Gaussian variational estimation for multidimensional item response theory

April E. Cho¹, Chun Wang², Xue Zhang³ and Gongjun Xu⁴*

¹Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA
²College of Education, University of Washington, Seattle, Washington, USA
³China Institute of Rural Education Development, Northeast Normal University, Changchun, China
⁴Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA

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 a 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 the latent variable structure. Various methods have been proposed that involve either 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 expectation–maximization (GVEM) algorithm which adopts 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 compared to the popular alternative Metropolis–Hastings Robbins–Monro algorithm. In addition, theoretical results are presented to establish the consistency of the estimator from the new GVEM algorithm.

1. Introduction

The increasing availability of rich educational survey data and the need to assess competencies in education pose great challenges to existing techniques used to handle and analyse 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 recent 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 (MHRM) algorithm, the stochastic expectation–maximization algorithm, and the fully Bayesian estimation methods, also help promote the use of MIRT models. However, even with these state-of-the-art algorithms, 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), MIRT models contain two or more parameters to describe the interaction between latent traits and responses to test items. In this paper we focus on the logistic model with dichotomous responses. Specifically, for the multidimensional two-parameter logistic (M2PL) model, there are $N$ individuals who respond independently to $J$ items with binary response variables $Y_{ij}$, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Then the item response function of the $i$th individual to the $j$th item is modelled by

$$P(Y_{ij} = 1|\theta_i) = \frac{\exp(\alpha_j^T \theta_i - b_j)}{1 + \exp(\alpha_j^T \theta_i - b_j)},$$

(1)

where $\alpha_j$ denotes a $K$-dimensional vector of item discrimination parameters for the $j$th item and $b_j$ specifies the corresponding difficulty level with item difficulty parameter as $b_j = \| \alpha_j \|_2$. $\theta_i$ denotes the $K$-dimensional vector of latent ability for student $i$.

For the multidimensional three-parameter logistic (M3PL) model, there is an additional parameter $c_j$, which denotes the guessing probability of the $j$th test item. The item response function is expressed as

$$P(Y_{ij} = 1|\theta_i) = c_j + (1 - c_j) \frac{\exp(\alpha_j^T \theta_i - b_j)}{1 + \exp(\alpha_j^T \theta_i - b_j)}.$$

(2)

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

$$l(M_p; Y) = \sum_{i=1}^{N} \log P(Y_i|M_p) = \sum_{i=1}^{N} \log \left( \int \prod_{j=1}^{J} P(Y_{ij}|\theta_i, M_p) \phi(\theta_i) d\theta_i \right).$$

(3)

where $Y_i = (Y_{ij}, j = 1, \ldots, J)$ is the $i$th subject’s response vector and $J$ is the total number of items in the test. $\phi$ denotes the $K$-dimensional Gaussian distribution of $\theta$ with mean 0 and covariance $\Sigma_\theta$. 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 $K$ dimensional integrals that are usually intractable. Direct numerical approximations 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 computationally 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 has a skewed shape. Other numerical approximation methods based on Monte Carlo simulation have also been developed in the literature, such as the Monte Carlo expectation–maximization (McCulloch, 1997), stochastic expectation–maximization (von Davier & Sinharay, 2010), and MHRM algorithms (Cai, 2010a, 2010b). These methods usually depend on sampling data points from a posterior distribution and would be computationally
intensive. Recently, Zhang, Chen, and Liu (2020) improved proposed using the stochastic EM (StEM) algorithm (Celeux & Diebolt, 1985) for 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 joint maximum likelihood (JML) estimation by treating the latent abilities as fixed effect parameters instead of random variables as in Equation (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 has been 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 (GLMMs) with nested random effects (Hall, Ormerod, & Wand, 2011; Ormerod & Wand, 2012). However, variational methods have only slowly gained recognition in psychometrics and educational measurement, with the pioneer papers by Rijmen and Jeon (2013) as well as Jeon et al. (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., Equation (3)) with a computationally feasible form, known as the variational lower bound, for 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. First, 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. Second, 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. Third, 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 this paper is organized as follows. Section 2 introduces the general framework of the Gaussian variational method and derivation of the 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. Sections 5 and 6...
illustrate the performance of the proposed GVEM method with simulation studies and on real data, respectively. Section 7 discusses future steps and concludes. The Online Supplementary Material includes the detailed mathematical derivations of the EM steps and the proofs of the theorem and proposition.

2. Gaussian variational EM

Henceforth, for the MIRT models in Equations (1) and (2), we denote the model parameters by $A = \{a_j, j = 1, \ldots, J\}$, $B = \{b_j, j = 1, \ldots, J\}$, and $C = \{c_j, j = 1, \ldots, J\}$. As defined in Section 1, we use the notation $M_p = \{A, B, C\}$ in the 3PL model and $M_p = \{A, B\}$ in the 2PL model for simplicity. Latent traits $\theta$ from different dimensions are correlated, resulting in a $K \times K$ covariance matrix $\Sigma_\theta$. To fix the origin and units of measurement, it is conventional to fix the mean and variance of all $\theta$'s to be 0 and 1, respectively. To remove rotational indeterminacy in the exploratory analysis, (i.e., to ensure model identifiability) researchers often assume $\Sigma_\theta = I_k$ and $A$ 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 $A$ is completely or partially specified while the remaining non-zero elements are left unknown. In this case, the correlation of latent traits $\theta$ is of interest and we need to estimate the covariance matrix $\Sigma_\theta$. In this paper we consider a general setting for $\Sigma_\theta$ 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 $Y$ is

$$l(M_p; Y) = \sum_{i=1}^{N} \log P(Y_i|M_p) = \sum_{i=1}^{N} \log \left( \prod_{j=1}^{J} P(Y_{ij}|\theta_i,M_p) \phi(\theta_i)d\theta_i \right),$$

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

$$l(M_p; Y) = \sum_{i=1}^{N} \int P(Y_i|M_p) \times q_i(\theta_i)d\theta_i,$$

for any arbitrary probability density function $q_i$ satisfying $\int q_i(\theta_i)d\theta_i = 1$. Since $P(Y_i|M_p) = P(Y_i, \theta_i|M_p)/P(\theta_i|Y_i,M_p)$, then we can further write

$$l(M_p; Y) = \sum_{i=1}^{N} \int \frac{P(Y_i, \theta_i|M_p)}{P(\theta_i|Y_i,M_p)} \times q_i(\theta_i)d\theta_i$$

$$= \sum_{i=1}^{N} \int \frac{P(Y_i, \theta_i|M_p)}{q_i(\theta_i)} \times q_i(\theta_i)d\theta_i$$

$$= \sum_{i=1}^{N} \int \frac{P(Y_i, \theta_i|M_p)}{q_i(\theta_i)} \times q_i(\theta_i)d\theta_i + KL\{q_i(\theta_i)||P(\theta_i|Y_i,M_p)\}$$

where $KL\{q_i(\theta_i)||P(\theta_i|Y_i,M_p)\} = \int \log \frac{q_i(\theta_i)}{P(\theta_i|Y_i,M_p)} \times q_i(\theta_i)d\theta_i$ is the Kullback–Leibler (KL) distance between the distributions $q_i(\theta_i)$ and $P(\theta_i|Y_i,M_p)$. The KL distance
Motivated by this observation, we use the Gaussian approximation procedure that
leads to a two-step iterative procedure. In the variational E-step, we choose
a distribution \( q_i(\theta_i) \) that gives the best approximation of the marginal likelihood
function, which corresponds to the estimated posterior distribution
\( P(\theta_i|Y_i,M_p) \), or equivalently to maximizing the lower bound in (4).

\[
l(\theta_i|Y_i,M_p) = \log P(Y_i,\theta_i|M_p)_{\theta_i}q_i(\theta_i)d\theta_i - \sum_{i=1}^N \log q_i(\theta_i)d\theta_i
\]

and equality holds when \( q_i(\theta_i) = P(\theta_i|Y_i,M_p) \) for \( i = 1, \ldots, N \).

