A}$ contains a K-dimensional triangular matrix of zeros (Reckase, 2009). On the other hand, in the confirmatory analysis, the zero structure of the loading matrix $\boldsymbol{A}$ is completely or partially specified while the remaining nonzero elements are left unknown. In this case, the correlation of latent traits $\boldsymbol{\theta}$ is of interest and we need to estimate the covariance matrix $\Sigma_{\boldsymbol{\theta}}$. In this paper, we consider a general setting of $\Sigma_{\boldsymbol{\theta}}$ that works for both exploratory and confirmatory analyses.

The idea of variational approximation is to approximate the intractable marginal likelihood function, which involves integration over the latent random variables, by a computationally feasible lower bound. We follow the approach of variational inference (Bishop, 2006) to derive this lower bound.

The marginal log-likelihood of responses $\mathbf{Y}$ is

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

where ϕ denotes a K-dimensional Gaussian distribution of $\boldsymbol{\theta}$ with mean 0 and covariance $\Sigma_{\boldsymbol{\theta}}$. Note that the log-likelihood function $l(M_p; \mathbf{Y})$ can be equivalently rewritten as

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

for any arbitrary probability density function q_i satisfying $\int_{\boldsymbol{\theta}_i} q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i = 1$. Since $P(Y_i \mid$

$M_p) = P(Y_i, \boldsymbol{\theta}_i \mid M_p)/P(\boldsymbol{\theta}_i \mid Y_i, M_p)$, then we can further write

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

where $KL\{q_i(\boldsymbol{\theta}_i) \| P(\boldsymbol{\theta}_i \mid Y_i, M_p)\} = \int_{\boldsymbol{\theta}_i} \log \frac{q_i(\boldsymbol{\theta}_i)}{P(\boldsymbol{\theta}_i | Y_i, M_p)} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$ is the Kullback-Leibler (KL) distance between the distributions $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$. The KL distance $KL\{q_i(\boldsymbol{\theta}_i) \| P(\boldsymbol{\theta}_i \mid Y_i, M_p)\} \geq 0$ with the equality holds if and only if $q_i(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i \mid Y_i, M_p)$. Therefore, we have a lower bound of the marginal likelihood as

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

and the equality holds when $q_i(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ for $i = 1, \ldots, N$.

The follow-up question is how to design the candidate distribution function $q_i(\boldsymbol{\theta}_i)$ that gives the best approximation of the marginal likelihood. From the above argument, the best choice is the unknown posterior distribution function $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$. Although this choice of $q_i(\boldsymbol{\theta}_i)$ is intractable, it provides a guideline to choose $q_i(\boldsymbol{\theta}_i)$ in the sense that a good choice of $q_i(\boldsymbol{\theta}_i)$ must approximate $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ well. The well-known EM algorithm follows this idea and can be interpreted as a maximization-maximization (MM) algorithm (Hunter & Lange, 2004) based on the above decomposition. In particular, the E-step chooses q_i to be

a distribution that minimizes the KL distance function, which corresponds to the estimated posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, \hat{M}_p)$ with $\hat{M}_p$ from the previous step estimates. The E-step then evaluates the expectation with respect to q_i's, i.e.,

$$\sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log P(Y_i, \boldsymbol{\theta}_i \mid M_p) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i, \tag{5}$$

which is equal to the lower bound in (4), except the additional constant term $-\sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log q_i(\boldsymbol{\theta}_i) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$ that does not depend on model parameters M_p. In the M-step, we maximize the above expectation term to estimate model parameters and this is equivalent to maximizing the lower bound in (4).

However, one challenge in the EM algorithm is to evaluate the expectation in (5) with respect to the posterior distribution of $\boldsymbol{\theta}_i$. In the MIRT model, it is known that this integral in (5) does not have an explicit form and in the literature, numerical approximation methods are often used, such as the Gauss–Hermite approximation, Monte Carlo expectation-maximization (McCulloch, 1997), and stochastic expectation-maximization (von Davier & Sinharay, 2010).

To avoid directly evaluating the posterior distribution of $\boldsymbol{\theta}_i$, the variational inference method uses alternative choices of the $q_i(\boldsymbol{\theta}_i)$'s to approximate the marginal likelihood function. The choices of $q_i(\boldsymbol{\theta}_i)$ not only approximate the posterior $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ well, but also are easy to compute and usually give closed form evaluations in the algorithm. In particular, from the MIRT literature, we know that as the number of items J becomes reasonably large, the posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ can be well approximated by a Gaussian distribution (Bishop, 2006). Motivated by this observation, we use the Gaussian approximation

procedure that chooses $q_i(\boldsymbol{\theta}_i)$ from a family of Gaussian distributions such that the KL distance between $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ is minimized. The estimation is then taken as a two-step iterative procedure. In the variational E-step, we choose $q_i(\boldsymbol{\theta}_i)$ by minimizing the KL distance between $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ and evaluate the expectation of the likelihood function with respect $q_i(\boldsymbol{\theta}_i)$, which is (5). In the M-step we update the unknown model parameters by maximizing the above expectation. The algorithm repeats the two steps until convergence. In the following sections, we present the detailed algorithm steps for the M2PL and M3PL models.

3 GVEM for the M2PL Model

In this section we present the GVEM algorithm for the M2PL model. Without loss of generality, we first focus on the ith subject's likelihood function due to the independence of different subjects' responses. The joint distribution function of $\boldsymbol{\theta}_i$ and Y_i is

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

The difficulty of handling the marginal distribution of Y_i mostly comes from the logistic sigmoid function, which makes the integration over $\boldsymbol{\theta}$ not in a closed form in the E-step (i.e., Eq. (5)).

To avoid dealing with intractable likelihood in E-step, we use a local variational method

initially proposed in the machine learning literature (Bishop, 2006; Jordan, Ghahramani, Jaakkola, & Saul, 1999), which finds bounds on functions over individual variables or groups of variables within a model instead of the full posterior distribution over all random variables. For notational simplicity, hereafter, we denote $x_{i,j} = b_j - \boldsymbol{\alpha}_i^\top \boldsymbol{\theta}_i$. Because of the concavity of the logistic sigmoid function $\log(1/(1+e^{-x_{i,j}}))$, by the local variational method, we have the following variational lower bound on the logistic sigmoid function,

$$\begin{aligned} \frac{e^{x_{i,j}}}{(1+e^{x_{i,j}})} &= \max_{\xi_{i,j}} \frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} \exp\left\{ \frac{(x_{i,j}-\xi_{i,j})}{2} - \eta(\xi_{i,j})(x_{i,j}^2 - \xi_{i,j}^2) \right\} \\ &\geq \frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} \exp\left\{ \frac{(x_{i,j}-\xi_{i,j})}{2} - \eta(\xi_{i,j})(x_{i,j}^2 - \xi_{i,j}^2) \right\}. \end{aligned} \tag{6}$$

where $\xi_{i,j}$ is a variational parameter that is introduced to approximate the objective function $e^{x_{i,j}}/(1+e^{x_{i,j}})$, and $\eta(\xi_{i,j}) = (2\xi_{i,j})^{-1}[e^{\xi_{i,j}}/(1+e^{\xi_{i,j}}) - 1/2]$. We then aim to estimate the parameter $\xi_{i,j}$ that achieves the equality of the above display. By introducing an additional variational parameter $\xi_{i,j}$, we successfully avoid the problem of estimating the intractable integral in the E-step. The values of $\xi_{i,j}$'s will be iteratively updated in the M-step.

Using the lower bound on the logistic sigmoid function, we obtain a closed-form lower bound for $\log P(Y_i, \boldsymbol{\theta}_i \mid \boldsymbol{A}, \boldsymbol{B})$ as follows

$$\begin{aligned} \log P(Y_i, \boldsymbol{\theta}_i \mid \boldsymbol{A}, \boldsymbol{B}) &\geq \sum_{j=1}^{J} \log \frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + \sum_{j=1}^{J} Y_{ij}(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + \sum_{j=1}^{J} \frac{(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \xi_{i,j})}{2} \\ &\quad - \sum_{j=1}^{J} \eta(\xi_{i,j})\{(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \xi_{i,j}^2\} + \log \phi(\boldsymbol{\theta}_i) \\ &=: l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}) \end{aligned}$$

where $\boldsymbol{\xi}_i = (\xi_{i,j}, j = 1, \ldots, J)^\top$.

The key step is to find the optimal variational distribution $q_i(\boldsymbol{\theta}_i)$, which we describe in detail in the next section.

3.1 Algorithm Details

Choice of q_i Conditional on the model parameters $\boldsymbol{A}, \boldsymbol{B}$ and the variational parameters $\xi_{i,j}$ for $i = 1, \ldots, N, j = 1, \ldots, J$, by the variational inference theory, it can be shown that the variational distributions $q_i(\boldsymbol{\theta}_i), i = 1, \ldots, N$ that minimize the KL divergence with the posterior distributions $P(\boldsymbol{\theta}_i | A, B), i = 1, \ldots, N$ take the following form:

$$\log q_i(\boldsymbol{\theta}_i) \quad \propto \quad \sum_{j=1}^{J} \Big(Y_{ij} - \frac{1}{2}\Big)\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \sum_{j=1}^{J} \eta(\xi_{i,j})(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \frac{\boldsymbol{\theta}_i^\top \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\theta}_i}{2}.$$

The standard nonlinear optimization technique is exploited to show that $q_i(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \mu_i, \Sigma_i)$ minimizes the KL divergence among all normal distributions where the mean and the covariance are

$$\mu_i = \Sigma_i \times \sum_{j=1}^{J} \Big\{2\eta(\xi_{i,j}) b_j + Y_{ij} - \frac{1}{2}\Big\}\boldsymbol{\alpha}_j^\top, \tag{7}$$

$$\Sigma_i^{-1} = \Sigma_{\boldsymbol{\theta}}^{-1} + 2\sum_{j=1}^{J} \eta(\xi_{i,j}) \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^\top. \tag{8}$$

With the variational densities $q_i(\boldsymbol{\theta}_i)$'s, we aim to estimate model parameters $\boldsymbol{\xi}_i$'s, $\boldsymbol{\alpha}_j$'s and b_j's by maximizing the lower bound of the marginal likelihood. Suppose we have $\boldsymbol{\xi}_i$'s from a previous step's estimation or the initial values, denoted by $\boldsymbol{\xi}_i^{(t)}$. Similarly, define $\boldsymbol{A}^{(t)} = \{\boldsymbol{\alpha}_j^{(t)}, j = 1, \ldots, J\}$, $\boldsymbol{B}^{(t)} = \{b_j^{(t)}, j = 1, \ldots, J\}$, $\Sigma_{\boldsymbol{\theta}}^{(t)}$, $\mu_i^{(t)}$ and $\Sigma_i^{(t)}$. The EM iteration

is presented below.

