A NOTE ON IMPROVING VARIATIONAL ESTIMATION FOR MULTIDIMENSIONAL ITEM RESPONSE THEORY

CHENCHEN MA AND JING OUYANG

UNIVERSITY OF MICHIGAN

CHUN WANG

UNIVERSITY OF WASHINGTON

GONGJUN XU

UNIVERSITY OF MICHIGAN

Survey instruments and assessments are frequently used in many domains of social science. When the constructs that these assessments try to measure become multifaceted, multidimensional item response theory (MIRT) provides a unified framework and convenient statistical tool for item analysis, calibration, and scoring. However, the computational challenge of estimating MIRT models prohibits its wide use because many of the extant methods can hardly provide results in a realistic time frame when the number of dimensions, sample size, and test length are large. Instead, variational estimation methods, such as Gaussian variational expectation–maximization (GVEM) algorithm, have been recently proposed to solve the estimation challenge by providing a fast and accurate solution. However, results have shown that variational estimation methods may produce some bias on discrimination parameters during confirmatory model estimation, and this note proposes an importance-weighted version of GVEM (i.e., IW-GVEM) to correct for such bias under MIRT models. We also use the adaptive moment estimation method to update the learning rate for gradient descent automatically. Our simulations show that IW-GVEM can effectively correct bias with modest increase of computation time, compared with GVEM. The proposed method may also shed light on improving the variational estimation for other psychometrics models.

Key words: multidimensional item response theory, Gaussian variational em, importance sampling.

Developing, refining, and validating survey questionnaires that measure target latent traits such as personality or cognitive abilities has always been a core agenda in education and psychology, and this focus is also extended to health measurement and culminates in a multi-decade initiative on patient-reported outcome measures. Psychometric methods and tools are an integral part of achieving this focus. When the constructs that these assessments try to measure become increasingly complex, multidimensional item response theory (MIRT), also known as item factor analysis, provides a unified framework and convenient statistical tool for item analysis, calibration, and scoring. However, the increasing scale and complexity of survey designs, especially in large-scale assessments (LSA), require MIRT models with many latent factors. For instance, the English Language Proficiency Assessment for the 21st Century (ELPA21) across two gradebands consists of eight domain-level traits measured by more than 600 items (CRESST, 2017). The existing computational algorithms for fitting high-dimensional MIRT models are insufficient to navigate the massive amount of assessment data, reflected by excessively long computation time and unstable estimation results.

Correspondence should be made to Chun Wang, College of Education, University of Washington, 312 E Miller Hall, 2012 Skagit Lane, Seattle, WA 98105, USA. Email: wang4066@uw.edu

Correspondence should be made to Gongjun Xu, Department of Statistics, University of Michigan, 456 West Hall, 1085 South University, Ann Arbor, MI 48109, USA. Email: gongjun@umich.edu

MIRT provides a powerful tool for enriching the information gained in educational assessment (Hartig & Höhler, 2009). For instance, cognitive instructional psychology considers "science knowledge" and "mathematical ability" as highly differentiated theoretical constructs that consist of both basic facts and skills as well as deeper or higher-order understanding (Hamilton et al., 1995; Kupermintz et al., 1995). As another example, the 2003 assessment framework of PISA (OECD, 2003) contains a hierarchy of ability dimensions with general "knowledge and skills" at the highest level, followed by reading, math, science, and problem solving. Then at the lowest level are the sub-domains such as "space and shape," "change and relationships," and "quantity" nested within math. Hence, dimensions on different levels vary in their degree of generality and abstraction. Oftentimes, the highest level represents a broad competency level, whereas lower levels represent narrower and more specific abilities. If the intention is to model both the overall and lower-level abilities simultaneously, the model will be high dimensional (Briggs & Wilson, 2003).