The follow-up question is how to design the candidate distribution function \( q_i(\theta_i) \)
that gives the best approximation of the marginal likelihood. From the above argument,
the best choice is the unknown posterior distribution function \( P(\theta_i|Y_i,M_p) \). Although
this choice of \( q_i(\theta_i) \) is intractable, it provides a guideline for choosing \( q_i(\theta_i) \) in the
sense that a good choice of \( q_i(\theta_i) \) must approximate \( P(\theta_i|Y_i,M_p) \) 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 a distribution \( q_i \) that minimizes the KL distance
function, which corresponds to the estimated posterior distribution \( P(\theta_i|Y_i,M_p) \) with
\( \hat{M}_p \) from the previous step estimates. The E-step then evaluates the expectation with
respect to the \( q_i \)'s, namely,

\[
\sum_{i=1}^N \log P(Y_i,\theta_i|M_p)_{\theta_i}q_i(\theta_i)d\theta_i
\]

which is equal to the lower bound in (4), except the additional constant term
\(-\sum_{i=1}^N \log q_i(\theta_i)_{\theta_i}q_i(\theta_i)d\theta_i \) 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 \( \theta_i \). In the MIRT model it is known that the integral in
(5) does not have an explicit form, and in the literature numerical approximation methods
are often used, such as 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 \( \theta_i \), the variational
inference method uses alternative choices of the \( q_i(\theta_i) \)'s to approximate the marginal
likelihood function. The choices of \( q_i(\theta_i) \) not only approximate the posterior
\( P(\theta_i|Y_i,M_p) \) 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
\( P(\theta_i|Y_i,M_p) \) can be well approximated by a Gaussian distribution (Bishop, 2006).
Motivated by this observation, we use the Gaussian approximation procedure that
chooses \( q_i(\theta_i) \) from a family of Gaussian distributions such that the KL distance
between \( q_i(\theta_i) \) and \( P(\theta_i|Y_i,M_p) \) is minimized. The estimation is then taken as a two-
step iterative procedure. In the variational E-step, we choose \( q_i(\theta_i) \) by minimizing the

KL distance between \( q_i(\theta_i) \) and \( P(\theta_i|Y_i,M_p) \) and evaluate the expectation of the likelihood function with respect \( q_i(\theta_i) \), 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 4th subject’s likelihood function due to the independence of different subjects’ responses. The joint distribution function of \( \theta_i \) and \( Y_i \) is:

\[
\log P(Y_i, \theta_i|A,B) = \log P(Y_i|\theta_i, A, B) + \log \phi(\theta_i)
\]

\[
= \sum_{j=1}^{J} \left[ Y_j \log \frac{\exp \left( a_j^T \theta_i - b_j \right)}{1 + \exp \left( a_j^T \theta_i - b_j \right)} + (1 - Y_j) \log \frac{1}{1 + \exp \left( a_j^T \theta_i - b_j \right)} \right] + \log \phi(\theta_i).
\]

The difficulty of handling the marginal distribution of \( Y_i \) mostly comes from the logistic sigmoid function, which results in the integration over \( \theta_i \) not being in closed form in the E-step (i.e., expression (5)).

To avoid dealing with an intractable likelihood in E-step, we use a local variational method initially proposed in the machine learning literature (Bishop, 2006; Jordan et al., 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, we henceforth denote \( x_{ij} = b_j - a_j^T \theta_i \). Because of the concavity of the logistic sigmoid function \( \log \left( 1/(1 + e^{-x_{ij}}) \right) \), by the local variational method, we have the following variational lower bound on the logistic sigmoid function:

\[
\frac{e^{x_{ij}}}{1 + e^{x_{ij}}} = \max_{\xi_{ij}} \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} \exp \left\{ \frac{(x_{ij} - \xi_{ij})^2}{2} - \eta(\xi_{ij}) \left( x_{ij}^2 - \xi_{ij}^2 \right) \right\}
\]

\[
\geq \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} \exp \left\{ \frac{(x_{ij} - \xi_{ij})^2}{2} - \eta(\xi_{ij}) \left( x_{ij}^2 - \xi_{ij}^2 \right) \right\},
\]

(6)

where \( \xi_{ij} \) is a variational parameter that is introduced to approximate the objective function \( e^{x_{ij}}/(1 + e^{x_{ij}}) \), and \( \eta(\xi_{ij}) = (2\xi_{ij})^{-1} \left[ e^{\xi_{ij}}/(1 + e^{\xi_{ij}}) - 1/2 \right] \). We then aim to estimate the parameter \( \xi_{ij} \) that achieves the equality in (6). By introducing an additional variational parameter \( \bar{\xi}_{ij} \), we successfully avoid the problem of estimating the intractable integral in the E-step. The values of \( \bar{\xi}_{ij} \) 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 \( \log P(Y_i, \theta_i|A,B) \),
\[
\log P(Y_i, \theta_i | A, B) \geq \sum_{j=1}^{J} \log \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} + \sum_{j=1}^{J} Y_{ij} (\alpha_j^T \theta_i - b_j) + \sum_{j=1}^{J} \left( \frac{b_j - \alpha_j^T \theta_i - \xi_{ij}}{2} \right) - \sum_{j=1}^{J} \eta(\xi_{ij}) \left[ (b_j - \alpha_j^T \theta_i)^2 - \xi_{ij}^2 \right] + \log \phi(\theta_i) =: l(Y_i, \theta_i, \xi_i | A, B),
\]

where \(\xi_i = (\xi_{ij}, j = 1, \ldots, J)^T\).

The key step is to find the optimal variational distribution \(q_i(\theta_i)\), which we describe in detail in the next section.