E-Step In E-step, we evaluate the closed-form lower bound of the expected log likelihood with respect to the variational distributions q_i's. With iteratively updated variational parameters $\mu_i^{(t)}$ and $\Sigma_i^{(t)}$, we easily evaluate the tth iteration's lower bound of the expected log-likelihood. Denote the tth iteration's variational density as $q_i^{(t)}(\boldsymbol{\theta}_i) = q_i(\boldsymbol{\theta}_i \mid \boldsymbol{\xi}_i^{(t)}, \boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \Sigma_{\boldsymbol{\theta}}^{(t)})$. Then, the tth iteration's lower bound can be derived as

$$
\begin{aligned}
E^{(t)}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}) := & \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}) \times q_i^{(t)}(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\
= & \sum_{i=1}^{N} \sum_{j=1}^{J} \Bigg(\log \frac{e^{\xi_{i,j}^{(t)}}}{(1+e^{\xi_{i,j}^{(t)}})} + (\frac{1}{2} - Y_{ij}) b_j^{(t)} + (Y_{ij} - \frac{1}{2}) \boldsymbol{\alpha}_j^{(t)\top} \mu_i^{(t)} - \frac{1}{2} \xi_{i,j}^{(t)} \\
& - \eta(\xi_{i,j}^{(t)}) \{ b_j^{(t)2} - 2 b_j^{(t)} \boldsymbol{\alpha}_j^{(t)\top} \mu_i^{(t)} + \boldsymbol{\alpha}_j^{(t)\top} [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top] \boldsymbol{\alpha}_j^{(t)} - \xi_{i,j}^{(t)2} \} \Bigg) \\
& + \frac{N}{2} \log |(\Sigma_{\boldsymbol{\theta}}^{(t)})^{-1}| - \sum_{i=1}^{N} \frac{1}{2} Tr((\Sigma_{\boldsymbol{\theta}}^{(t)})^{-1} [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]).
\end{aligned}
$$

M-Step In M-step, we maximize the estimated lower bound to update the model parameters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}, \Sigma_{\boldsymbol{\theta}})$. This is achieved by simply setting the derivative of the lower bound with respect to $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}, \Sigma_{\boldsymbol{\theta}})$ to be zero. As a result, it can be shown that each update of the model parameters are done in a closed form, which makes the proposed GVEM algorithm computationally efficient. The updating step is presented below. The most recently updated

copies of the parameters are used for each iterative update.

$$\boldsymbol{\alpha}_j = \frac{1}{2}\Big[\sum_{i=1}^{N}\eta(\xi_{i,j})\Sigma_i + \eta(\xi_{i,j})\mu_i\mu_i^\top\Big]^{-1}\sum_{i=1}^{N}\Big[\Big(Y_{ij} - \frac{1}{2} + 2b_j\eta(\xi_{i,j})\Big)\mu_i^\top\Big], \tag{9}$$

$$b_j = \frac{\sum_{i=1}^{N}\Big[(\frac{1}{2} - Y_{ij}) + 2\eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top\mu_i\Big]}{\sum_{i=1}^{N}2\eta(\xi_{i,j})}, \tag{10}$$

$$\xi_{i,j}^2 = b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top\mu_i + \boldsymbol{\alpha}_j^\top[\Sigma_i + \mu_i\mu_i^\top]\boldsymbol{\alpha}_j. \tag{11}$$

For the covariance matrix $\Sigma_{\boldsymbol{\theta}}$, in the exploratory analysis, we can keep $\Sigma_{\boldsymbol{\theta}} = I_K$ during the GVEM estimation and then later performed proper rotation; in the confirmatory analysis, we update $\Sigma_{\boldsymbol{\theta}}$ by

$$\Sigma_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}[\Sigma_i + \mu_i\mu_i^\top]. \tag{12}$$

Note that if the $\Sigma_{\boldsymbol{\theta}}$ is assumed to be the correlation matrix with diagonals being 1, then we need to standardize the estimated $\Sigma_{\boldsymbol{\theta}}$ to get correlation matrix. Detailed derivations regarding the above EM steps are given in the Supplementary Material.

In light of the above exposition, the GVEM algorithm for M2PL can be summarized as follows.

Algorithm 1 GV-EM algorithm

1: Initialize $M_p^{(0)} = \{\boldsymbol{A}_0, \boldsymbol{B}_0\}, \boldsymbol{\xi}^{(0)}$.
2: **repeat**
3: E step : For step $t \geq 1$, update $\mu_i^{(t)}$ and $\Sigma_i^{(t)}$ according to closed-form equations (7) and (8).
4: M step : Further update $M_p^{(t)}$ and $\boldsymbol{\xi}^{(t)}$ according to closed-form equations (9), (10), and (11), iteratively. Fix $\Sigma_{\boldsymbol{\theta}}^{(t)} = I_K$ in the exploratory analysis or update $\Sigma_{\boldsymbol{\theta}}^{(t)}$ according to (12) in the confirmatory analysis.
5: **until** convergence

Remark 1 *The algorithm complexity increases with the sample size N, which makes the algorithm computationally inefficient for large data sets. Thus, we can stochastically optimize the EM algorithm by sub-sampling the data to form noisy estimates of the variational lower bound and model parameters. Please refer to Section 4.2 for detailed explanation of the stochastic GVEM.*

Remark 2 *Under the IRT framework, test dimensionality is one of the major issues explored in order to validate the design of a test and help practitioners with test development. As a byproduct of the algorithm, we can empirically estimate the number of latent dimensions from data. Specifically, the information criteria such as AIC or BIC can be used to compare the model fit with varying number of dimensions. Because we approximate the true log-likelihood by its lower bound in GVEM, the information criteria also need to be modified by replacing the true log-likelihood with the variational lower bound, resulting in the following modified AIC and BIC, denoted as $AIC^\star$ and $BIC^\star$. The approximated information criteria are as follows, $AIC^\star = 2(\|\boldsymbol{A}\|_0 + \|\boldsymbol{B}\|_0 + \|\Sigma_{\boldsymbol{\theta}}\|_0) - 2E(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\xi}})$ and $BIC^\star = ln(N)(\|\boldsymbol{A}\|_0 + \|\boldsymbol{B}\|_0 + \|\Sigma_{\boldsymbol{\theta}}\|_0) - 2E(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\xi}})$ where $E(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\xi}})$ is the estimated variational lower bound and $\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\xi}}$ are the final estimates from GVEM estimation procedure. The notation $\|\boldsymbol{A}\|_0$ of matrix $\boldsymbol{A}$ denotes the zero norm of the matrix $\boldsymbol{A}$, which is simply the number of non-zero entries of $\boldsymbol{A}$. The advantage of using GVEM to estimate test dimensionality is that it is computationally more efficient especially under high dimensional data and more complex model. This procedure can be easily applied in both the 2PL and the 3PL models. Please see the simulation study for more discussions.*

3.2 Theoretical Properties

In this section, we establish theoretical bounds on the estimation of the model parameters under the high-dimensional setting where both N and J go to infinity. The dimension of latent traits, K, is assumed known for this analysis and thus fixed. As defined in Section 2, $\boldsymbol{A} = [\alpha_{jk}]_{J\times K}$ denotes a matrix of factor loadings. Additionally, let $\Theta = [\theta_{ij}]_{N\times K}$ denote a matrix of random variables following $q_i(\boldsymbol{\theta}_i)$ and let $\hat{\Theta} = [\hat{\theta}_{ij}]_{N\times K}$ denote a matrix of estimated latent abilities from data. Define $E_{\hat{\boldsymbol{\theta}}\sim\hat{q}}$ to be the expectation with respect to the estimated variational densities $\{\hat{q}_i(\hat{\boldsymbol{\theta}}_i) \sim N(\hat{\mu}_i, \hat{\Sigma}_i) : i = 1, \ldots, N\}$ from data. Lastly, a superscript $*$ denote a true parameter. For example, $\boldsymbol{\theta}_i^*$ denotes the i^{th} person's true latent ability, which is a deterministic realization from its population distribution. We assume that the true parameters Θ^* and $\boldsymbol{A}^*$ satisfy

(A1). $\|\boldsymbol{\theta}_i^*\|^2 \leq C$ and $\|\boldsymbol{\alpha}_j^*\|^2 \leq C$ for all i, j for some positive constant C

Theorem 1 derives the bound on the expected Frobenius norm of the error, $\|\hat{\Theta}\hat{\boldsymbol{A}}^\top - \Theta^*(\boldsymbol{A}^*)^\top\|_F$, where $\|M\|_F = \sqrt{\sum_{i,j} M_{ij}^2}$ denotes the Frobenius norm of a matrix M.

Theorem 1 *Suppose that condition **(A1)** is satisfied for the true parameters Θ^* and $\boldsymbol{A}^*$. With optimally estimated variational densities $\hat{q}_i$ from data and estimated parameter matrix $\hat{\boldsymbol{A}}$ that maximizes the variational lower bound, there exists absolute constants C_1 and C_2 such that*

$$\frac{1}{NJ} E_{\hat{\boldsymbol{\theta}}\sim\hat{q}}[\|\hat{\Theta}\hat{\boldsymbol{A}}^\top - \Theta^*(\boldsymbol{A}^*)^\top\|_F] \leq C_2 C e^C \sqrt{\frac{J+N}{JN}} \sqrt{1 + \frac{\log(N+J)}{N+J}}$$

is satisfied with probability $1 - C_1/(N+J)$.

The proof of Theorem 1 can be found in the Supplementary Material.

Remark 3 *Theorem 1 states that the expected estimation error measured by Frobenius norm goes to 0 as both $N \to \infty$ and $J \to \infty$. The proof of Theorem 1 follows a similar argument from Davenport, Plan, Van Den Berg, and Wootters (2014) and Theorem 1 in Chen et al. (2019). However, the previous work by Chen et al. (2019) treats $\boldsymbol{\theta}_i$ as fixed effects while this work follows the conventional MIRT model setting with $\boldsymbol{\theta}_i$ random effects and following a normal population distribution.*

Remark 4 *The Gaussian family as the candidate choice of q is reasonable according to Laplace approximation of the posterior distribution $P(\boldsymbol{\theta}_i|Y_i)$. The Laplace approximation of $P(\boldsymbol{\theta}_i|Y_i)$ is a normal distribution with MLE $\hat{\boldsymbol{\theta}}_i$ as mean and inverse of observed Fisher information $I^{-1}(\hat{\boldsymbol{\theta}}_i)$ as variance. Denote $\boldsymbol{\theta}_i^*$ as the true parameter. By Bernstein-von Mises Theorem, since $P(Y_i \mid \boldsymbol{\theta}_i), i = 1, \ldots, N$ have same support and $\boldsymbol{\theta}_i \to \log P(Y_i \mid \boldsymbol{\theta}_i)$ is twice continuously differentiable, then $\hat{\boldsymbol{\theta}}_i \to \boldsymbol{\theta}_i^*$ almost surely and the Laplace approximated distribution $N(\hat{\boldsymbol{\theta}}_i, I^{-1}(\hat{\boldsymbol{\theta}}_i))$ converges in distribution to the true limiting normal distribution $N(\boldsymbol{\theta}_i^*, I^{-1}(\boldsymbol{\theta}_i^*))$ as $J \to \infty$ where $I^{-1}(\boldsymbol{\theta}_i^*)$ is the inverse of expected Fisher information. This supports our choice of variational density q_i as a multivariate Gaussian distribution provides an asymptotically good approximation for the true posterior distribution of $\boldsymbol{\theta}$.*

Remark 5 *Compared with the existing stochastic estimation algorithms, such as the Metropolis-Hastings Robbins-Monro algorithm and the stochastic EM algorithm, the proposed estimation method has the advantage that each of the estimation iterations has simple closed-form update and it does not involve the stochastic samplings from some intermediate posterior distributions as in the current stochastic estimation algorithms. As discussed in Remark 4, even*

though variational distributions are used to approximate the posterior distributions in our method, the normal approximation is asymptotically valid. Simulation studies in Section 5 further illustrate this. Moreover, the above variational EM development can be easily generalized to the M3PL model and can also be naturally combined with the idea of the stochastic EM, as illustrated in the next section.

4 GVEM for the M3PL Model

Derivation of the variational lower bound is trickier in the M3PL function since the cancellation of log and exponential function, which was essential in simplifying the variational lower bound in M2PL, is impossible due to the addition of a guessing parameter. To solve this problem, we introduce another latent variable, Z_{ij} which is an indicator function of whether ith individual answered jth item based on their latent abilities or guessed it correctly (von Davier, 2009). We define $Z_{ij} = 1$ if ith individual solved item j based on his or her latent ability, and $Z_{ij} = 0$ if he or she guessed item j correctly. Notice here that for the case of $Z_{ij} = 1$, Y_{ij} can be either 0 or 1. However, when $Z_{ij} = 0$, Y_{ij} has to be 1 by the definition of Z_{ij}. Hence, $\{Y_{ij} = 0, Z_{ij} = 0\}$ cannot occur.

Proposition 1 *Given the two latent variables $\boldsymbol{\theta}_i$ and Z_{ij}, then $P(Y_{ij} \mid \boldsymbol{\theta}_i)$ under the follow-*

ing hierarchical model is equivalent to (2) *of the 3PL model.*

$$
\begin{aligned}
Z_{ij} &\sim Bernoulli(1 - c_j), \\
Y_{ij} \mid \boldsymbol{\theta}_i, Z_{ij} = 1 &\sim Bernoulli(\Big[\frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}\Big]), \\
Y_{ij} \mid \boldsymbol{\theta}_i, Z_{ij} = 0 &\sim Bernoulli(I(Y_{ij} = 1)).
\end{aligned}
$$

The distribution of observation Y_{ij} given latent variables $\boldsymbol{\theta}_i$ and Z_{ij} is then

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

Without loss of generality we first focus on the ith subject's likelihood function due to the independence of different subjects. 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 is

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

Following the result from Proposition 1, the hierarchical formulation of the 3PL model

with the new latent variable Z_{ij} could be used to derive the GVEM algorithm for the 3PL model. Please refer to the Supplementary Material for the proof of Proposition 1. Similar data augmentation scheme was proposed in Albert (1992) in the Bayesian framework.