Even though the research and development in statistics and psychometrics have provided increasingly sophisticated measurement models to better assess constructs in social sciences, the practice still lags behind (Cai & Hansen, 2018). Unidimensional IRT models continue to dominate the current applications in many domains. One reason is that when the number of items, sample size, and the number of dimensions are all large, the current computational algorithms for MIRT estimation may not be powerful enough to produce results in a reasonable time frame (or ever) (CRESST, 2017). For instance, due to the large number of students and items within each gradeband, the operational analysis approach used for ELPA21 is a two-step approach: in the first step, a unidimensional IRT model is fitted to the item response data for each domain subtest to obtain item parameter estimates; then in the second step, a restricted hierarchical model [i.e., testlet model, Wainer et al. (2007), Gibbons and Hedeker (1992), Cai et al. (2011)] is fitted to estimate the correlations between the four domains (Thissen, 2013). Such a two-step process has two limitations: (1) the item parameter calibration errors are ignored in the second step, and (2) the restricted hierarchical model is only an approximation to the independent-cluster MIRT model. Various full-information methods have been proposed to deal with the computational challenge, which are listed below with pros and cons. The list is by no means exhaustive, but it includes some of the most popular methods that are available in commercial software packages or R packages.[1]

1. Adaptive Gaussian quadrature. Compared to the regular Gauss-Hermite quadrature [e.g., Bock and Aitkin (1981)], even though the number of quadrature points per dimension is reduced, the total number of quadrature points still increases exponentially with the number of dimensions. Moreover, an extra step is needed to compute the posterior mode and variance of latent factors in each iteration, which adds additional computation costs (Pinheiro & Bates, 1995).

2. Monte Carlo techniques. This family of methods include, for instance, the Monte Carlo EM algorithm (McCulloch, 1997; Wang & Xu, 2015), stochastic EM algorithm (von Davier & Sinharay, 2010; Zhang et al., 2020), or Metropolis–Hastings Robbins–Monro algorithm (Cai 2010a,b). These methods circumvent intractable integrations by sampling from the posterior distributions; however, they may still computationally intensive for complicated high-dimensional models. Fully Bayesian estimation methods, such as Markov chain Monte Carlo (MCMC; Albert, 1992; Patz & Junker, 1999) can also be considered in this category. The Bayesian approach is also computationally costly as it needs a long chain to converge for complex models, though it is preferable with smaller sample sizes.

---

[1] The limited-information method such as weighted least squares is not reviewed here as it handles high-dimensional models very differently, and it cannot handle missing data very well.

3. Analytic dimension reduction. For models assuming certain conditional independence among factors (such as the bi-factor models), the conditional independence relations can be used to partition the joint space of all latent variables into smaller subsets. As a result, brute force numerical integration over the joint latent space can be replaced by a sequence of integrations over smaller subsets of latent variables, which helps reduce the computation burden dramatically. This strategy to deal with high-dimensional integration challenges is known as analytic dimension reduction (Cai et al., 2011; Gibbons & Hedeker, 1992; Rijmen et al., 2008). One limitation, though, is that the algebraic manipulations of the likelihood of a specific model might become very complicated, and they differ for different models (e.g., Cai et al., 2011; Gibbons & Hedeker, 1992). Hence, there is no universal rule that applies to any model.

4. Laplace approximation. This method is based on second-order Taylor expansion of the log-integrand around its mode (Lindstrom & Bates, 1988) such that the high-dimensional integral becomes tractable. This method is a classical and popularly used method for generalized linear mixed-effects models (GLMM), and it is available in many software packages, such as the "lem4" R package (Bates et al., 2014). However, this approximation may not be accurate when the dimension increases to 3 or higher, the sample size is small (Jeon et al., 2017), or the likelihood function is skewed.

Besides the full-information methods above, a recent constraint joint maximum likelihood estimation (CJMLE) was proposed by Chen et al. (2019), which is more computationally efficient than many marginal maximum likelihood methods, and the estimator has the theoretical guarantee to be consistent under high-dimensional settings. Extending CJMLE, the singular value decomposition (SVD)-based estimator was proposed by Zhang et al. (2020a), which further improves the performance of CJMLE. These joint maximum likelihood methods enjoy the low computational cost but sacrifice the flexibility of latent factors by treating them as fixed effects. For instance, it would be hard conceptually to generalize the algorithm to a multiple-group condition in which unbiased estimation of group-specific population distributions is often needed than estimation of individual person's latent trait as a fixed effect.