### 3.1. Algorithm details

#### 3.1.1. Choice of \(q_i\)

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

\[
\log q_i(\theta_i) \propto \sum_{j=1}^{J} \left( Y_{ij} - \frac{1}{2} \right) \alpha_j^T \theta_i - \sum_{j=1}^{J} \eta(\xi_{ij}) \left( b_j - \alpha_j^T \theta_i \right)^2 - \frac{\theta_i^T \Sigma_i^{-1} \theta_i}{2}.
\]

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

\[
\mu_i = \Sigma_i \times \sum_{j=1}^{J} \left\{ 2\eta(\xi_{ij}) b_j + Y_{ij} - \frac{1}{2} \right\} \alpha_j^T,
\]

\[
\Sigma_i^{-1} = \Sigma_0^{-1} + 2 \sum_{j=1}^{J} \eta(\xi_{ij}) \alpha_j \alpha_j^T
\]

(7) (8)

With the variational densities \(q_i(\theta_i)\), we aim to estimate model parameters \(\xi_i, \alpha_j\) and \(b_j\) by maximizing the lower bound of the marginal likelihood. Suppose we have \(\xi_i\) from a previous step’s estimation or the initial values, denoted by \(\xi^{(t)}\). Similarly, define \(A^{(t)} = \{a_j^{(t)}, j = 1, \ldots, J\}\), \(B^{(t)} = \{b_j^{(t)}, j = 1, \ldots, J\}\), \(\Sigma_0^{(t)}\), \(\mu_i^{(t)}\) and \(\Sigma_i^{(t)}\). The EM iteration is presented below

#### 3.1.2. E-step

In the E-step, we evaluate the closed-form lower bound of the expected log-likelihood with respect to the variational distributions \(q_i\). With iteratively updated variational parameters \(\mu_i^{(t)}\) and \(\Sigma_i^{(t)}\), we easily evaluate the \(t\)th iteration’s lower bound of the
expected log-likelihood. Denote the $t$th iteration’s variational density by $q_i^{(t)}(\theta_i) = q_i(\theta_i|\xi_i^{(t)}, A^{(t)}, B^{(t)}, \Sigma^{(t)})$. Then, the $t$th iteration’s lower bound can be derived as

$$
E^{(t)}(A, B, \xi) := \sum_{i=1}^{N} \int l(Y_i, \theta_i|\xi_i) A \times q_i^{(t)}(\theta_i) d\theta_i
$$

$$
= \sum_{i=1}^{N} \sum_{j=1}^{J} \left( \log \frac{e^{\xi_{ij}^{(t)}}}{1 + e^{\xi_{ij}^{(t)}}} + \left( \frac{1}{2} - Y_{ij} \right) b_j^{(t)} - \frac{1}{2} \right) b_j^{(t)} + \left( Y_{ij} - \frac{1}{2} \right) \alpha_j^{(t)} T \mu_i^{(t)}
$$

$$
- \frac{1}{2} \left( \xi_{ij}^{(t)} - \eta(\xi_{ij}^{(t)}) \right) - \left( b_j^{(t)} - 2 b_j^{(t)} \alpha_j^{(t)} T \mu_i^{(t)} + \alpha_j^{(t)} T \Sigma^{(t)} + \left( \mu_i^{(t)} \right) \left( \mu_i^{(t)} \right)^T \right) + \frac{1}{2} \left( \alpha_j^{(t)} - \xi_{ij}^{(t)} \right)^2
$$

$$
+ \frac{N}{2} \log \left( \Sigma_\theta^{(t)} \right) - \sum_{i=1}^{N} \frac{1}{2} \text{Tr} \left( \left( \Sigma_\theta^{(t)} \right)^{-1} \left[ \Sigma_i^{(t)} + \left( \mu_i^{(t)} \right) \left( \mu_i^{(t)} \right)^T \right] \right).
$$

### 3.1.3. M-step

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

$$
\alpha_j = \frac{1}{2} \left[ \sum_{i=1}^{N} \eta(\xi_{ij}) \Sigma_i + \eta(\xi_{ij}) \mu_i \right]^{-1} \sum_{i=1}^{N} \left[ Y_{ij} - \frac{1}{2} + 2 b_j \eta(\xi_{ij}) \mu_i \right] \mu_i^T
$$

$$
b_j = \frac{\sum_{i=1}^{N} \left[ \frac{1}{2} - Y_{ij} + 2 \eta(\xi_{ij}) \alpha_j^T \mu_i \right]}{\sum_{i=1}^{N} 2 \eta(\xi_{ij})},
$$

$$
\xi_{ij}^2 = b_j^2 - 2 b_j \alpha_j \mu_i + \alpha_j^T \Sigma_i + \mu_i \mu_i^T \alpha_j.
$$

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

$$
\Sigma_\theta = \frac{1}{N} \sum_{i=1}^{N} \Sigma_i + \mu_i \mu_i^T
$$

Note that if the $\Sigma_\theta$ is assumed to be the correlation matrix with unit diagonal, then we need to standardize the estimated $\Sigma_\theta$ to get the correlation matrix. Detailed derivations regarding the above EM steps are given in the Online 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 $M_p^{(0)} = \{A_0, B_0, \xi(0)\}$.
2: repeat
3: E-step: For step $t \geq 1$, update $\mu^{(t)}_i$ and $\Sigma^{(t)}_i$ according to closed-form equations (7) and (8).
4: M-step: Further update $M^{(t)}_p$ and $\xi^{(t)}$ iteratively according to closed-form equations (9), (10), and (11). Fix $\Sigma^{(t)}_\theta = I_K$ in the exploratory analysis or update $\Sigma^{(t)}_\theta$ 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 subsampling the data to form noisy estimates of the variational lower bound and model parameters. See Section 4.2 for a 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 by-product of the algorithm, we can empirically estimate the number of latent dimensions from data. Specifically, the Akaike information criterion (AIC) or Bayesian information criterion (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:

$$
\text{AIC}^* = 2(\| A \|_0 + \| B \|_0 + \| \Sigma_\theta \|_0) - 2E(\hat{A}, \hat{B}, \hat{\xi}),
$$

$$
\text{BIC}^* = \ln(N)(\| A \|_0 + \| B \|_0 + \| \Sigma_\theta \|_0) - 2E(\hat{A}, \hat{B}, \hat{\xi}),
$$

where $E(\hat{A}, \hat{B}, \hat{\xi})$ is the estimated variational lower bound and $\hat{A}, \hat{B}, \hat{\xi}$ are the final estimates from the GVEM estimation procedure. $\| A \|_0$ denotes the zero norm of the matrix $A$, which is simply the number of non-zero entries of $A$. The advantage of using GVEM to estimate test dimensionality is that it is computationally more efficient, especially with high-dimensional data and more complex models. This procedure can be easily applied in both the 2PL and the 3PL models. See Section 5 for more discussion.