In this section, we derive the optimal choices of the variational densities for the latent variables Z_{ij} and $\boldsymbol{\theta}_i$. The approach is similar to that of the 2PL model. For any arbitrary density functions q_i and r_{ij} of the latent variables $\boldsymbol{\theta}_i$ and Z_{ij}, the following equation always holds

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

where $r_i(\boldsymbol{Z}_i) = \prod_{j=1}^{J} r_{ij}(Z_{ij})$.

Note that $P(Y_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) = P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) / P(\boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid Y_i, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$. We can write

$$\begin{aligned}
\log P(Y_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &= \int_{\boldsymbol{\theta}_i} \sum_{\boldsymbol{Z}_i} \log \frac{P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})}{P(\boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid Y_i, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})} \times q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) d\boldsymbol{\theta}_i \\
&= \int_{\boldsymbol{\theta}_i} \sum_{\boldsymbol{Z}_i} \log \frac{P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})}{q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i)} \times q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) d\boldsymbol{\theta}_i \\
&\quad + KL\{q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) \| P(\boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid Y_i, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})\}.
\end{aligned}$$

Since the KL distance is ≥ 0 by definition, we get a lower bound on the marginal likelihood similarly as in the 2PL model.

$$\begin{aligned}
\log P(Y_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) &\geq \int_{\boldsymbol{\theta}_i} \sum_{\boldsymbol{Z}_i} \log P(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \times q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) d\boldsymbol{\theta}_i \qquad (13) \\
&\quad - \int_{\boldsymbol{\theta}_i} \sum_{\boldsymbol{Z}_i} \log \left(q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) \right) \times q_i(\boldsymbol{\theta}_i) r_i(\boldsymbol{Z}_i) d\boldsymbol{\theta}_i \qquad (14)
\end{aligned}$$

Since (14) doesn't depend on parameters $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, we focus on (13) for the derivation of the lower bound. Again, the ith subject's likelihood function is

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

Using the same variational lower bound (6) on the logistic sigmoid function as in the 2PL model, we show

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

Recall that if $Y_{ij} = 0$, then we always have $Z_{ij} = 1$ by the design of our model. In other words, $\{Y_{ij}, Z_{ij}\} = \{0, 0\}$ cannot occur. To accommodate this constraint, we replace Z_{ij} by $Z'_{ij} = 1 - Y_{ij} + Z_{ij} Y_{ij}$ so that $Z'_{ij} = Z_{ij}$ if $Y_{ij} = 1$ and $Z'_{ij} = 1$ if $Y_{ij} = 0$. This makes sure that the case of $\{Y_{ij}, Z_{ij}\} = \{0, 0\}$ is not included as a possible scenario during the estimation

procedure. By this substitution, we have

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

where $\log p(Z'_{ij}) = (1 - Y_{ij} + Z_{ij}Y_{ij}) \log(1 - c_j) + Y_{ij}(1 - Z_{ij}) \log(c_j)$.

With variational distributions q_i's and r_i's, we have the following expression for the variational lower bound of the marginal likelihood, which is an expectation of the joint distribution with respect to q_i's and r_i's, i.e.,

$$
E^{(t)}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}) := \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \left[\sum_{\boldsymbol{Z}_i} l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \times r_i^{(t)}(\boldsymbol{Z}_i) \right] \times q_i^{(t)}(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \quad (15)
$$

Appropriate choices of the variational distributions will lead to a closed form expression of the lower bound expressed in (15). As in the 2PL model, we choose the variational distributions for each latent variable by finding a distribution that best approximates the posterior distribution of each latent variable.

4.1 Algorithm Details

Choice of q_i Let E_r denote the expectation with respect to. the variational densities of Z_{ij}'s, i.e. $r_{ij}(Z_{ij})$'s. We can write

$$
\begin{aligned}
& E_r(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}) := \sum_{i=1}^{N} \sum_{Z_{ij}} l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \times r_{ij}(Z_{ij}) \\
= & \sum_{i=1}^{N} \Bigg[\sum_{j=1}^{J} (1 - Y_{ij} + E_r[Z_{ij}]Y_{ij}) \log \frac{e^{\xi_{i,j}}}{(1 + e^{\xi_{i,j}})} + \sum_{j=1}^{J} (1 - Y_{ij} + E_r[Z_{ij}]Y_{ij}) Y_{ij} (\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) \\
& + \sum_{j=1}^{J} (1 - Y_{ij} + E_r[Z_{ij}]Y_{ij}) \frac{1}{2} (b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \xi_{i,j}) \\
& - \sum_{j=1}^{J} (1 - Y_{ij} + E_r[Z_{ij}]Y_{ij}) \eta(\xi_{i,j}) \{ (b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \xi_{i,j}^2 \} \\
& + \sum_{j=1}^{J} \{ Y_{ij} (1 - E_r[Z_{ij}]) \log I(Y_{ij} = 1) \} + \log \phi(\boldsymbol{\theta}_i) + \sum_{j=1}^{J} E_r[\log p(Z'_{ij})] \Bigg]
\end{aligned}
$$

Conditional on the model parameters $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and the variational parameters $\boldsymbol{\xi}_i$ where $i = 1, \ldots, N$, by the variational inference theory, we can show that the variational distributions $q_i(\boldsymbol{\theta}_i), i = 1, \ldots, N$ that minimize the distances between them and the posterior distributions take the following form;

$$
\begin{aligned}
\log q_i(\boldsymbol{\theta}_i) \quad \propto \quad & \sum_{j=1}^{J} (1 - Y_{ij} + E_r(Z_{ij})Y_{ij}) \Big(Y_{ij} - \frac{1}{2} \Big) \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i \\
& - \sum_{j=1}^{J} (1 - Y_{ij} + E_r(Z_{ij})Y_{ij})) \eta(\xi_{i,j}) (b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \frac{1}{2} \boldsymbol{\theta}_i^\top \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\theta}_i.
\end{aligned}
$$

The above likelihood function implies that $q_i(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \mu_i, \Sigma_i)$ where the mean and covariance are

$$\mu_i = \Sigma_i \times \sum_{j=1}^{J} \Big\{ 2\eta(\xi_{i,j}) b_j + Y_{ij} - \frac{1}{2} \Big\} (1 - Y_{ij} + E_r(Z_{ij}) Y_{ij}) \boldsymbol{\alpha}_j^\top, \tag{16}$$

$$\Sigma_i^{-1} = \Sigma_{\boldsymbol{\theta}}^{-1} + 2 \sum_{j=1}^{J} (1 - Y_{ij} + E_r(Z_{ij}) Y_{ij}) \eta(\xi_{i,j}) \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^\top. \tag{17}$$

Choice of r_{ij} We follow the similar steps as q_i. That is, we take the expectation of the lower bound $l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ with respect to the variational density of $\boldsymbol{\theta}_i$, $q_i(\boldsymbol{\theta}_i)$ and derive the variational distributions for $Z_{ij}, i = 1, \ldots, N, j = 1, \ldots, J$. The variational distribution minimizes the distances between them and the posterior distributions of Z_{ij} given model parameters $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and the variational parameters $\boldsymbol{\xi}_i$.

Let E_q denote the expectation with respect to. the variational densities q_i's and E_{q_i} denote the expectation with respect to q_i. Taking expectation of the lower bound $l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ with respect to $q_i(\boldsymbol{\theta}_i)$, we have

$$\begin{aligned}
& E_q(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi}) \\
= & \sum_{i=1}^{N} \Bigg[\sum_{j=1}^{J} (1 - Y_{ij} + Z_{ij} Y_{ij}) \log \frac{e^{\xi_{i,j}}}{(1 + e^{\xi_{i,j}})} + \sum_{j=1}^{J} (1 - Y_{ij} + Z_{ij} Y_{ij}) Y_{ij} (\boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i] - b_j) \\
& + \sum_{j=1}^{J} (1 - Y_{ij} + Z_{ij} Y_{ij}) \frac{1}{2} (b_j - \boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i] - \xi_{i,j}) \\
& - \sum_{j=1}^{J} (1 - Y_{ij} + Z_{ij} Y_{ij}) \eta(\xi_{i,j}) \{ E_{q_i}[(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2] - \xi_{i,j}^2 \} \\
& + \sum_{j=1}^{J} \{ Y_{ij} (1 - Z_{ij}) \log I(Y_{ij} = 1) \} + E_{q_i}[\log \phi(\boldsymbol{\theta}_i)] + \sum_{j=1}^{J} \log p(Z'_{ij}) \Bigg]
\end{aligned} \tag{18}$$

This implies that the variational distributions $r_{ij}(Z_{ij})$ are

$$\begin{aligned}\log r_{ij}(Z_{ij}) \propto\ & Z_{ij}Y_{ij}\Big[\log\frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})}+Y_{ij}(\boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i]-b_j)+\frac{1}{2}(b_j-\boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i]-\xi_{i,j})\\ & -\eta(\xi_{i,j})\{E_{q_i}[(b_j-\boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i)^2]-\xi_{i,j}^2\}+\log(1-c_j)\Big]\\ & +Y_{ij}(1-Z_{ij})\Big[\log I(Y_{ij}=1)+\log(c_j)\Big].\end{aligned}$$

Thus, $r_{ij}(Z_{ij})\sim Bernoulli(s_{ij})$ where $s_{ij}=1$ if $Y_{ij}=0$ and

$$\begin{aligned}s_{ij}^{-1}= & \ 1+\frac{c_j}{1-c_j}\frac{1+e^{\xi_{i,j}}}{e^{\xi_{i,j}}}\exp\Big\{-Y_{ij}(\boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i]-b_j)+\\ & \frac{1}{2}(b_j-\boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i]-\xi_{i,j})-\eta(\xi_{i,j})\{E_{q_i}[(b_j-\boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i)^2]-\xi_{i,j}^2\}\Big\}\end{aligned}\tag{19}$$

if $Y_{ij}=1$ where $E_{q_i}[\boldsymbol{\theta}_i]=\mu_i$ and $E_{q_i}[(b_j-\boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i)^2]=b_j^2-2b_j\boldsymbol{\alpha}_j^\top\mu_i+\boldsymbol{\alpha}_j^\top[\Sigma_i+\mu_i\mu_i^\top]\boldsymbol{\alpha}_j$.

With the chosen q_i's and r_{ij}'s, we aim to estimate model parameters $\boldsymbol{\xi}$, $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$, by maximizing the variational lower bound of the marginal likelihood, i.e., (15). The EM steps for 3PL model follow the same procedure as in 2PL case.