In light of the limitations of the above-mentioned methods, variational estimation methods that leverage advances in statistical and machine learning have recently gained increasing interests in psychometrics (Cho et al., 2021, 2022; Jeon et al., 2017). Among numerous variational estimation methods, Rijmen & Jeon (2013) was one of the first to use a variational estimation technique for MIRT models that approximates the likelihood function by a computationally tractable lower bound, but it only studied MIRT models with discrete latent factors. Later, a wide range of studies on variational methods were conducted for the estimation of more complex models (Hui et al., 2017; Natesan et al., 2016). Recently, Jeon et al. (2017) proposed variational maximization–maximization (VMM) algorithm for the generalized linear mixed models (GLMMs), which outperforms Laplace approximation with a small sample size. However, they rely on some iterative numerical algorithms to attain the solutions in each maximization step, resulting in a slow speed in running the algorithm. To further increase computational efficiency, many researchers brought up variational autoencoder (VAE), a deep learning-based variational method to tackle the estimation problems in MIRT models (Curi et al., 2019; Wu et al., 2020). Extending from VAE, the importance-weighted VAE (IW-VAE) is developed and exhibits competitive performances to other estimation methods (Liu et al., 2022; Urban & Bauer, 2021) at large sample sizes. However, the two IW-VAE methods lack theoretical support for the consistency of estimators. In addition, although they are powerful in handling large-scale data, their performances in small to medium-sample data may not be as well (see Appendix for more details). Cho et al. (2021, 2022) proposed a Gaussian variational expectation–maximization (GVEM) algorithm, which has shown to be computationally fast and produces comparable and sometimes more accurate param-

eter estimates than the MH-RM algorithm and than the CJMLE method in high-dimensional exploratory item factor analysis models (i.e., M2PL and M3PL in Cho et al., 2021). Moreover, Cho et al. (2021) proved that the estimated parameters from GVEM algorithm are consistent under the high-dimensional setting. However, we found that directly applying the GVEM algorithm in confirmatory MIRT models would generate relatively large bias on discrimination parameters, especially when the correlations among factors are high and the sample size is not large (please see Sect. 2 for the detailed simulation results). Such a bias issue happens commonly to variational estimation for various statistical models (Bishop, 2006).

To correct the bias in the variational algorithms for MIRT models, we propose an importance-weighted GVEM algorithm (denoted as IW-GVEM hereafter) , which is an extension of GVEM algorithm by performing additional steps after GVEM convergence. The primary idea is to use an importance-weighted variational inference technique to create a tighter variational lower bound to the target, otherwise intractable, marginal likelihood. Because the variational lower bound proposed in Cho et al. (2021, 2022) is replaced by a weighted average based on importance sampling (Domke & Sheldon, 2018), the desirable closed-form solution in the M-step is no longer applicable. Instead, we propose to use *Adam* (Kingma & Ba, 2014), a popular algorithm for first-order gradient-based optimization. This computationally efficient algorithm updates the objective function stochastically based on adaptive estimates of lower-order moments, and it is especially well suited for large data and complex models. Moreover, different from the IW-VAE methods rooted in deep neural network models where substantial theoretical works on the consistency of the estimators remain to be done, our proposed IW-GVEM is a more transparent method that comes with theoretical guarantees under the high-dimensional setting.

In what follows, this note briefly describes the M2PL model and the original GVEM algorithm and then introduces the IW-GVEM algorithm in Sect. 1, followed by a comprehensive simulation study in Sect. 2. We end the paper with discussions and future directions.

1. Methods

1.1. M2PL

Multidimensional 2PL model is one of the most widely used MIRT models in practice (Reckase, 2009). With M2PL, the item response function of the ith individual to the jth item is modeled by

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

where Y_{ij} for $i = 1, ..., N$ and $j = 1, ..., J$ is a binary response, $\boldsymbol{a}_j$ denotes a K-dimensional vector of item discrimination parameters for item j, and b_j specifies the corresponding difficulty level with item difficulty parameter as $b_j / \|\boldsymbol{a}_j\|_2$. Following notations in Cho et al. (2021), we use $\boldsymbol{Y}_i$ to denote the response vector of the ith subject, and $\boldsymbol{\theta}_i$ to denote the latent trait vector of the ith subject. We write $A = (\boldsymbol{\alpha}_j, j = 1, \ldots, J)$ and $B = (b_j, j = 1, \ldots, J)$. For model identification, oftentimes the means and variances of $\boldsymbol{\theta}$ are fixed as zeros and ones, respectively, and the covariance (which is actually correlation) of $\boldsymbol{\theta}$ is freely estimated.