### 3.2. Theoretical properties

In this section we establish theoretical bounds on the estimation of the model parameters in 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,
\[ A = [\alpha_{jk}]_{f \times k} \] denotes a matrix of factor loadings. Additionally, let \( \Theta = [\theta_{ij}]_{N \times K} \) denote a matrix of random variables following \( q_i(\theta_i) \) and let \( \hat{\Theta} = [\hat{\theta}_{ij}]_{N \times K} \) denote a matrix of estimated latent abilities from data. Define \( E_{\hat{\theta} \sim \hat{q}} \) to be the expectation with respect to the estimated variational densities \( \{q_i(\theta_i) \sim N(\mu_i, \Sigma_i) : i = 1, \ldots, N \} \) from data. Lastly, a superscript * denote a true parameter. For example, \( \theta_i^* \) denotes the \( i \)th person's true latent ability, which is a deterministic realization from its population distribution. We assume that the true parameters \( \Theta^* \) and \( A^* \) satisfy the following condition.

\[ \langle A1 \rangle. \| \theta_i^* \|^2 \leq C \text{ and } \| \alpha_{ij}^* \|^2 \leq C \text{ for all } i, j \text{ for some positive constant } C. \]

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

\[
\frac{1}{NJ}E_{\hat{\theta} \sim \hat{q}} \left[ \| \hat{\Theta}^T - \Theta^*(A^*)^T \|_F \right] \leq C_2Ge^C \sqrt{\frac{J+N}{NJ}} \sqrt{1 + \frac{\log(N + J)}{N + J}}
\]

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

The proof of Theorem 1 can be found in the Online 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 due to Davenport et al. (2014) and Theorem 1 in Chen et al. (2019). However, the previous work by Chen et al. (2019) treats \( \theta_i \) as fixed effects while this work follows the conventional MIRT model setting with \( \theta_i \) random effects and following a normal population distribution.

Remark 4. The Gaussian family is reasonable as the candidate choice of \( q \) according to Laplace approximation of the posterior distribution \( P(\theta_i | Y_i) \). The Laplace approximation of \( P(\theta_i | Y_i) \) is a normal distribution with the maximum likelihood estimator \( \hat{\theta}_i \) as mean and the inverse of the observed Fisher information \( I^{-1}(\hat{\theta}_i) \) as variance. Denote the true parameter by \( \theta_i^* \). By the Bernstein–von Mises theorem, since the \( P(Y_i | \theta_i) \), \( i = 1, \ldots, N \), have the same support and \( \theta_i \to \log P(Y_i | \theta_i) \) is twice continuously differentiable, it follows that \( \hat{\theta}_i \to \theta_i^* \) almost surely and the Laplace approximated distribution \( N(\hat{\theta}_i, I^{-1}(\hat{\theta}_i)) \) converges in distribution to the true limiting normal distribution \( N(\theta_i^*, I^{-1}(\theta_i^*)) \) as \( J \to \infty \). 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 \( \theta \).

Remark 5. Compared with the existing stochastic estimation algorithms, such as the MHMR algorithm and the STEM algorithm, the proposed estimation method has the advantage that each of the estimation iterations has simple closed-form update and it does
not involve stochastic sampling from some intermediate posterior distributions as in the current stochastic estimation algorithms. As discussed in Remark 3.2, 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 StEM, as illustrated in the next section.

4. GVEM for the M3PL model

Derivation of the variational lower bound is trickier got the M3PL function since the cancellation of log and exponential functions, 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 the $i$th individual answered the $j$th item based on their latent abilities or guessed it correctly (van Davier, 2009). We define $Z_{ij} = 1$ if $i$th 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 $\theta_i$ and $Z_{ij}$, $P(Y_{ij}|\theta_i)$ under the following hierarchical model is equivalent to equation (2) of the 3PL model:

$$Z_{ij} \sim \text{Bernoulli}(1 - c_j),$$

$$Y_{ij}|\theta_i, Z_{ij} = 1 \sim \text{Bernoulli}\left(\frac{\exp(\alpha_j^\top \theta_i - b_j)}{1 + \exp(\alpha_j^\top \theta_i - b_j)}\right),$$

$$Y_{ij}|\theta_i, Z_{ij} = 0 \sim \text{Bernoulli}(1(Y_{ij} = 1)).$$

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

$$P(Y_{ij}|Z_{ij}, \theta_i) = \left\{ \frac{\exp(\alpha_j^\top \theta_i - b_j)}{1 + \exp(\alpha_j^\top \theta_i - b_j)} \right\}^{Y_{ij}} \left\{ \frac{1}{1 + \exp(\alpha_j^\top \theta_i - b_j)} \right\}^{1 - Y_{ij}} Z_{ij} I(Y_{ij} = 1)^{1 - Z_{ij}}.$$

Without loss of generality we first focus on the $i$th subject’s likelihood function due to the independence of different subjects. Denote $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ij})$ and its distribution by $p(Z_i) = \prod_{j=1}^I p(Z_{ij})$. Then the complete data likelihood of the $i$th subject is
\[ \log P(Y_i, \theta_i, Z_i | A, B, C) \]
\[ = \log P(Y_i | \theta_i, Z_i, A, B, C) + \log p(\theta_i) + \log p(Z_i) \]
\[ = \sum_{j=1}^{I} \left\{ \begin{array}{l}
Y_{ij} Z_{ij} \log \left( \frac{\exp \left( \alpha_j^T \theta_i - b_j \right)}{1 + \exp \left( \alpha_j^T \theta_i - b_j \right)} \right) + \left( 1 - Y_{ij} \right) Z_{ij} \log \left( \frac{1}{1 + \exp \left( \alpha_j^T \theta_i - b_j \right)} \right) \\
+ \sum_{j=1}^{I} \left\{ (1 - Z_{ij}) \log f(Y_{ij} = 1) \right\} + \log p(\theta_i) + \log p(Z_i) \\
\end{array} \right. 
\]
\[ = \sum_{j=1}^{I} \begin{array}{l}
Y_{ij} Z_{ij} \left( \alpha_j^T \theta_i - b_j \right) + Z_{ij} \log \frac{1}{1 + \exp \left( \alpha_j^T \theta_i - b_j \right)} + \left( 1 - Z_{ij} \right) \log f(Y_{ij} = 1) \\
+ \log p(\theta_i) + \log p(Z_i). 
\end{array} \]

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. See the Online Supplementary Material for the proof of Proposition 1. A 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 \( \theta_i \). The approach is similar to that of the 2PL model. For any arbitrary density functions \( q_i \) and \( r_i \) of the latent variables \( \theta_i \) and \( Z_{ij} \), the following equation always holds:

\[ \log P(Y_i | A, B, C) = \int \sum_{Z_i} \log P(Y_i | A, B, C) \times q_i(\theta_i) r_i(Z_i) d\theta_i. \]

where \( r_i(Z_i) = \prod_{j=1}^{I} r_{ij}(Z_i) \).