E-Step In every E step, we choose the optimal variational distributions q_i's and r_{ij}'s, which is equivalent to estimating variational parameters μ_i, Σ_i, and s_{ij} for every i and j. With iteratively updated variational parameters, (i.e. $\mu_i^{(t)}$, $\Sigma_i^{(t)}$, and $s_{ij}^{(t)}$) and most recent updates of model parameters (i.e. $M_p^{(t)}=\{\boldsymbol{A}^{(t)},\boldsymbol{B}^{(t)},\boldsymbol{C}^{(t)}\}$), we derive a closed form expression of

variational lower bound at tth step as follows;

$$
\begin{aligned}
& E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi}) \\
= & \sum_{i=1}^{N}\sum_{j=1}^{J}(1-Y_{ij}+s_{ij}^{(t)}Y_{ij})\Big(\log\frac{e^{\xi_{i,j}^{(t)}}}{(1+e^{\xi_{i,j}^{(t)}})}+(\frac{1}{2}-Y_{ij})b_j^{(t)}+(Y_{ij}-\frac{1}{2})\boldsymbol{\alpha}_j^{(t)\top}\mu_i^{(t)} \\
& -\frac{1}{2}\xi_{i,j}^{(t)}-\eta(\xi_{i,j}^{(t)})\{b_j^{(t)2}-2b_j^{(t)}\boldsymbol{\alpha}_j^{(t)\top}\mu_i^{(t)}+(\boldsymbol{\alpha}_j^{(t)})^\top[\Sigma_i^{(t)}+(\mu_i^{(t)})(\mu_i^{(t)})^\top]\boldsymbol{\alpha}_j^{(t)}-\xi_{i,j}^{(t)2}\}\Big) \\
& +\sum_{i=1}^{N}\sum_{j=1}^{J}Y_{ij}(1-s_{ij}^{(t)})\log I(Y_{ij}=1)-\sum_{i=1}^{N}\frac{1}{2}Tr((\Sigma_{\boldsymbol{\theta}}^{(t)})^{-1}[\Sigma_i^{(t)}+(\mu_i^{(t)})(\mu_i^{(t)})^\top]) \\
& +\frac{N}{2}\log|(\Sigma_{\boldsymbol{\theta}}^{(t)})^{-1}|+\sum_{i=1}^{N}\sum_{j=1}^{J}\{(1-Y_{ij}+s_{ij}^{(t)}Y_{ij})\log(1-c_j^{(t)})+Y_{ij}(1-s_{ij}^{(t)})\log(c_j^{(t)})\}.
\end{aligned}
$$

M-Step In this step, we again maximize the $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ to update the parameters $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$. This is achieved by setting the derivative of $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ with respect to $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ to be zero. Since we have a closed-form expression of the lower bound, updates of the model parameters are also in closed-form. Detailed derivation is provided in the Supplementary Material.

For $\boldsymbol{\xi}$ and $\Sigma_{\boldsymbol{\theta}}$, the update is the same as in 2PL model. For other parameters, we derive the updating rule by taking derivative of the variational lower bound $E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ derived

in E step. As a result, we have the following updating rule for $\boldsymbol{\alpha}_j$, b_j and c_j;

$$\begin{aligned}
\boldsymbol{\alpha}_j &= \frac{1}{2}\Big[\sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})\eta(\xi_{i,j})[\Sigma_i+\mu_i\mu_i^\top]\Big]^{-1} \\
&\qquad \times \sum_{i=1}^{N}\Big[(1-Y_{ij}+s_{ij}Y_{ij})\Big(Y_{ij}-\frac{1}{2}+2b_j\eta(\xi_{i,j})\Big)\mu_i^\top\Big], \qquad (20)\\
b_j &= \frac{\sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})\Big[(\frac{1}{2}-Y_{ij})+2\eta(\xi_{i,j})\boldsymbol{\alpha}_j^{(t)\top}\mu_i\Big]}{\sum_{i=1}^{N}2(1-Y_{ij}+s_{ij}Y_{ij})\eta(\xi_{i,j})}, \qquad (21)\\
c_j &= \frac{\sum_{i=1}^{N}(Y_{ij}-s_{ij}Y_{ij})}{\sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})+\sum_{i=1}^{N}(Y_{ij}-s_{ij}Y_{ij})} = \frac{1}{N}\sum_{i=1}^{N}Y_{ij}(1-s_{ij}). \qquad (22)
\end{aligned}$$

The Algorithm 2 summarizes the EM steps for GVEM algorithm in M3PL.

Algorithm 2 GV-EM algorithm for M3PL

1: Initialize $M_p^{(0)} = \{\boldsymbol{A}_0, \boldsymbol{B}_0, \boldsymbol{C}_0\}, \boldsymbol{\xi}^{(0)}$.
2: **repeat**
3: E step : For step $t \geq 1$, update variational parameters $\mu_i^{(t+1)}$, $\Sigma_i^{(t+1)}$, and $s_{ij}^{(t+1)}$ according to closed-form equations (16), (17), and (19).
4: M step : Further update $M_p^{(t+1)}$ according to closed-form equations (20), (21), and (22) iteratively. Update $\boldsymbol{\xi}^{(t+1)}$ and $\Sigma_{\boldsymbol{\theta}}^{(t+1)}$ same as in M2PL.
5: **until** convergence

Remark 6 *The theoretical property of the M3PL is more challenging to derive rigorously due to the addition of the guessing parameters c_j's. From Theorem 2 in Davenport et al. (2014) we can show that the Hellinger distance of error between estimated probability distributions and the true probability distributions is bounded above. For this discussion, we define Hellinger distance for probability distributions and matrices. Hellinger distance for two scalars $p, q \in [0,1]$ is defined as $d_H^2(p,q) := (\sqrt{p}-\sqrt{q})^2 + (\sqrt{1-p}-\sqrt{1-q})^2$. Following Davenport et al. (2014), we also allow the Hellinger distance to act on matrices by averaging Hellinger distances over their entries. For matrices $P, Q \in [0,1]^{d_1 \times d_2}$, we de-*

fine $d_H^2(P,Q) = \frac{1}{d_1 d_2}\sum_{i,j} d_H^2(P_{ij},Q_{ij})$. *Let* $M = [M_{ij}]_{N\times J}$ *be the matrix with entries* M_{ij} *satisfying* $\frac{\exp(M_{ij})}{1+\exp(M_{ij})} = c_j + (1-c_j)\frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1+\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}$. *Let* $P(\boldsymbol{Y}|M)$ *be a matrix of probability distributions* $P(Y_{ij}|M_{ij})$*'s where* M_{ij} *denotes a collection of model parameters* $\boldsymbol{\alpha}_{ij}, b_j, c_j$. *Again,* M^* *denotes a matrix of true parameters and* $\hat{M}$ *denotes estimated model parameters. Then by Theorem 2 of Davenport et al. (2014)*

$$d_H^2(P(\boldsymbol{Y}|\hat{M}), P(\boldsymbol{Y}|M^*)) \leq C_2 C\sqrt{\frac{K(N+J)}{NJ}}\sqrt{1+\frac{(N+J)\log(NJ)}{NJ}}$$

with probability $1-\frac{C_1}{N+J}$ *for absolute constants* C_1 *and* C_2. *Hence, the Hellinger distance between estimated probability distribution and true probability distribution goes to* 0 *as both* $N\to\infty$ *and* $J\to\infty$. *However, the consistency result for model parameter* $\{\boldsymbol{\alpha}_j, b_j, c_j : j = 1,\ldots,J\}$ *in M3PL is more challenging to derive and thus left for the future research.*

4.2 Stochastic Optimization of GVEM

In M3PL, the proposed GVEM algorithm may become computationally inefficient as sample size increases because of the additional variational parameters and model parameters to estimate compared to M2PL. Especially in the E step, variational parameters (i.e. $\mu_i, \Sigma_i, \xi_{i,j}, s_{ij}$) need to be optimized for every data points $i = 1,\ldots,N$. Thus, the computational burden increases with larger sample size N. To improve the computational efficiency of the GVEM algorithm, we can stochastically optimize the variational approximation in the E step (Hoffman, Blei, Wang, & Paisley, 2013). That is, at each iteration of the E step, we subsample the data to form noisy estimate of the variational lower bound and iteratively update the estimate with a decreasing step size. Then M step in Algorithm 2 follows using this

stochastically estimated variational lower bound. The stochastic optimization only affects the E step, thus with minor changes to the original GVEM algorithm we can stochastically optimize the algorithm for M3PL. The noisy estimates of the variational lower bound are cheaper to compute as it only requires small subset of the data at each iteration. Also, for complicated models like M3PL, following such noisy estimates can also help the algorithm to escape local optima of complex objective functions. Specifically, the stochastic EM steps can be summarized as follows.

Stochastic E step For step $t \geq 1$, choose a subset of data S_t with desired size. Choose a decreasing step size ϵ_t. Update $\mu_i^{(t)}$, $\Sigma_i^{(t)}$, $\boldsymbol{\xi}_i^{(t)}$ and $s_{ij}^{(t)}$ for data point $i \in S_t$ only, according to closed-form equations (16) and (17). Since we only update variational parameters for $i \in S_t$, the algorithm is computationally more efficient than GVEM approach without stochastic optimization, especially when the size of the subset S_t is chosen to be small.

With updated variational parameters partially for $i \in S_t$, calculate noisy estimate of tth iteration's expected variational lower bound $\hat{Q}_t$ as follows;

$$\hat{Q}_t = \sum_{i \in S_t} \int_{\boldsymbol{\theta}_i} \left[\sum_{\boldsymbol{Z}_i} l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}) \times r_i^{(t)}(\boldsymbol{Z}_i) \right] \times q_i^{(t)}(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$$

Then we obtain a stochastic approximation of the variational lower bound by a weighted average of previous and current step's noisy estimates of the lower bound, i.e. $(1-\epsilon_t)\hat{Q}_{t-1} + \epsilon_t \hat{Q}_t$.

M step Once E step is done, we follow the previous M step. That is, estimate $\hat{\boldsymbol{A}}^{(t)}$, $\hat{\boldsymbol{B}}^{(t)}$, $\hat{\boldsymbol{C}}^{(t)}$, and $\hat{\Sigma}_{\boldsymbol{\theta}}^{(t)}$ that maximizes the stochastic approximation of the variational lower bound.

Notice that this stochastic optimization idea is different from the stochastic component in the stochastic EM (StEM) algorithm (Nielsen, 2000). In the StEM algorithm, random samples of the unobserved latent variables $\boldsymbol{\theta}_i$ are drawn from the conditional distribution of $\boldsymbol{\theta}_i$ given observed variable Y_i, and these random samples are used to approximate the otherwise intractable expectation in the E step. In our algorithm, the stochastic component instead refers to the random sub-sampling of the observed data $\{Y_{ij}, i = 1, \ldots, N\}$ to form a noisy approximation of the variational lower bound $E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{\xi})$ in E step.

In theory, if a sequence of step sizes satisfies the conditions such that

$$\sum \epsilon_t = \infty \text{ and } \sum \epsilon_t^2 < \infty, \tag{23}$$

which results in a sequence of decreasing step sizes, the algorithms provably converge to an optimum (Robbins & Monro, 1951). Following the approach in Hoffman et al. (2013), we set the tth step size as $\epsilon_t = (t + \tau)^{-r}$ where *forget rate* $r \in (0.5, 1]$ and *delay* $\tau \geq 0$. The *forget rate* controls how quickly old information is forgotten and the *delay* down-weights early iterations to decrease the effect of the earlier noisy estimations. This step size obviously satisfies the conditions (23). Thus the iterative stochastic optimization of E step converges to a local optimum of the variational lower bound. In simulation, we fix the delay to be one and try various forget rates as different values of delay didn't play a big role for our model. Although in theory the stochastic optimization of GVEM converges to a stationary point for any valid forget rate r, the quality and speed of the convergence may depend on r in practice.

5 Simulations

5.1 Design

A series of simulation studies were conducted to evaluate the performance of the proposed GVEM algorithm in comparison to the Metropolis-Hastings Robbins-Monro (MHRM) algorithm implemented in the R package, 'mirt' (Chalmers, 2012). The Metropolis-Hastings sampler is used to draw missing data (which is $\boldsymbol{\theta}$ in MIRT) in the stochastic imputation step of the MHRM algorithm (Cai, 2008, 2010a). In the 'mirt' package, MHcand is a vector of values used to tune the MH sampler, with larger values yielding lower acceptance rate. By default, these values are determined internally and adjusted on-the-fly, attempting to tune the acceptance of the draws to be between .1 and .4. In addition, the default number of Metropolis-Hastings draws at each iteration is 5, which is considered sufficient by Cai (2010a). Only the exploratory item factor analysis will be presented since it is a computationally more challenging scenario than the confirmatory analysis. That is, in the confirmatory analysis, many of the item loading parameters (or discrimination parameters) are constrained to 0 based on the pre-specified item factor loading structure. Hence, the update equation for $\boldsymbol{\alpha}$ (i.e., (9) for the 2PL model and (20) for the 3PL model) only needs minimum updates to reflect the constraints specified in the factor loading structure. In the exploratory analysis, we do not assume any constraint on the item discrimination parameter $\boldsymbol{A}$ while fix $\Sigma_{\boldsymbol{\theta}} = I_K$ during the estimation. A post-hoc rotation can then follow to rotate the factors and allow them to be correlated. The best-known rotation methods available in most commercial software packages are varimax (Kaiser, 1958) in orthogonal rotation or promax (Hendrickson & White, 1964) in oblique rotation. Other popular methods include,

for instance, the CF-Quartimax rotation (Browne, 2001). In the simulations studies, the promax rotation was used such that the factors were allowed to be correlated. Both the M2PL and M3PL were considered in the simulation studies. The number of dimensions was fixed at 3 and test length was fixed at 45.