1.2. GVEM

Let $\boldsymbol{\Delta} = (A, \boldsymbol{B}, \boldsymbol{\rho})$ denote the set of unknown parameters for M2PL, where $\boldsymbol{\rho}$ denotes the correlations of $\boldsymbol{\theta}$. As discussed, the population means of $\boldsymbol{\theta}$ are fixed at 0, and the population

variances are fixed at 1. The correlations among $\boldsymbol{\theta}$'s can be freely estimated. Then, the log-marginal likelihood of responses $\boldsymbol{Y}$ is

$$l(\boldsymbol{\Delta} \mid \boldsymbol{Y}) = \sum_{i=1}^{N} \log P(\boldsymbol{Y}_i \mid \boldsymbol{\Delta}) = \sum_{i=1}^{N} \log \int \prod_{j=1}^{J} P(Y_{ij} \mid \boldsymbol{\Delta}, \boldsymbol{\theta}_i) \phi(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i, \qquad (2)$$

where ϕ denotes a K-dimensional Gaussian distribution of $\boldsymbol{\theta}$ with mean 0 and covariance $\Sigma_{\boldsymbol{\theta}}$. It is the potentially high-dimensional integration in Eq. (2) that makes direct maximization of the log-marginal likelihood computationally prohibitive. The log-likelihood of response $\boldsymbol{Y}$ has an equivalent form

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

where $q_i(\boldsymbol{\theta}_i)$ can be any probability density function satisfying $\int_{\boldsymbol{\theta}_i} q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i = 1$.

The main idea behind variational inference is to approximate the intractable integral in Eq. (2) with a computationally feasible form, known as the evidence lower bound (ELBO; Blei et al., 2017; Ormerod & Wand, 2010). Because $P(\boldsymbol{Y}_i \mid \boldsymbol{\Delta}) = P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i \mid \boldsymbol{\Delta})/P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta})$, we write $l(\boldsymbol{\Delta} \mid \boldsymbol{Y})$ as

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

where $KL\{q_i(\boldsymbol{\theta}_i) \mid P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta})\} = \int_{\boldsymbol{\theta}_i} \log \frac{q_i(\boldsymbol{\theta}_i)}{P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta})} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$ is nonnegative. This is because

$$\begin{aligned} -KL\{q_i(\boldsymbol{\theta}_i) \mid P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta})\} &= \int_{\boldsymbol{\theta}_i} \log \frac{P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta})}{q_i(\boldsymbol{\theta}_i)} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &\leqslant \int_{\boldsymbol{\theta}_i} \left(\frac{P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta})}{q_i(\boldsymbol{\theta}_i)} - 1 \right) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &\leqslant \int_{\boldsymbol{\theta}_i} P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta}) d\boldsymbol{\theta}_i - \int_{\boldsymbol{\theta}_i} q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\ &= 1 - 1 = 0 \end{aligned}$$

Therefore, we have a lower bound of log-likelihood that

$$l(\boldsymbol{\Delta} \mid \boldsymbol{Y}) \geqslant \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \log \frac{P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i \mid \boldsymbol{\Delta})}{q_i(\boldsymbol{\theta}_i)} \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$$

$$= \sum_{i=1}^{N} E_{q_i(\boldsymbol{\theta}_i)} \left[\log \frac{P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i \mid \boldsymbol{\Delta})}{q_i(\boldsymbol{\theta}_i)} \right] =: ELBO, \tag{3}$$

where the last term $\sum_{i=1}^{N} E_{q_i(\boldsymbol{\theta}_i)} \left[\log \frac{P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i|\boldsymbol{\Delta})}{q_i(\boldsymbol{\theta}_i)} \right]$ is the ELBO for $l(\boldsymbol{\Delta}|\boldsymbol{Y})$ in Equation (2). Maximizing the log-marginal likelihood is then approximated by maximizing ELBO, and $q_i(\boldsymbol{\theta}_i)$, the variational distribution, needs to be carefully chosen to minimize the gap between the log-marginal likelihood and its ELBO.