Note that \( P(Y_i | A, B, C) = P(Y_i, \theta_i, Z_i | A, B, C) / P(\theta_i, Z_i | Y_i, A, B, C) \). We can write

\[ \log P(Y_i | A, B, C) = \int \sum_{\theta_i, Z_i} \frac{P(Y_i, \theta_i, Z_i | A, B, C)}{P(\theta_i, Z_i | Y_i, A, B, C)} \times q_i(\theta_i) r_i(Z_i) d\theta_i 
\]
\[ = \int \sum_{\theta_i, Z_i} \frac{P(Y_i, \theta_i, Z_i | A, B, C)}{q_i(\theta_i) r_i(Z_i)} \times q_i(\theta_i) r_i(Z_i) d\theta_i 
\]
\[ + \text{KL} \left\{ q_i(\theta_i) r_i(Z_i) || P(\theta_i, Z_i | Y_i, A, B, C) \right\}. \]

Since the KL distance is non-negative by definition, we get a lower bound on the marginal likelihood similarly as in the 2PL model:

\[ \log P(Y_i | A, B, C) = \int \sum_{\theta_i, Z_i} \log P(Y_i, \theta_i, Z_i | A, B, C) \times q_i(\theta_i) r_i(Z_i) d\theta_i \]
\[ \geq \int \sum_{\theta_i, Z_i} \log (q_i(\theta_i) r_i(Z_i)) \times q_i(\theta_i) r_i(Z_i) d\theta_i \]  \hspace{1cm} (13)
\[ - \int \sum_{\theta_i, Z_i} \log (q_i(\theta_i) r_i(Z_i)) \times q_i(\theta_i) r_i(Z_i) d\theta_i \]  \hspace{1cm} (14)

Since the second line of inequality (14) does not depend on parameters \( A, B \) and \( C \), we focus on the first line for the derivation of the lower bound. Again, the \( i \)th subject’s likelihood function is
\[
\log P(Y_i, \theta_i, Z_i|A, B, C) \\
= \sum_{j=1}^{J} \left\{ Y_{ij}Z_{ij}(\alpha_j^T \theta_i - b_j) + Z_y \log \frac{1}{1 + \exp(\alpha_j^T \theta_i - b_j)} + (1 - Z_y) \log I(Y_{ij} = 1) \right\} \\
+ \log \phi(\theta_i) + \log p(Z_i).
\]

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

\[
\log P(Y_i, \theta_i, Z_i|A, B, C) \\
\geq \sum_{j=1}^{J} Z_y \log \left( \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} \right) + \sum_{j=1}^{J} Z_y Y_{ij}(\alpha_j^T \theta_i - b_j) \\
+ \sum_{j=1}^{J} \frac{1}{2} Z_y \left( b_j - \alpha_j^T \theta_i - \xi_{ij} \right) - \sum_{j=1}^{J} Z_y \eta(\xi_{ij}) \left\{ (b_j - \alpha_j^T \theta_i)^2 - \xi_{ij}^2 \right\} \\
+ \sum_{j=1}^{J} \{ (1 - Z_y) \log(1 - c_j) \} + \log \phi(\theta_i) + \log p(Z_i) \\
=: l(Y_i, \theta_i, Z_i, \xi_i|A, B, C).
\]

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

\[
l(Y_i, \theta_i, Z_i, \xi_i|A, B, C) = \sum_{j=1}^{J} (1 - Y_{ij} + Z_y Y_{ij}) \log \left( \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} \right) \\
+ \sum_{j=1}^{J} \left( 1 - Y_{ij} + Z_y Y_{ij} \right) Y_{ij} \left( \alpha_j^T \theta_i - b_j \right) + \sum_{j=1}^{J} \frac{1}{2} \left( 1 - Y_{ij} + Z_y Y_{ij} \right) \left( b_j - \alpha_j^T \theta_i - \xi_{ij} \right) \\
- \sum_{j=1}^{J} \left( Y_{ij} \{1 - Z_y\} \log(1 - c_j) \right) + \log \phi(\theta_i) + \sum_{j=1}^{J} \log p(Z_{ij}),
\]

where \(\log p(Z_{ij}) = (1 - Y_{ij} + Z_y Y_{ij}) \log(1 - c_j) + Y_{ij} (1 - Z_y) \log(c_j)\).

With variational distributions \(q_i\) and \(r_i\), 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 \(q_i\) and \(r_i\):

\[
E^{(t)}(A, B, C, \xi) := \sum_{i=1}^{N} \int_{\theta_i} \left[ \sum_{j} l(Y_i, \theta_i, Z_i, \xi_i|A, B, C) \times r^{(t)}(Z_i) \right] \times q^{(t)}(\theta_i) d\theta_i.
\]

Appropriate choices of the variational distributions will lead to a closed-form expression for 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

4.1.1. Choice of \( q_i \)

Let \( E_r \) denote the expectation with respect to the variational densities of the \( Z_{ij} \), that is, \( r_i(Z_{ij}) \). We can write

\[
E_r(A, B, C, \xi) := \sum_{i=1}^{N} \sum_{j=1}^{J} I(Y_i, \theta_i, Z_i, \xi_i | A, B, C) \times r_i(Z_{ij})
\]

\[
= \sum_{i=1}^{N} \left[ \sum_{j=1}^{J} \left( 1 - Y_{ij} + E_r[Z_{ij}] Y_{ij} \right) \log \frac{\exp\left( \xi_{ij} \right)}{1 + \exp\left( \xi_{ij} \right)} + \sum_{j=1}^{J} \left( 1 - Y_{ij} + E_r[Z_{ij}] Y_{ij} \right) Y_{ij} \left( \alpha_j^T \theta_i - b_j \right) \right]
\]

\[
+ \sum_{j=1}^{J} \left( 1 - Y_{ij} + E_r[Z_{ij}] Y_{ij} \right) \frac{1}{2} \left( b_j - \alpha_j^T \theta_i - \xi_{ij} \right)
\]

\[
- \sum_{j=1}^{J} \left( 1 - Y_{ij} + E_r[Z_{ij}] Y_{ij} \right) \eta(\xi_{ij}) \left\{ \left( b_j - \alpha_j^T \theta_i \right)^2 - \xi_{ij}^2 \right\}
\]

\[
+ \sum_{j=1}^{J} \left( Y_{ij} (1 - E_r[Z_{ij}]) \log I(Y_{ij} = 1) \right) + \log \phi(\theta_i) + \sum_{j=1}^{J} E_r[\log p(Z_{ij})].
\]

Conditional on the model parameters \( A, B, C \) and the variational parameters \( \xi_i, i = 1, \ldots, N \), by the variational inference theory, we can show that the variational distributions \( q_i(\theta_i) \), \( i = 1, \ldots, N \), that minimize the distances between them and the posterior distributions take the form

\[
\log q_i(\theta_i) \propto \sum_{j=1}^{J} \left( 1 - Y_{ij} + E_r[Z_{ij}] Y_{ij} \right) \left( Y_{ij} \frac{1}{2} \right) \alpha_j^T \theta_i
\]

\[
- \sum_{j=1}^{J} \left( 1 - Y_{ij} + E_r[Z_{ij}] Y_{ij} \right) \eta(\xi_{ij}) \left( b_j - \alpha_j^T \theta_i \right)^2 - \frac{1}{2} \theta_i^T \Sigma^{-1}_\theta \theta_i.
\]

The above likelihood function implies that \( q_i(\theta_i) \sim N(\theta_i | \mu_i, \Sigma_i) \), where the mean and covariance are

\[
\mu_i = \Sigma_i \sum_{j=1}^{J} \left\{ 2 \eta(\xi_{ij}) b_j + Y_{ij} \frac{1}{2} \right\} (1 - Y_{ij} + E_r[Z_{ij}] Y_{ij}) \alpha_j^T,
\]

\[
\Sigma^{-1}_i = \Sigma^{-1}_\theta + 2 \sum_{j=1}^{J} \left( 1 - Y_{ij} + E_r[Z_{ij}] Y_{ij} \right) \eta(\xi_{ij}) \alpha_j \alpha_j^T.
\]
4.1.2. Choice of \( r_{ij} \)

We follow similar steps to those for \( q_i \). That is, we take the expectation of the lower bound \( l(Y_t, \theta_t, Z_i, \xi_i | A, B, C) \) with respect to the variational density of \( \theta_t, q_t(\theta_t) \), 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 \( A, B, C \) and the variational parameters \( \xi_i \).