Additionally, we compared the performance of GVEM to the joint maximum likelihood (JML) estimator given that the JML estimator is also shown to be consistent under the same high-dimensional setting presented in Theorem 1 and efficient (Chen et al., 2019). The JML estimator was computed using the default settings in the R package 'mirtjml' implemented by Chen et al. (2019). Since Chen et al. (2019) did not study M3PL, here we only compared the performances for M2PL.

The manipulated conditions include: (i) multidimensional structure, i.e. between-item multidimensionality and within-item multidimensionality; (ii) correlations among the latent traits, and (iii) sample size. In particular, for the between-item multidimensional structure, there were 15 items loaded onto each factor; whereas for the within-item multidimensional structure, about one third of the items were loaded onto one, two, and three factors respectively. In all cases, item discrimination parameters were simulated from $Unif(1, 2)$ distribution, and difficulty parameter b_j was simulated from the standard normal distribution. For the M3PL model, the true guessing parameters were fixed at 0.2 for all test items. The latent traits $\boldsymbol{\theta}_i$ were generated from multivariate normal distribution, $N(0, \Sigma_{\boldsymbol{\theta}})$, where $\Sigma_{\boldsymbol{\theta}}$ is a covariance matrix whose diagonal elements were 1 and the off-diagonals were drawn from Uniform distribution. For the high correlation condition, the correlations were drawn from $Unif(0.5, 0.7)$ and for the low correlation condition, they were drawn from $Unif(0.1, 0.3)$. Sample size was set at either 200 or 500.

The convergence criterion for the GVEM algorithm is $\|M_p\|_2 < 0.0001$, where $\|M_p\|_2$ refers to the L_2 norm of all model parameters. The number of Markov chain samples drawn in the MHRM algorithm is by default 5000 in the R package 'mirt'. Lastly, the JML method adopts sequential change in log-likelihood as the convergence criterion and the tolerance of convergence is by default 5 in the R package 'mirtjml'. 100 replications were conducted for each condition. Evaluation criteria include the average bias, root mean squared error (RMSE), and computation time of both methods. The parameter recovery for $\Sigma_{\boldsymbol{\theta}}$ is calculated by taking differences between each entries of the true $\Sigma_{\boldsymbol{\theta}}$ and estimated $\hat{\Sigma}_{\boldsymbol{\theta}}$. Both bias and RMSE were obtained for each model parameter across all items within a condition first and then averaged over 100 replications.

5.2 Results for the M2PL model

Figures 1 and 2 compare the distributions of bias and RMSE of the model parameters from the two methods under the four manipulated conditions for the between-item and within-item M2PL model respectively. As shown, GVEM generally produces comparable or more accurate model parameter estimates than MHRM run by the R package 'mirt' in all conditions for both between-item and within-item models. With respect to the manipulated conditions, increasing sample sizes helps reduce the RMSE and bias of the parameter estimates in both GVEM and MHRM in 'mirt'. Moreover, the RMSE and bias are generally higher when the correlations among factors are higher. This may be because higher correlation introduce multicollinearity among factors, making the parameter recovery more difficult (Wang & Nydick, 2015). Last but not least, the parameter recovery from the between-item

multidimensional model is better than the parameter recovery from the within-item multidimensional model. This is not surprising since the loading structure $\boldsymbol{A}$ is more complex in the within-item model. Figures 3 and 4 compare the distribution of bias and RMSE of the model parameters from GVEM and the JML method under the four manipulated conditions for the between-item and within-item M2PL models respectively. We observe that GVEM produces much lower RMSE and bias than the JML estimation under all conditions for both between-item and within-item models. The performance of the JML estimator is especially worse in small sample and high correlation settings and under more complex within-item multidimensionality structure. This could be due to the fact that the JML estimator assumes $\boldsymbol{\theta}_i$'s as fixed effects whereas GVEM models them as random effects with multivariate Gaussian distributions which account for the factor correlations. This result suggests that our proposed estimation method not only is theoretically consistent but also performs better in practice particularly under these complex simulation settings with correlated latent factors.

Figure 5 shows the average computation times in seconds for GVEM and MHRM in 'mirt' over 100 replications. To demonstrate a thorough comparison of the computation time, additional simulation settings were added for Figure 5; three different sample sizes (N = 200, 500, and 1000) and three different test dimensions (K = 3, 4, and 5) were considered as the simulation settings, resulting in 9 conditions in total. Each column presents the results for the between-item and withinin-item model respectively. Overall, GVEM algorithm is computationally more efficient than MHRM in both low and high correlation settings with varying sample sizes. The most reduction in computation time was observed in between-item model with low correlation setting. Unsurprisingly, computation time increases for both

methods when the number of dimensions increases or when sample sizes increase.

We would like to emphasize that the above observations regarding the MHRM algorithm are based on the implementation of the algorithm in the 'mirt' package under the default setting. Researchers using other packages may get slightly different results. We also tried other tuning methods in *flex*MIRT and found that a more careful tuning can improve the performance of MHRM as in 'mirt'; on the other hand, the estimation results can be very sensitive to the tuning, and the optimal tuning of MHRM could vary case by case, depending on the model setting and the correlation of the latent traits. For instance, following one reviewer's kind suggestion, we found that the strategy of combining mirt's default Stage 3 setup with *flex*MIRT's default Stages 1 and 2 setup provides slightly better estimation results than the proposed GVEM under the high correlation and between item model setting (while still slightly worse under the low correlation and within item model settings); please see Figure 1 in the Supplementary Material. Based on these observations, we clarify that the simulation study does not intend to conclude that the proposed GVEM outperforms the MHRM algorithm, but rather to show that GVEM provides a good alternative estimation method for MIRT, which does not rely on much tuning. Thoroughly evaluating the optimal tuning of MHRM algorithm is an interesting research problem, yet it is beyond the scope of the current paper, and we would like to leave that as a future study.

5.3 Results for the M3PL model

For the M3PL model, the sample size and forget rate for stochastically optimized 3PL algorithm were chosen based on pilot testing of various sample sizes and forget rates. We

observed that using the whole data set for the initial estimation step helped a lot with the estimation precision. Hence the forget rate was fixed at a small value so that the information from entire data set in the first iteration was weighted more heavily in the subsequest iterations (i.e. forget the information from entire data set slowly with small forget rate). After the first iteration, only 5 data points were sampled at a time, resulting in a huge reduction in computation time.

Figures 6 and 7 present the distributions of bias and RMSE of the model parameters from the two methods under the four manipulated conditions for the between-item and within-item M3PL model, respectively. During simulation studies, we observed that the performance of MHRM was quite unstable and the model did not converge well in M3PL under all manipulated conditions. Specifically, model did not converge in about 30 to 45% of the total experiments in most conditions. In another 15 to 20% of the experiments, the model converged but the estimates of the model parameters exploded to surprisingly high values, which implies the instability of the parameter estimation. For MHRM method, we excluded these results from the total of 100 experiments and reported only the values that seem more meaningful. On the other hand, we report the results for all 100 experiments for the GVEM method. Precisely, in Figure 6, 40 cases for (a), 41 for (b), 28 for (c), and 40 for (d) were reported. In Figure 7, 48 cases for (a), 46 for (b), 54 for (c), and 47 for (d) were reported. Note again that in both Figures, we report all 100 experiments for GVEM method because they all converged successfully. Similarly as in the simulation studies for M2PL, increasing sample sizes helps reduce the RMSE and bias of the parameter estimates in both GVEM and MHRM. However, the RMSE for MHRM method is quite high with large variation under most conditions. Overall, we observe that for varying sample sizes and correlations between

latent traits, GVEM performs better than MHRM, even after excluding unstable estimation results for MHRM. Given that the inclusion of guessing parameter poses model estimation challenge is well-documented in literature (e.g, Lord, 1968; Thissen & Wainer, 1982; Yen, 1987), it is not too surprising to note the large proportion of non-converged replications from MHRM. However, the stable performance of GVEM further reinforces its promise as a robust alternative method to the current status-quo, in particular when guessing parameter is included in the model. Also note that GVEM does not need much tuning for good performance, hence it is more accessible to broader audience who may not have the technical capacity to manually tune certain parameters, as may required by other algorithms. One last note to make is, for M3PL or 3PL models in general, marginal maximum a posteriori estimation (MMAP) is sometimes preferred over the maximum likelihood approach. That is, prior distributions are specified for constrained estimation of the a and c parameters to improve estimation stability (Kim, 2006). Therefore, one can compare GVEM with MMAP in a future study as well.

5.4 Estimating the Number of Dimensions

In this section, a separate simulation study was conducted to evaluate if $AIC^\star$ and $BIC^\star$ could help identify the correct number of factors from data. The simulation design is the same as illustrated in Section 5.1. The result is presented for different sample sizes and degrees of correlation between latent traits. A total of 100 independent samples were generated for each setting, and the proportion of replications in which the correct number of factors identified by $AIC^\star$ and $BIC^\star$ were recorded.

Table 1 and Table 2 present the correct estimation rate of the number of dimensions for M2PL and M3PL models respectively. As shown, increasing sample size help increase the correct estimation rate. In addition, similar to the findings in the previous sections, lower correlation is more preferable as it usually produced higher correct estimation rates. There is only one exception, though, for the within-item M3PL model, in which both $AIC^\star$ and $BIC^\star$ performed better for the higher correlation scenario regardless of the sample size. There is no appreciable difference between $AIC^\star$ and $BIC^\star$ except a few cells in Table 1: $AIC^\star$ performed better than $BIC^\star$ for large Σ_θ with sample size of 200, whereas $BIC^\star$ performed better for small Σ_θ with sample size of 200.

6 Real Data Analysis

In this section, the GVEM and MHRM algorithms were used to conduct an exploratory item factor analysis on the National Education Longitudinal Study of 1988 (NELS:88) data. In this data set, a nationally representative sample of approximately 24,500 students were tracked via multidimensional cognitive batteries 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). For the science subject, there are 25 items and four factors emerged 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 reasoning (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. Because the factor structure was analyzed using normal theory factor analysis more than two decades ago, we plan to reanalyze the data using the proposed new methods. In addition, pooling together both math and science domains result in potentially high dimensional data. First, both GVEM and MHRM were conducted assuming the number of factors were 6. The focus is on the recovery of the correlation matrix $\Sigma_{\boldsymbol{\theta}}$ and its comparison between two methods. Since the exploratory item factor analysis was conducted, in both GVEM and MHRM we assumed that $\Sigma_{\boldsymbol{\theta}} = I_K$ during GVEM estimation and later performed the same promax rotation to estimate the correlation matrix $\hat{\Sigma}_{\boldsymbol{\theta}}$. Second, GVEM was used to explore the dimension of latent traits from the data.

Table 3 shows the estimated $\Sigma_{\boldsymbol{\theta}}$ from both methods assuming the number of factors is 6. The correlations in $\hat{\Sigma}_{\boldsymbol{\theta}}$ from two algorithms look comparable although most values from GVEM are slightly smaller than those from MHRM. The negative correlations on the last row, especially, are similar between two correlation matrices. Please note that $\hat{\Sigma}_{\boldsymbol{\theta}}$ is invariant to the ordering of the latent traits (i.e., the factor labels are arbitrary), hence it is possible to reduce the differences between two matrices by further reordering their columns in Table 3.