The key is to find $q_i(\boldsymbol{\theta}_i)$ so that ELBO approximates the marginal likelihood $l(\boldsymbol{\Delta}|\boldsymbol{Y})$ as close as possible. Note that when $q_i(\boldsymbol{\theta}_i)$ is the posterior density of $\boldsymbol{\theta}_i$, i.e., $q_i(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i \mid \boldsymbol{Y}_i, \boldsymbol{\Delta})$, maximizing ELBO is equivalent to Bock & Aitkin (1981)'s marginal maximum likelihood/expectation–maximization (MML/EM) algorithm. Instead, as the choice of $q_i(\boldsymbol{\theta}_i)$ determines the computational cost and success of the algorithm, Cho et al. (2021, 2022) proposed a choice of $q_i(\boldsymbol{\theta}_i)$ that satisfied two criteria: (1) it is easy to maximize, and (2) it approximates the true log-marginal likelihood well. Due to the independence of the students' responses in general IRT models, $q_i(\boldsymbol{\theta}_i)$ is selected for each individual separately. Specifically, under M2PL, the joint distribution of $\boldsymbol{\theta}_i$ and $\boldsymbol{Y}_i$ is,

$$\log P(Y_i, \boldsymbol{\theta}_i \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})$$

$$= \sum_{j=1}^{J} \left\{ Y_{ij}(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + \log \frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} \right\} + \log \phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)$$

$$\geq \sum_{j=1}^{J} \log \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} + \sum_{j=1}^{J} Y_{ij}(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + \sum_{j=1}^{J} \frac{b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \xi_{ij}}{2}$$

$$- \sum_{j=1}^{J} \eta(\xi_{ij})\{(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \xi_{ij}^2\} + \log \phi_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \tag{4}$$

$$:= l(Y_i, \boldsymbol{\theta}_i \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho}, \xi_{ij}), \tag{5}$$

where ξ_{ij} is the variational parameter for the ith subject, which will be updated iteratively in the M-step of GVEM, and $\eta(\xi_{ij}) = (2\xi_{i,j})^{-1}[e^{\xi_{i,j}}/(1 + e^{\xi_{i,j}}) - 1/2]$. The derivation is as follows. Because the difficulty of handling the marginal distribution of $P(\boldsymbol{Y}_i)$ mostly comes from the logistic sigmoid function, which makes the integration over θ not a closed form in the E-step. As a result, Cho et al. (2021) used a local variational approximation method (Jordan, 2004). Denote $x_{ij} = b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i$, the local variational method gives the following variational lower bound for the sigmoid function:

$$\frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} = \frac{\exp(x_{ij})}{1 + \exp(x_{ij})}$$

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

$$\geqslant \frac{\exp(\xi_{ij})}{1 + \exp(\xi_{ij})} \exp \left\{ \frac{x_{ij} - \xi_{ij}}{2} - \eta(\xi_{ij})(x_{ij}^2 - \xi_{ij}^2) \right\},$$

and by applying the above lower bound to Eq. (4), we get Eq. (5), which provides a variational lower bound for $\log P(Y_i, \boldsymbol{\theta}_i \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})$.

By variational inference theory, we can show that the variational distributions $q_i(\boldsymbol{\theta}_i)$ (for $i = 1, ..., N$) that minimize the distances between the lower bound and the joint distribution follow a Gaussian distribution with closed-form mean and variance, i.e., $q_i(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \mu_i, \Sigma_i)$ where the mean parameter of the normal distribution is

$$\mu_i = \Sigma_i \times \sum_{j=1}^{J} \left\{ 2\eta(\xi_{i,j})b_j + Y_{ij} - \frac{1}{2} \right\} \boldsymbol{\alpha}_j \tag{6}$$

and the covariance matrix is

$$\left(\Sigma_i \right)^{-1} = \left(\Sigma_{\boldsymbol{\theta}} \right)^{-1} + 2 \sum_{j=1}^{J} \eta(\xi_{i,j}) \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^{\top}. \tag{7}$$

In the confirmatory model estimation, we update population covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ by

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\top}). \tag{8}$$

But because we need to fix the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ during estimation to fix the scale, we propose to rescale $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ after the M-step converges, i.e.,

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^* = \left(\left(\sqrt{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})} \right)^{-1} \right)^{\top} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left(\sqrt{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})} \right)^{-1},$$

and the discrimination parameter needs to be rescaled accordingly, i.e., $\boldsymbol{\alpha}_j^* = \boldsymbol{\alpha}_j \sqrt{\mathrm{diag}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})}$. For the exploratory analysis, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is fixed at an identity matrix during estimation, and a post hoc rotation will then produce proper nonzero correlations. In the following, we assume that the GVEM algorithm has converged and we fix the variational parameter ξ_{ij} as the final estimates. In other words, we do not update ξ_{ij} in the later iterative steps and ξ_{ij} is fixed at the initialization GVEM step in Algorithm 1.