Let \( E_q \) denote the expectation with respect to the variational densities \( q_i \), and let \( E_{q_i} \) denote the expectation with respect to \( q_i \). Taking expectation of the lower bound \( l(Y_t, \theta_t, Z_i, \xi_i | A, B, C) \) with respect to \( q_i(\theta_t) \), we have

\[
E_q(A, B, C, \xi) = \sum_{i=1}^{N} \left( \frac{J}{\sum_{j=1}^{J} (1 - Y_{ij} + Z_{ij}Y_{ij}) \log \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} + \sum_{j=1}^{J} (1 - Y_{ij} + Z_{ij}Y_{ij}) Y_{ij} \left( \alpha^T_j E_{q_i}[\theta_i] - b_j \right) + \frac{J}{\sum_{j=1}^{J} (1 - Y_{ij} + Z_{ij}Y_{ij}) \frac{1}{2} \left( b_j - \alpha^T_j E_{q_i}[\theta_i] - \xi_{ij} \right) - \frac{1}{2} \left( \alpha^T_j \theta_i \right)^2 - \xi_{ij}^2 \right)}{1 + e^{\xi_{ij}}} \right) + \sum_{j=1}^{J} \log \left( \sum_{i=1}^{N} \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} \right)
\]

(18)

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

\[
\log r_{ij}(Z_{ij}) \propto Z_{ij} \left[ \log \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} + Y_{ij} \left( \alpha^T_j E_{q_i}[\theta_i] - b_j \right) + \frac{1}{2} \left( b_j - \alpha^T_j E_{q_i}[\theta_i] - \xi_{ij} \right) - \eta(\xi_{ij}) \left\{ E_{q_i} \left[ \left( b_j - \alpha^T_j \theta_i \right)^2 \right] - \xi_{ij}^2 \right\} + \log(1 - c_j) \right] + Y_{ij} \left( 1 - Z_{ij} \right) \log l(Y_{ij} = 1) + \log(1 - c_j).
\]

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

\[
s_{ij}^{-1} = 1 + \frac{c_j}{1 - s_{ij}} \frac{1 + e^{\xi_{ij}}}{e^{\xi_{ij}}} \exp \left\{ -Y_{ij} \left( \alpha^T_j E_{q_i}[\theta_i] - b_j \right) + \frac{1}{2} \left( b_j - \alpha^T_j E_{q_i}[\theta_i] - \xi_{ij} \right) - \eta(\xi_{ij}) \left\{ E_{q_i} \left[ \left( b_j - \alpha^T_j \theta_i \right)^2 \right] - \xi_{ij}^2 \right\} \right\}
\]

(19)

if \( Y_{ij} = 1 \), where \( E_{q_i}[\theta_i] = \mu_i \) and \( E_{q_i}[\left( b_j - \alpha^T_j \theta_i \right)^2] = b_j^2 - 2b_j \alpha^T_j \mu_i + \alpha^T_j [\Sigma_i + \mu_i \mu_i^T] \alpha_j \).

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

4.1.3. E-step

In every E-step, we choose the optimal variational distributions \( q_i \) and \( r_{ij} \), which is equivalent to estimating variational parameters \( \mu_i, \Sigma_i \), and \( s_{ij} \) for every \( i \) and \( j \). With
achieved by setting the derivative of $j$ iteratively updated variational parameters, (i.e., $M^{(t)}_p = \{A^{(t)}, B^{(t)}, C^{(t)}\}$), we derive the updating rule by taking derivative of the variational lower bound at $t$th step as follows:

$$E^{(t)}(A,B,C,\xi)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{I} \left( 1 - Y_{ij} + s_{ij} Y_{ij} \right) \left( \log \frac{e^{\xi_{ij}^{(t)}}}{1 + e^{\xi_{ij}^{(t)}}} + \left( \frac{1}{2} - Y_{ij} \right) b_{ij}^{(t)} + \left( Y_{ij} - \frac{1}{2} \right) \alpha_{ij}^{(t)} \mu_{ij}^{(t)} \right)$$

$$- \frac{1}{2} \xi_{ij}^{(t)} - \eta \left( \xi_{ij}^{(t)} \right) \left\{ b_{ij}^{(t)^2} - 2b_{ij}^{(t)} \alpha_{ij}^{(t)} \mu_{ij}^{(t)} + \left( \alpha_{ij}^{(t)} \right)^T \left[ \Sigma_{ij}^{(t)} + \mu_{ij}^{(t)} \left( \mu_{ij}^{(t)} \right)^T \right] \alpha_{ij}^{(t)} - \xi_{ij}^{(t)^2} \right\}$$

$$+ N \log \left( \Sigma_{ij}^{(t)} \right)^{-1} + \sum_{i=1}^{N} \sum_{j=1}^{I} \left\{ \left( 1 - Y_{ij} + s_{ij} Y_{ij} \right) \log \left( 1 - c_{ij}^{(t)} \right) + Y_{ij} \left( 1 - s_{ij}^{(t)} \right) \log \left( c_{ij}^{(t)} \right) \right\}.$$

4.1.4. M-step

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

For $\xi$ and $\Sigma_\theta$, the update is the same as in the 2PL model. For other parameters, we derive the updating rule by taking derivative of the variational lower bound $E(A,B,C,\xi)$ derived in the E-step. As a result, we have the following updating rule for $\alpha_j$, $b_j$ and $c_j$:

$$\alpha_j = \frac{1}{2} \left[ \sum_{i=1}^{N} \left( 1 - Y_{ij} + s_{ij} Y_{ij} \right) \eta \left( \xi_{ij} \right) \left( \Sigma_{ij} + \mu_{ij} \mu_{ij}^T \right) \right]^{-1} \sum_{i=1}^{N} \left( 1 - Y_{ij} + s_{ij} Y_{ij} \right) \left( Y_{ij} - \frac{1}{2} + 2b_{ij} \eta \left( \xi_{ij} \right) \mu_{ij} \right),$$