To further explore the optimal number of factors from data, we applied the GVEM algorithm with the information criteria for dimension selection. Figure 8 presented the results of latent dimension selection under M2PL and M3PL models. By fitting the M2PL model to the data, the optimal dimensionality of the latent traits was estimated to be six by both $AIC^{\star}$ and $BIC^{\star}$ as shown in Figure 8. This corresponds to the number of latent traits identified in prior research. However, the dimensionality of the latent traits was estimated to

be five under the M3PL model. This result implies that some of the six latent traits may be highly correlated under the M3PL model that they are merged. Comparing the information criteria values across both M2PL and M3PL, it appears that $AIC^\star$ and $BIC^\star$ were smallest for the M2PL model with six factors. Hence, our results further validate the number of latent factors that could be extracted from the NELS:88 data. In addition, it suggests that the guessing didn't play a significant role in students' performance on the math and science cognitive test data.

7 Discussions

Variational methods are first introduced in psychometrics by Rijmen and Jeon (2013) for high dimensional IRT model with discrete latent traits, and later by Jeon et al (2017) in a form of a variational maximization-maximization algorithm for GLMMs with crossed random effects. Although their findings demonstrate great promise of variational methods as they apply in psychometrics, their methods are not ready for calibrating high-dimensional MIRT models with correlated latent factors and guessing parameters. In this paper, a new method based on variational approximation is proposed for the parameter estimation in the M2PL and M3PL models. Compared to the existing methods, it has the advantage of avoiding the calculation of intractable log-likelihood by approximating the lower bound to the log-likelihood. It also greatly reduces the computation complexity by deriving the closed-form updates in the every EM step. Moreover, the efficiency of the algorithm is further improved in the stochastic version. Simulation studies demonstrate that the proposed methods show better performance in terms of parameter recovery and computation time in both M2PL and

M3PL compared to the widely used MHRM method. Theoretical results are provided on the convergence rate, which shows that the estimation error goes to 0 as both the sample size and number of test items go to infinity. As byproducts of the GVEM algorithm, both $AIC^\star$ and $BIC^\star$ could be used to help identify the optimal number of latent factors from data, as reflected by the simulation results.

Although the current simulation study and data analysis focused on the exploratory item factor analysis, the GVEM algorithm can also be easily applied to the confirmatory item factor analysis. In the latter case, the loading matrix $\boldsymbol{A}$ will have structural 0's implying that certain items are irrelevant to certain factors. Similar to the approach in Cai (2010b), these user-defined restrictions can be incorporated in the estimation via linear constraints. Reflecting in the GVEM algorithm, due to the closed-form solutions in the M-step, handling the structural 0's basically means multiplying $\hat{\boldsymbol{A}}$ by a same size matrix of binary entries with 1's indicating the corresponding element is estimable.

Taking one step further, the GVEM algorithm could be coupled with latent variable selection (Sun, Chen, Liu, Ying, & Xin, 2016). Traditional approaches for identifying item factor loading structure proceeds in two steps: (i) allowing all item factor loadings to be freely estimated, and (ii) conducting a post-hoc rotation (Browne, 2001). While these rotation methods intend to produce a near-simple structure for the ease of interpretation, an arbitrary cut-off for the rotated factor loadings is often needed. In contrast, the latent variable selection avoids setting subjective cut-offs. The principle idea is to estimate the non-zero elements in the $\boldsymbol{A}$ matrix. Specifically, a penalty will be added to elements in $\boldsymbol{A}$ and when a factor is not associated with an item, the corresponding element in $\boldsymbol{A}$ will shrink to 0. Hence, this is a one-step approach where model calibration and factor selection are

completed simultaneously. This idea was first proposed by Sun et al. (2016), but they still used a traditional EM algorithm that can hardly be generalized to higher dimensions due to the computation burden. The GVEM algorithm proposed in this study is a good candidate for such one-step latent variable selection, and future studies could explore this direction.

Despite its computational efficiency and comparable estimation accuracy, GVEM does not produce standard errors of the model parameters as a byproduct of the estimation procedure. However, one can derive standard errors of the model parameters similarly following the existing works (Jamshidian & Jennrich, 2000). Relevant future research is needed on exploring the accuracy and efficiency of the estimation of standard errors in the GVEM framework. In addition, extending the GVEM framework to polytomous response models would be of another interest for the future research as polytomous response models have a wider range of applications including psychological and social science assessments with likert scales.

References

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

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer-Verlag.

Blei, D. M., & Jordan, M. I. (2004). Variational methods for the Dirichlet process. In *Proceedings of the twenty-first international conference on machine learning* (p. 12).

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

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

Cai, L. (2008). A MetropolisHastings RobbinsMonro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model. *Unpublished doctoral dissertation. Department of Psychology, University of North Carolina at Chapel Hill.*.

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

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

Celeux, G., & Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, *2*, 73-82.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29.

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

Davenport, M. A., Plan, Y., Van Den Berg, E., & Wootters, M. (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, *3*(3), 189–223.

Hall, P., Ormerod, J. T., & Wand, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, *21*(1), 369-389.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology*, *17*, 65-70.

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

Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, *58*(1), 30-37.

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 257–270.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2017). A variational maximization–maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, *82*(3), 693–716.

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.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355-381.

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS: 88 mathematics achievement to 12th grade. *American Educational Research Journal*, *34*(1), 124-150.

Lindstrom, M. J., & Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*(404), 1014–1022.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistic model. *Educational and Psychological Measurement*, *28*, 989-1020.

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

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, *6*(3), 457–489.

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.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*(2), 140-153.

Ormerod, J. T., & Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *21*(1), 2-17.

Parisi, G. (1988). *Statistical field theory.* Redwood City, CA: Addison-Wesley.

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

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research*, *206*(1), 647–662.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400–407.

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.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*, 397-412.

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

Titterington, D. (2004). Bayesian methods for neural networks and related models. *Statistical Science*, *19*(1), 128–139.

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.

Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, *48*(3-4), 233-243.

Yen, M., Wendy. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*, 275-291.

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.

		between-item		within-item	
Correlation($\Sigma_{\boldsymbol{\theta}}$)	N	$AIC^{\star}$	$BIC^{\star}$	$AIC^{\star}$	$BIC^{\star}$
small	200	76	92	69	94
	500	82	91	76	83
	1000	88	93	79	85
large	200	59	25	69	58
	500	66	41	82	81
	1000	83	52	84	89

Table 1: Simulation: correct estimation rate(%) in the M2PL model

		between-item		within-item	
Correlation($\Sigma_{\boldsymbol{\theta}}$)	N	$AIC^{\star}$	$BIC^{\star}$	$AIC^{\star}$	$BIC^{\star}$
small	200	47	47	63	63
	500	83	87	93	93
	1000	93	93	84	84
large	200	40	43	83	83
	500	60	60	97	97
	1000	73	73	97	97

Table 2: Simulation: correct estimation rate(%) in the M3PL model

GVEM	MHRM
$\begin{bmatrix} 1 & & & & & \\ .622 & 1 & & & & \\ .566 & .298 & 1 & & & \\ .472 & .112 & .426 & 1 & & \\ .489 & .869 & .424 & .248 & 1 & \\ -.767 & -.388 & -.701 & -.512 & -.595 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & & & & & \\ .549 & 1 & & & & \\ .697 & .432 & 1 & & & \\ .635 & .532 & .682 & 1 & & \\ .697 & .478 & .740 & .544 & 1 & \\ -.607 & -.497 & -.602 & -.525 & -.592 & 1 \end{bmatrix}$

Table 3: Real Data: comparison of estimated $\hat{\Sigma}_{\boldsymbol{\theta}}$

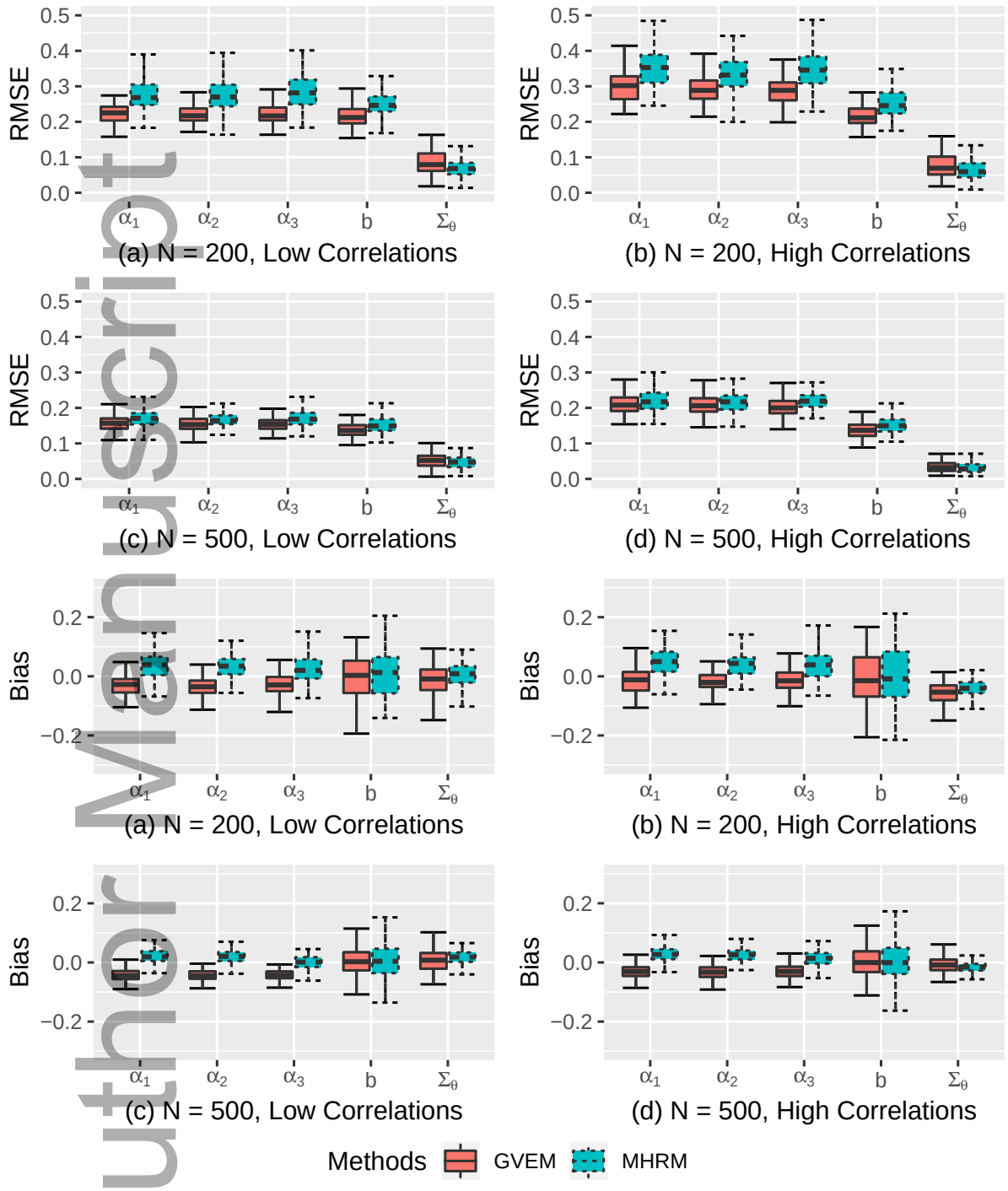

Figure 1: Parameter recovery of the between-item M2PL models from exploratory factor analysis