1.3. Importance Sampling

Referring back to the basic idea underlying variational inference, i.e., the ELBO for log-likelihood of response $l(\boldsymbol{\Delta} \mid \boldsymbol{Y})$ in the inequality (3), it can be seen that a tighter lower bound is attained when $R \equiv P(\boldsymbol{Y}_i, \boldsymbol{\theta}_i \mid \boldsymbol{\Delta})/q_i(\boldsymbol{\theta}_i)$ around its mean $P(\boldsymbol{Y}_i \mid \boldsymbol{\Delta})$. Therefore, we can consider different random variables with the same mean that are more concentrated. For example, we can draw M i.i.d. samples from $q(\boldsymbol{z})$, and average the estimates as in importance sampling (IS):

$$R_M = \frac{1}{M} \sum_{m=1}^{M} R_m = \frac{1}{M} \sum_{m=1}^{M} \frac{p(\boldsymbol{x}, z_m)}{q(z_m)}, \quad z_m \sim q(\cdot). \tag{9}$$

This leads to a tighter "importance-weighted ELBO" (IW-ELBO) on $\log P(\boldsymbol{x})$,

$$\text{IW-ELBO}_M = E_{q(\boldsymbol{Z})} \left[\log \frac{1}{M} \sum_{m=1}^{M} \frac{p(z_m, \boldsymbol{x})}{q(z_m)} \right] := \mathcal{L}_M(\boldsymbol{x}). \tag{10}$$

It is shown that $\mathcal{L}_M(\boldsymbol{x})$ converge to $\log p(\boldsymbol{x})$ as M goes to infinity (Burda et al., 2015), which is summarized in the following result.

Proposition 1. *For all M, the lower bounds satisfy*

$$\log p(\boldsymbol{x}) \geq \mathcal{L}_{M+1} \geq \mathcal{L}_M.$$

Moreover, if $p(\boldsymbol{x}, \boldsymbol{z})/q(\boldsymbol{z}|\boldsymbol{x})$ is bounded, then $\mathcal{L}_M$ approaches $\log p(\boldsymbol{x})$ as M goes to infinity.

Motivated by this result, we use the importance sampling method and calculate the derivatives of $\mathcal{L}_M$ to further perform gradient-based optimization. Specifically, denote $w_m = p(\boldsymbol{x}, \boldsymbol{z}_m)/q(\boldsymbol{z}_m)$, then the derivatives of $\mathcal{L}_M$ with respect to $\boldsymbol{\theta}$ are

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_M(\boldsymbol{x}) = \nabla_{\boldsymbol{\theta}} E_{q(\boldsymbol{Z})} \left[\log \frac{1}{M} \sum_{m=1}^{M} w_m \right]$$

$$= E_{q(\boldsymbol{Z})} \left[\nabla_{\boldsymbol{\theta}} \log \frac{1}{M} \sum_{m=1}^{M} w_m \right]$$

$$= E_{q(\boldsymbol{Z})} \left[\sum_{m=1}^{M} \tilde{w}_m \nabla_{\boldsymbol{\theta}} \log w_m \right],$$

where $\tilde{w}_m = w_m / \sum_{m'=1}^{M} w_{m'}$ and

$$\nabla_{\boldsymbol{\theta}} \log w_m = \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x}, \boldsymbol{z}_m) - \nabla_{\boldsymbol{\theta}} \log q(\boldsymbol{z}_m). \tag{11}$$

1.4. IW-GVEM

The primary idea of IW-GVEM is to replace Eq. (3) with importance-weighted ELBO as in Eq. (10). That is, for each $i = 1, \ldots, N$, we draw M samples from $q_i(\boldsymbol{\theta}_i)$ for S times:

$$\boldsymbol{\theta}_i^{(s,m)} \sim q_i(\boldsymbol{\theta}_i), \text{ for } s = 1, \ldots, S, m = 1, \ldots, M.$$

Define $w_i^{(s,m)} = p(Y_i, \boldsymbol{\theta}_i^{(s,m)})/q_i(\boldsymbol{\theta}_i^{(s,m)})$, where $p(Y_i, \boldsymbol{\theta}_i^{(s,m)}) = P(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})$ as in Eq. (4), and $q_i(\boldsymbol{\theta}_i^{(s,m)}) \sim N(\boldsymbol{\theta}_i^{(s,m)} \mid \mu_i, \Sigma_i)$, then $\mathcal{L}_M(\boldsymbol{Y})$ can be approximated by

$$\mathcal{L}_M(\boldsymbol{Y}) \approx \sum_{i=1}^{N} \left(\frac{1}{S} \sum_{s=1}^{S} \left[\log \frac{1}{M} \sum_{m=1}^{M} w_i^{(s,m)} \right] \right).$$

Note $w_i^{(s,m)}$ is a function of parameters $(\xi_i, \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})$.