(20)

$$b_j = \frac{\sum_{i=1}^{N} \left( 1 - Y_{ij} + s_{ij} Y_{ij} \right) \left( \frac{1}{2} - Y_{ij} \right) + 2\eta \left( \xi_{ij} \right) \alpha_j^T \mu_{ij}}{\sum_{i=1}^{N} 2 \left( 1 - Y_{ij} + s_{ij} Y_{ij} \right) \eta \left( \xi_{ij} \right)},$$

(21)

$$c_j = \frac{\sum_{i=1}^{N} \left( Y_{ij} - s_{ij} Y_{ij} \right)}{\sum_{i=1}^{N} \left( 1 - Y_{ij} + s_{ij} Y_{ij} \right) + \sum_{i=1}^{N} \left( Y_{ij} - s_{ij} Y_{ij} \right)} = \frac{1}{N} \sum_{i=1}^{N} Y_{ij} (1 - s_{ij}).$$

(22)

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

1: Initialize $M_p^{(0)} = \{A_0, B_0, C_0\}, \xi^{(0)}$.
2: repeat
3: E-step: For step $t \geq 1$, update variational parameters $\mu^{(t+1)}_i$, $\Sigma^{(t+1)}_i$, and $s^{(t+1)}_{ij}$ according to closed-form equations (16), (17), and (19).
4: M-step: Further update $M_p^{(t+1)}$ iteratively according to closed-form equations (20), (21) and (22).
5: Update $\xi^{(t+1)}$ and $\Sigma^{(t+1)}$ as in M2PL.
6: Untile convergence

Remark 6. The theoretical property of the M3PL is more challenging to derive rigorously due to the addition of the guessing parameters $c_j$. 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 the Hellinger distance for probability distributions and matrices. The Hellinger distance for two scalars $p, q \in [0, 1]$ is defined as $d^2_H(p, q) := (\sqrt{p} - \sqrt{q})^2 + (\sqrt{1-p} - \sqrt{1-q})^2$. 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 define

$$d^2_H(P, Q) = \frac{1}{d_1 d_2} \sum_{i,j} d^2_H(P_{ij}, Q_{ij}).$$

Let $M = [M_{ij}]_{N \times J}$ be the matrix with entries $M_{ij}$ satisfying

$$\frac{\exp(M_{ij})}{1 + \exp(M_{ij})} = c_j + (1 - c_j) \frac{\exp(a_j^T \theta_i - b_j)}{1 + \exp(a_j^T \theta_i - b_j)}.$$

Let $P(Y|M)$ be a matrix of probability distributions $P(Y_{ij}|M_{ij})$, where $M_{ij}$ denotes a collection of model parameters $\alpha_{ij}, b_j, c_j$. Again, $M^*$ denotes a matrix of true parameters and $\hat{M}$ denotes estimated model parameters. Then by Theorem 2 of Davenport et al. (2014)

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

with probability $1 - C_1/N + J$ for absolute constants $C_1$ and $C_2$. Hence, the Hellinger distance between the estimated probability distribution and the true probability distribution goes to 0 as both $N \to \infty$ and $J \to \infty$. However, the consistency result for model parameters $\{\alpha_j, b_j, c_j : j = 1, \ldots, J\}$ 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., $\mu_i, \Sigma_i, \xi_{ij}, s_{ij}$) need to be optimized for every data point $i = 1, \ldots, N$. 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 et al., 2013). That is, at each iteration of the E-step, we subsample the data to form a noisy estimate of the variational lower bound and iteratively update the estimate with a decreasing step size. Then the M-step in Algorithm 4.1 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 a 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 StEM steps can be summarized as follows.

4.2.1. Stochastic E-step

For step $t \geq 1$, choose a subset of data $S_t$ with desired size. Choose a decreasing step size $\epsilon_t$. Update $\mu_i^{(t)}, \Sigma_i^{(t)}, \xi_{ij}^{(t)}$ 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 the 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 a noisy estimate of $t$th iteration’s expected variational lower bound $\hat{Q}_t$ as follows:

$$\hat{Q}_t = \sum_{i \in S_t} \int \left[ \Sigma_i^{(t)}(Y_i, \theta_i, Z_i, \xi_i, A, B, C) \times r_i^{(t)}(Z_i) \right] \times q_i^{(t)}(\theta_i) d\theta_i.$$ 

Then we obtain a stochastic approximation of the variational lower bound by a weighted average of the previous and current steps’ noisy estimates of the lower bound: $(1 - \epsilon_t)\hat{Q}_{t-1} + \epsilon_t \hat{Q}_t$.

4.2.2. M-step

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

Notice that this stochastic optimization idea is different from the stochastic component in the StEM algorithm (Nielsen, 2000). In the StEM algorithm, random samples of the unobserved latent variables $\theta_i$ are drawn from the conditional distribution of $\theta_i$ 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 subsampling of the observed data $\{Y_{ij}, i = 1, \ldots, N\}$ to form a noisy approximation of the variational lower bound $E(A, B, C, \xi)$ in the E-step.
In theory, if a sequence of step sizes satisfies the conditions such that

$$\sum \varepsilon_t = \infty \text{ and } \sum \varepsilon_t^2 < \infty,$$

(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 $t$th step size as $\varepsilon_t = (t + \tau)^{-r}$, where the forget rate $r \in (0, 1]$ and the delay $\tau \geq 0$. The forget rate controls how quickly old information is forgotten and the delay downweights early iterations to decrease the effect of the earlier noisy estimations. This step size obviously satisfies conditions (23). Thus, the iterative stochastic optimization of the E-step converges to a local optimum of the variational lower bound. In simulation, we fix the delay to be 1 and try various forget rates as different values of delay did not 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 MHRM algorithm implemented in the R package mirt (Chalmers, 2012). The Metropolis–Hastings sampler is used to draw missing data (which is $\theta$ 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 Metropolis–Hastings 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 $\alpha$ (i.e., equation (9) for the 2PL model and equation (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 $A$ while fix $\Sigma_\theta = I_K$ 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 and promax (Hendrickson & White, 1964) in oblique rotation. Other popular methods include 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 three and test length was fixed at 45.

Additionally, we compared the performance of GVEM to the 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 compare the performances for M2PL.
The manipulated conditions include: multidimensional structure (i.e., between-item and within-item multidimensionality); correlations among the latent traits; and 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 1, 2, and 3 factors, respectively. In all cases, item discrimination parameters were simulated from the Unif(1,2) distribution, and the 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 $\theta_i$ were generated from a multivariate normal distribution, $N(0, \Sigma)$, where $\Sigma$ is a covariance matrix whose diagonal elements were 1 and the off-diagonals were drawn from a 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 $Unif(0.1,0.3)$. The 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 5,000 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. One hundred 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 $\Sigma$ is calculated by taking differences between each entry of the true $\Sigma$ and estimated $\hat{\Sigma}$. 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 the 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 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 $A$ is more complex in the within-item model. Figures 3 and 4 compare the distribution of the 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 JML estimation under all conditions for both between-item and within-item models. The performance of the JML estimator is particularly 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 the $\theta_i$ 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 ($N = 200, 500, \text{ and } 1,000$) and three different test dimensions ($K = 3, 4, \text{ and } 5$) were considered as the simulation settings, resulting in nine conditions in total. Each column presents the results for the between-item and within-item model, respectively. Overall, the GVEM algorithm is computationally more efficient than MHRM in both low- and high-

![Figure 1](image-url). Parameter recovery of the between-item M2PL models from exploratory factor analysis.
correlation settings with varying sample sizes. The biggest reduction in computation time was observed for the between-item model in the 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 \textit{mirt} package under the default setting. Researchers using other packages may get slightly different results. We also tried other tuning methods in flexMIRT and found that a more careful tuning can improve the performance of MHRM as in \textit{mirt}; on the other hand, the estimation results

\textbf{Figure 2.} Parameter recovery of the within-item M2PL models from exploratory factor analysis.
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 flexMIRT’s default stage 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 setting); see Figure S1. 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.