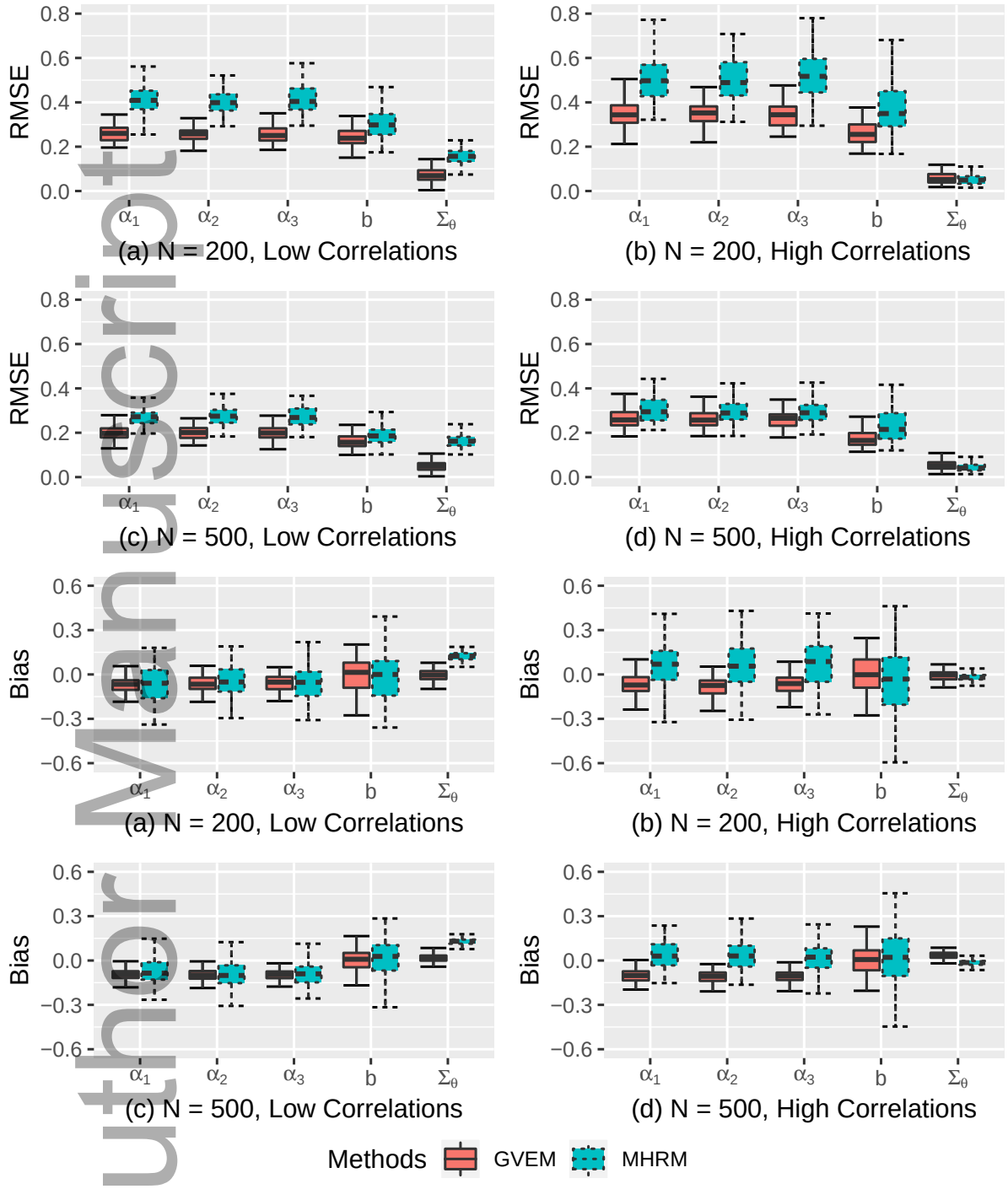

Figure 2: Parameter recovery of the within-item M2PL models from exploratory factor analysis

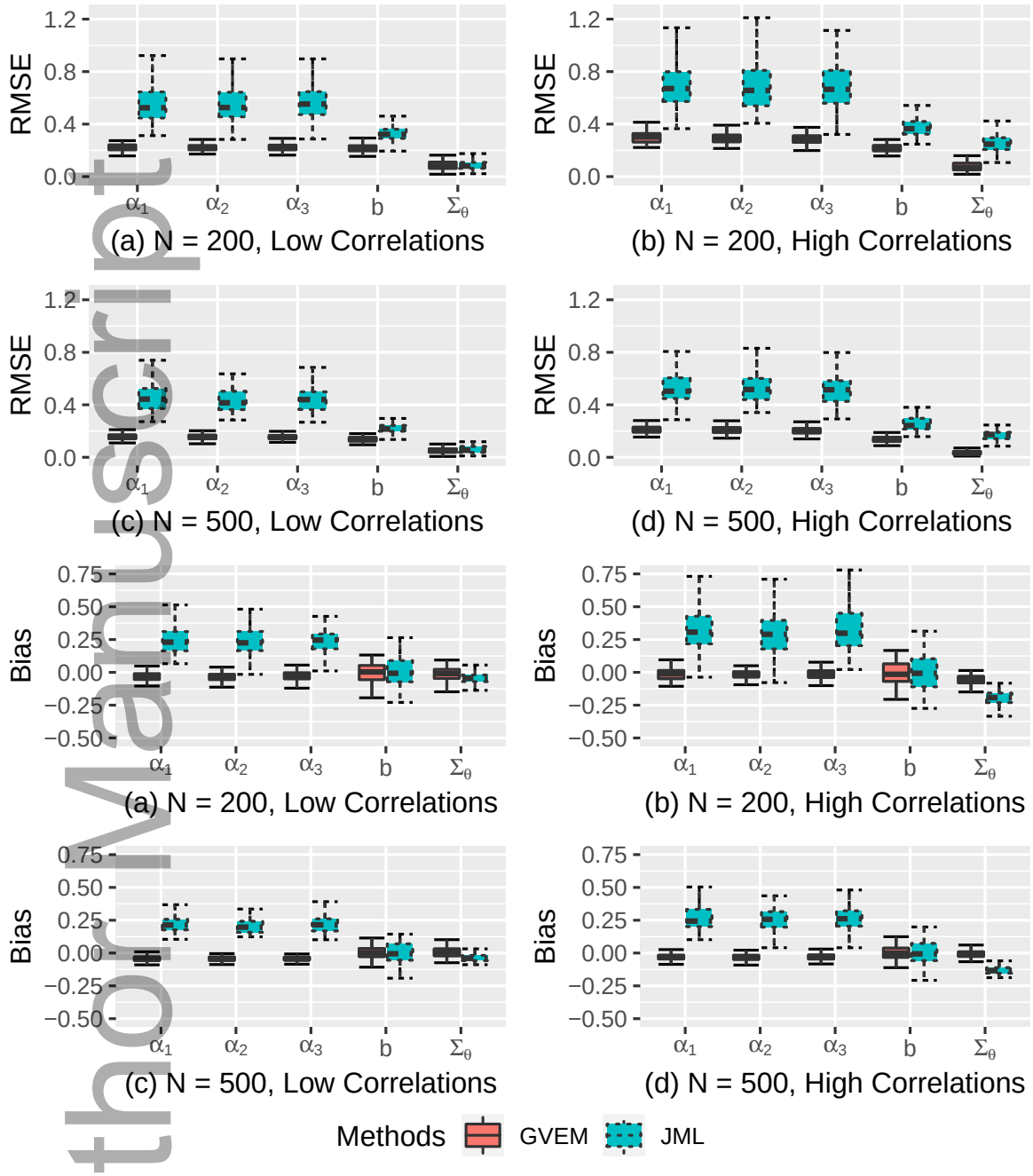

Figure 3: Parameter recovery of the between-item M2PL models from exploratory factor analysis using GVEM and Joint Maximum Likelihood (JML) estimator

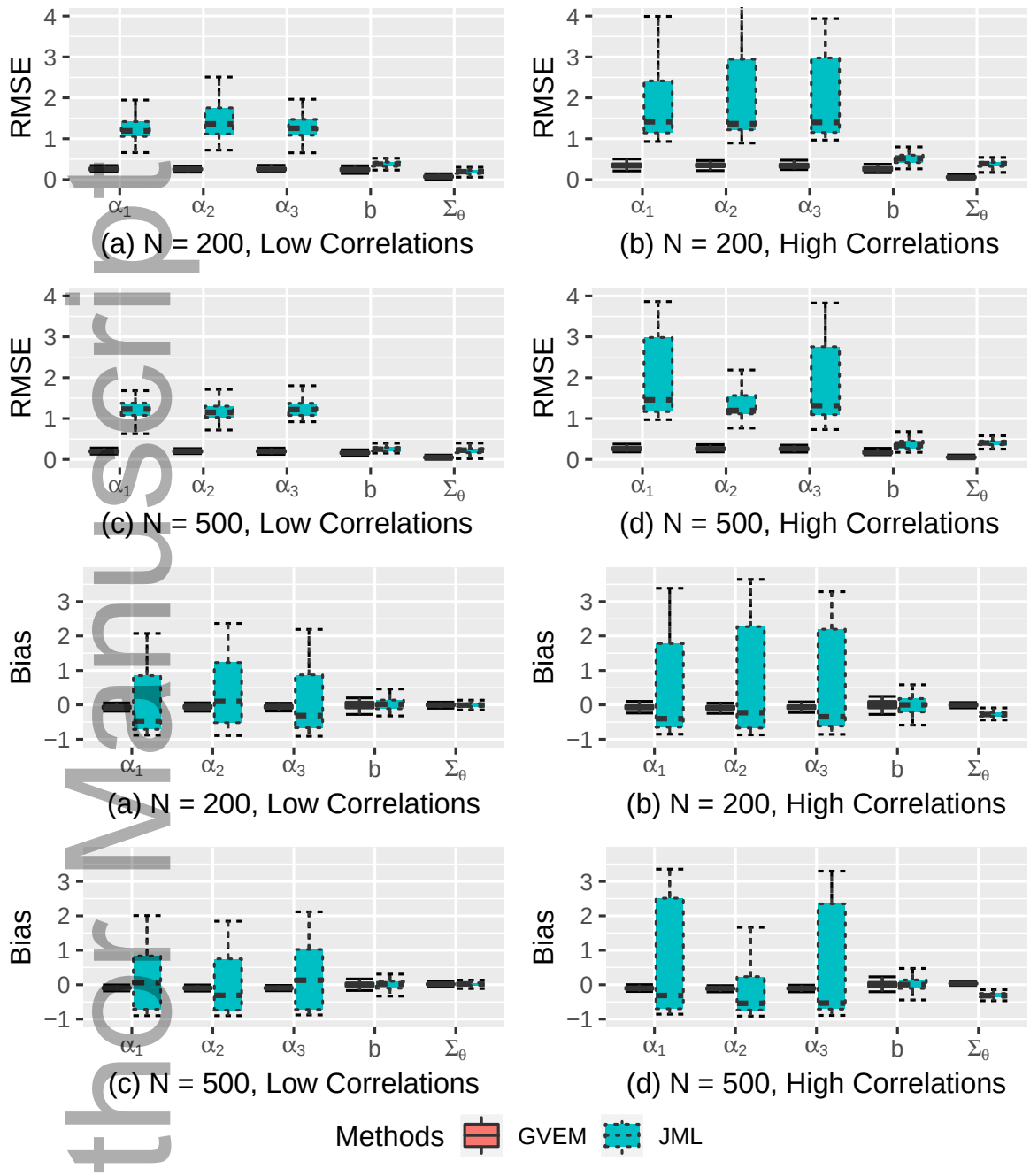

Figure 4: Parameter recovery of the within-item M2PL models from exploratory factor analysis using GVEM and Joint Maximum Likelihood (JML) estimator

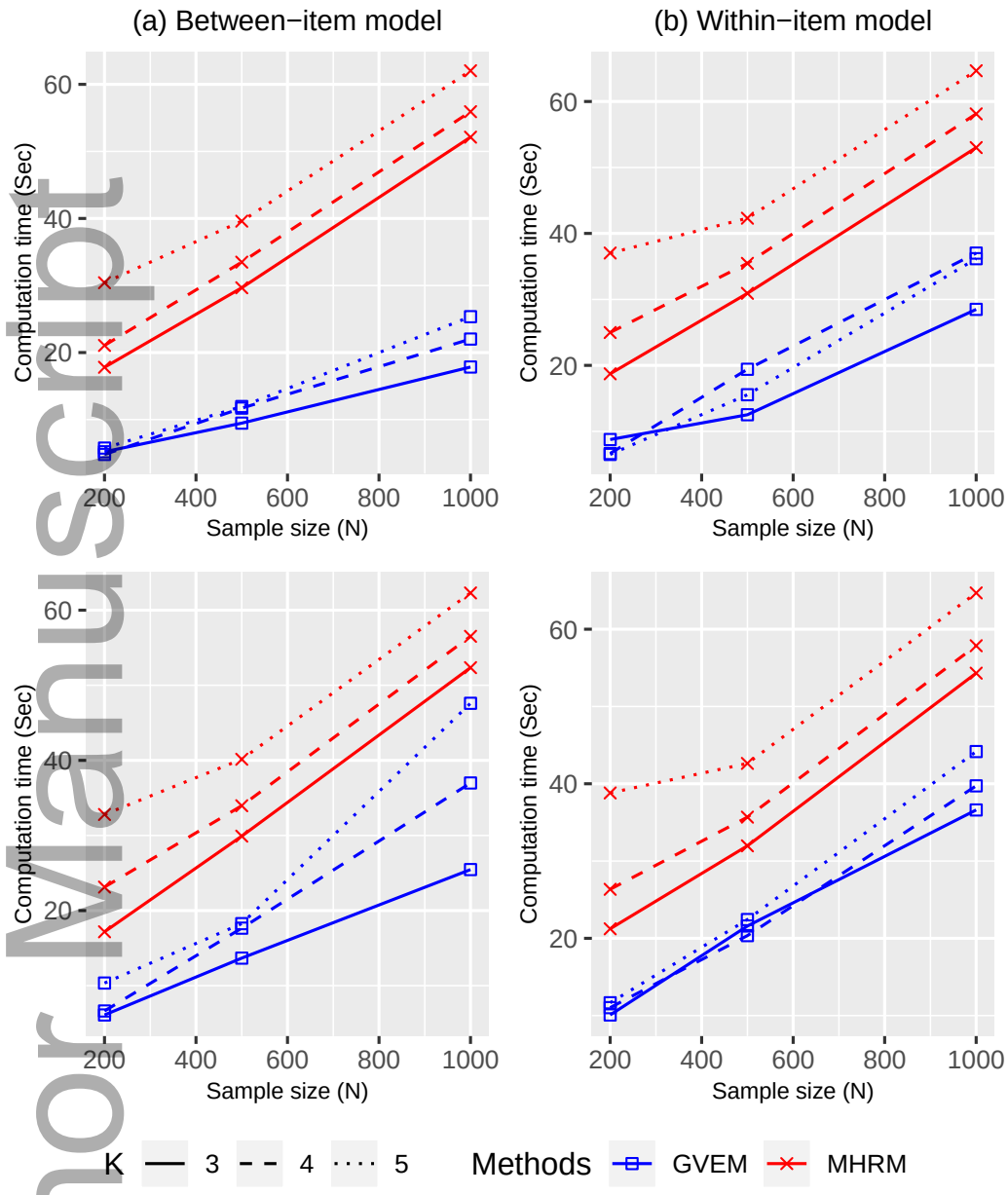

Figure 5: Average computation time for (a) between-item model (first column) and (b) within-item model (second column) with low correlation (first row) and high correlation (second row).

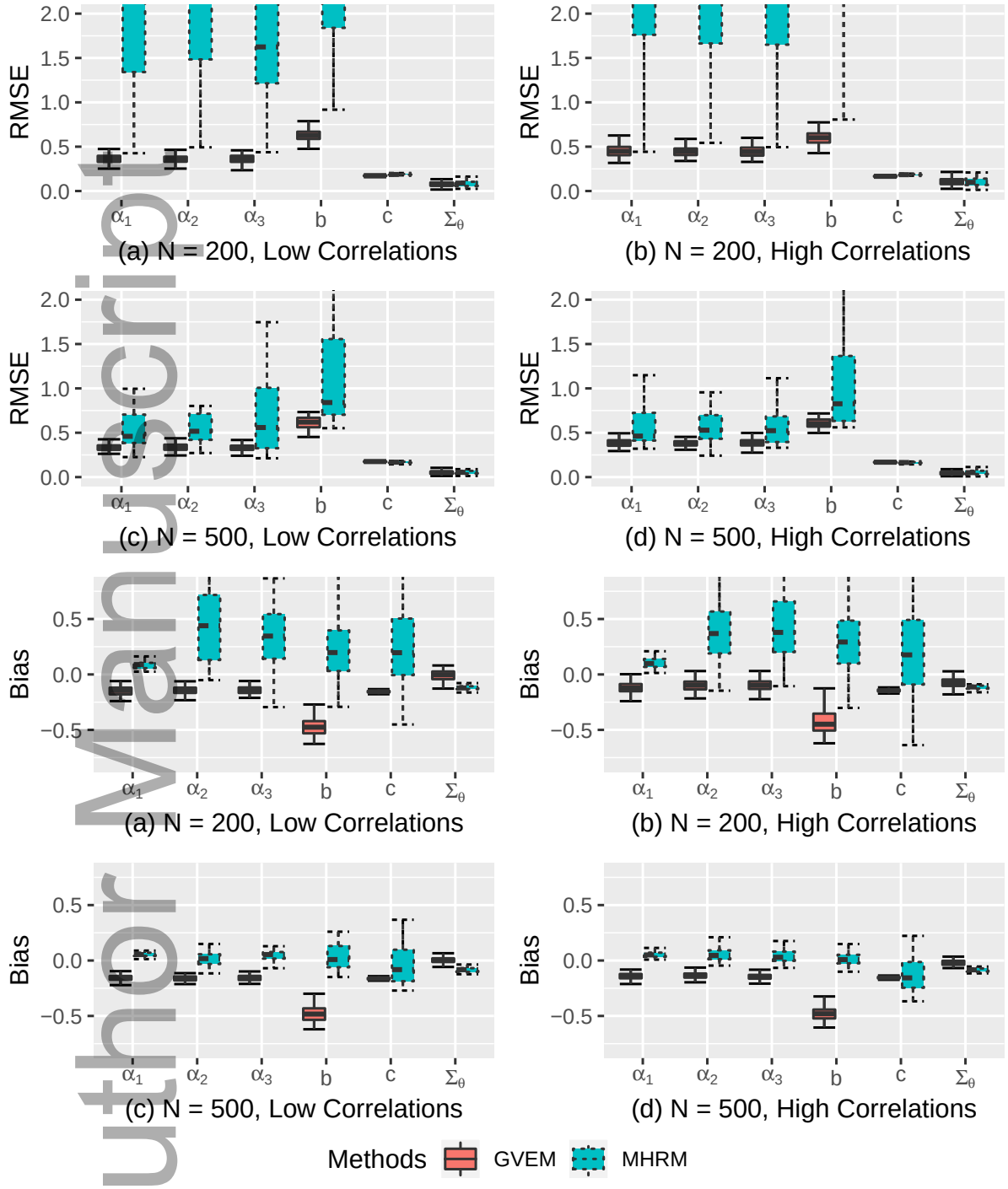

Figure 6: Parameter recovery of the between-item M3PL models from exploratory factor analysis. For MHRM, (a) 40, (b) 41, (c) 28, (d) 40 cases of simulation results were reported due to convergence issue. For GVEM, all 100 cases were reported under all conditions.

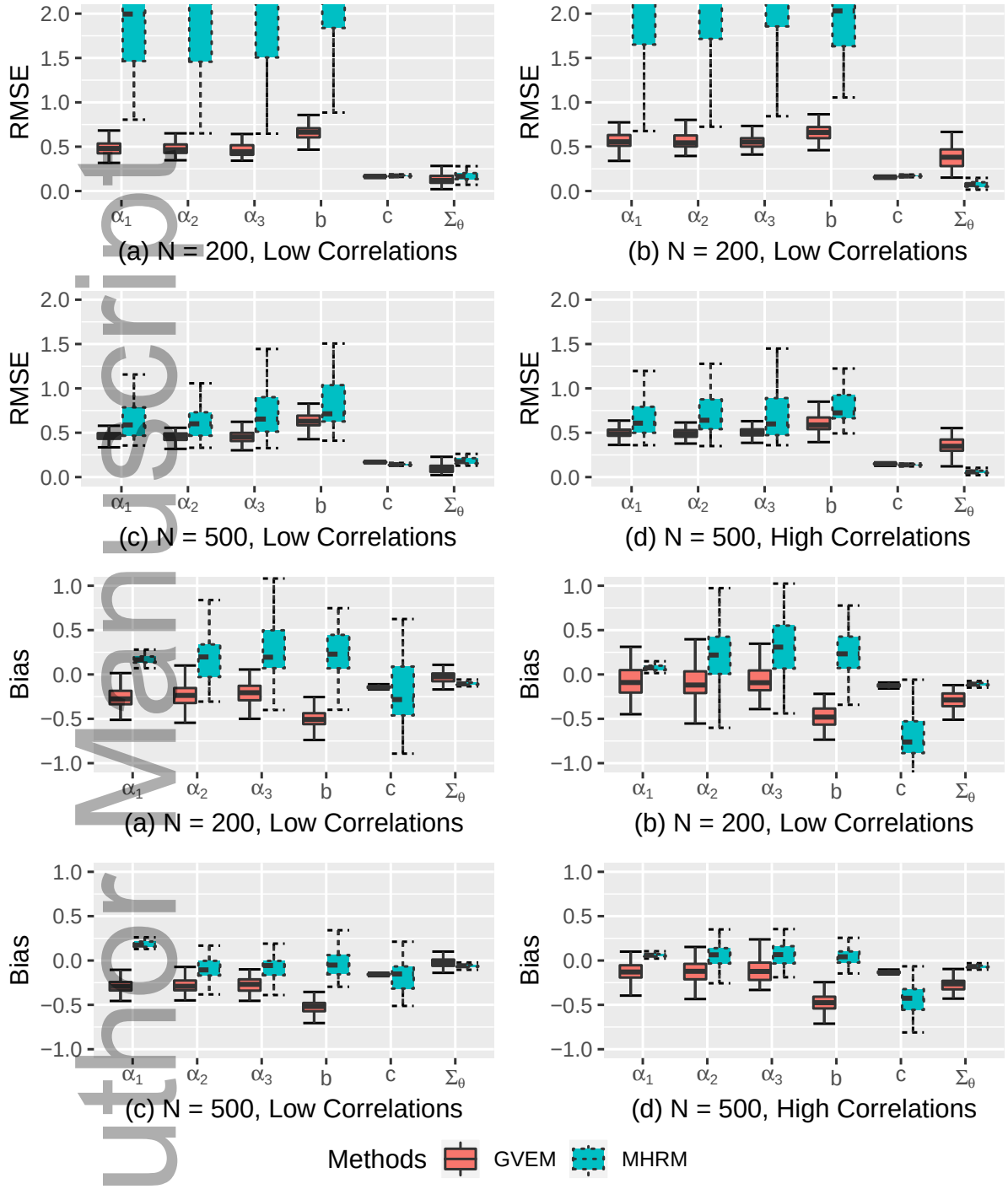

Figure 7: Parameter recovery of the within-item M3PL models from exploratory factor analysis. (a) 48, (b) 46, (c) 54, (d) 47 cases of simulation results were reported due to convergence issue. For GVEM, all 100 cases were reported under all conditions.

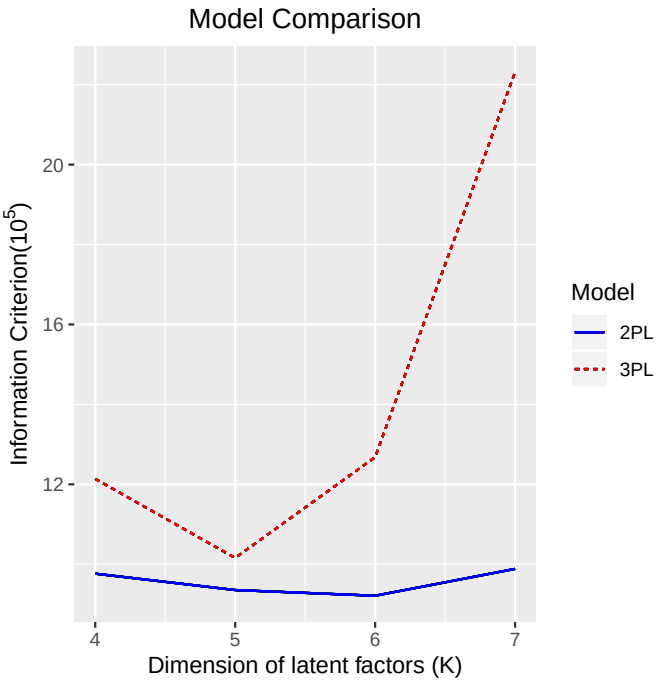

Figure 8: Real Data: $BIC^\star$ for both M2PL and M3PL ($AIC^\star$ shows the same trend).