**Figure 3.** Parameter recovery of the between-item M2PL models from exploratory factor analysis using GVEM and joint maximum likelihood (JML) estimator.
which does not rely on much tuning. Thoroughly evaluating the optimal tuning of the MHRM algorithm is an interesting research problem, yet it is beyond the scope of the current paper, and we leave it to a future study.

### 5.3. Results for the M3PL model

For the M3PL model, the sample size and forget rate for the 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
estimation precision. Hence the forget rate was fixed at a small value so that the information from the entire data set in the first iteration was weighted more heavily in subsequent iterations (i.e., the information from the entire data set is forgotten at a slow forget rate). After the first iteration, only five data points were sampled at a time, resulting in a huge reduction in computation time.

Figures 6 and 7 present the distributions of the bias and RMSE of the model parameters from the two methods under the four manipulated conditions for the between-item and

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).
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, the model did not converge in about 30–45% of the total experiments in most conditions. In another 15–20% of the experiments, the model converged but the estimates of the model parameters exploded to surprisingly high values, which implies instability of the parameter estimation. For the

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 issues. For GVEM, all 100 cases were reported under all conditions.
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. Specifically, 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 to 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 the 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 fact that inclusion of guessing parameters poses a challenge to model estimation 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 a 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 a broader audience who may not have the technical capacity to manually tune certain parameters, as may required by other algorithms.

One last observation is that, 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 could also compare GVEM with MMAP in a future study.

5.4. Estimating the number of dimensions

In this section a separate simulation study was conducted to evaluate whether AIC* and BIC* 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 was identified by AIC* and BIC* was recorded.

Tables 1 and 2 present the correct estimation rate of the number of dimensions for the 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* and BIC* performed better for the higher-correlation scenario regardless of the sample size. There is no appreciable difference between AIC* and BIC*, except for a few cells in

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

<table>
<thead>
<tr>
<th>Correlation (σ_θ)</th>
<th>N</th>
<th>Between-item</th>
<th></th>
<th></th>
<th>Within-item</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>AIC*</td>
<td>BIC*</td>
<td></td>
<td>AIC*</td>
<td>BIC*</td>
</tr>
<tr>
<td>Small</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>76</td>
<td>92</td>
<td></td>
<td>69</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>82</td>
<td>91</td>
<td></td>
<td>76</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>88</td>
<td>93</td>
<td></td>
<td>79</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Large</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>59</td>
<td>25</td>
<td></td>
<td>69</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>66</td>
<td>41</td>
<td></td>
<td>82</td>
<td>81</td>
<td></td>
</tr>
<tr>
<td>1,000</td>
<td>83</td>
<td>52</td>
<td></td>
<td>84</td>
<td>89</td>
<td></td>
</tr>
</tbody>
</table>
Table 1: AIC* performed better than BIC* for large Σθ with a sample size of 200, whereas BIC* performed better for small Σθ with a 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 eighth to 12th grade (the first three studies) in the 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 et al., 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 mathematics subject, there are 40 items in 1988 and two factors emerged, 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 analysed using normal theory factor analysis more than two decades ago, we reanalyse the data using the proposed new methods. In addition, pooling together both mathematics and science domains results in potentially high-dimensional data. First, both GVEM and MHRM were conducted assuming the number of factors was 6. The focus is on the recovery of the correlation matrix \( \Sigma_{\theta} \) and its comparison between two methods. Since an exploratory item factor analysis was conducted, in both GVEM and MHRM we assumed that \( \Sigma_{\theta} = I_K \) during GVEM estimation and later performed the same promax rotation to estimate the correlation matrix \( \hat{\Sigma}_{\theta} \). Second, GVEM was used to explore the dimension of latent traits from the data.

Table 3 shows the estimated \( \Sigma_{\theta} \) from both methods assuming the number of factors is 6. The correlations in \( \hat{\Sigma}_{\theta} \) from the 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. Note that \( \hat{\Sigma}_{\theta} \) 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 the data, we applied the GVEM algorithm with the information criteria for dimension selection. Figure 8 presents the results of latent dimension selection under the 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* and BIC*, 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 and are merged. Comparing the information criteria values across both M2PL and M3PL, it appears that AIC* and BIC* 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 did not play a significant role in students’ performance on the mathematics and science cognitive test data.

### Table 3. Real data: comparison of estimated $\Sigma$

<table>
<thead>
<tr>
<th>GVEM</th>
<th>MHRM</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://example.com/table.png" alt="Table" /></td>
<td><img src="https://example.com/table.png" alt="Table" /></td>
</tr>
</tbody>
</table>

7. **Discussions**

Variational methods were first introduced in psychometrics by Rijmen and Jeon (2013) for high-dimensional IRT models with discrete latent traits, and later by Jeon et al. (2017) in

![Figure 8](https://example.com/figure.png)

**Figure 8.** Real data: BIC* for both M2PL and M3PL (AIC* shows the same trend).
the 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 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 closed-form updates at 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 by-products of the GVEM algorithm, both AIC\* and BIC\* 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 $A$ will have structural zeros, implying that certain items are irrelevant to certain factors. Similarly 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 zeros basically means multiplying $A$ by a conformable matrix of binary entries with 1s indicating that the corresponding element is estimable.

This work does not study the standard errors of the GVEM 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 and four-parameter IRT models (Meng, Xu, Zhang, & Tao, 2019) would be useful topics of future research.

**Acknowledgements**

This research is partially supported by National Science Foundation grants CAREER SES-1846747, SES-1659328, DMS-1712717 and Institute of Education Sciences grant R305D200015.

**Conflicts of interest**

All authors declare no conflict of interest.

**Author contributions**

April Cho (Formal analysis; Methodology; Visualization; Writing – original draft; Writing – review & editing) Chun Wang (Methodology; Project administration; Supervision; Writing – review & editing) Xue Zhang (Formal analysis; Writing – review & editing) Gongjun Xu
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 licence for this study. Data are available with the permission of Institute of Education Sciences.

References


April E. Cho et al.


Received 11 October 2019; revised version received 30 July 2020

**Supporting Information**

The following supporting information may be found in the online edition of the article:

The supplementary material presents the derivations of the algorithms and additional simulation results.