To learn parameters, we use a stochastic gradient ascent method, which needs to calculate the gradients of $\mathcal{L}_M(\boldsymbol{Y})$. Based on Eq. (11), the gradients can be approximated by

$$\nabla_{\boldsymbol{\alpha}} \mathcal{L}_M(\boldsymbol{Y}) \approx \sum_{i=1}^{N} \left(\frac{1}{S} \sum_{s=1}^{S} \sum_{m=1}^{M} \tilde{w}_i^{(s,m)} \nabla_{\boldsymbol{\alpha}} \left[\log P(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho}) - \nabla_{\boldsymbol{\alpha}} \log q_i(\boldsymbol{\theta}_i^{(s,m)} \mid Y_i) \right] \right),$$

where $\tilde{w}_i^{(s,m)} = w_i^{(s,m)} / \sum_{m'=1}^{M} w_i^{(s,m')}$. Note that $q_i(\theta_i^{(s,m)} \mid Y_i)$ does not depend on the parameters in the current iteration. Therefore, we only need to calculate $\tilde{w}_i^{(s,m)}$ and $\nabla_{\boldsymbol{\alpha}} P(Y_i, \theta_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})$. Similarly we can calculate $\nabla_{\boldsymbol{b}} \mathcal{L}_M(\boldsymbol{Y})$ and $\nabla_{\Sigma_{\theta}} \mathcal{L}_M(\boldsymbol{Y})$. Specifically, we have

$$
\begin{aligned}
\nabla_{\boldsymbol{\alpha}_j} \log P\left(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho}\right) &= \tilde{w}_i^{(s,m)} \nabla_{\boldsymbol{\alpha}_j}\left[\log P(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})\right] \\
&= \tilde{w}_i^{(s,m)}\left[\left(Y_{ij} - 1 + \frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i^{(s,m)} - b_j)}\right) \boldsymbol{\theta}_i^{(s,m)}\right],
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{b_j} \log P\left(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho}\right) &= \tilde{w}_i^{(s,m)} \nabla_{b_j}\left[\log P(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})\right] \quad (12) \\
&= \tilde{w}_i^{(s,m)}\left[1 - Y_{ij} - \frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i^{(s,m)} - b_j)}\right],
\end{aligned}
$$

$$
\nabla_{\Sigma_{\theta}} \log P\left(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho}\right) = \tilde{w}_i^{(s,m)} \nabla_{\Sigma_{\theta}}\left[\log P(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})\right] \quad (13)
$$

$$
= \tilde{w}_i^{(s,m)}\left[\frac{1}{2}\Sigma_{\theta} - \frac{1}{2}\boldsymbol{\theta}_i^{(s,m)}(\boldsymbol{\theta}_i^{(s,m)})^\top\right]. \quad (14)
$$

To summarize, in the $(t+1)th$ iteration, we perform the following:

1. For $i = 1, \ldots, N$, draw M samples from $q_i(\boldsymbol{\theta}_i)$ for S times.
2. Calculate $w_i^{(s,m)} = P(Y_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho})/q_i(\boldsymbol{\theta}_i^{(s,m)})$ and $\tilde{w}_i^{(s,m)}$ $= w_i^{(s,m)} / \sum_{m'=1}^{M} w_i^{(s,m')}$.
3. Calculate the gradients according to Eqs. (12), (13), and (14).

Proper learning rate scheduling is important in gradient-based algorithms. In this work, we apply the **Ada**ptive **m**oment estimation (Adam) method (Kingma & Ba, 2014), which has been extensively used in deep learning research and applications, to adjust the learning rate in our training process. In Adam, we compute individual adaptive learning rates for each parameter from estimates of the first and second moments of the gradients. Specifically in the tth iteration, we calculate exponential moving averages of the gradient (denoted as $\boldsymbol{v}_t$) and the squared gradient (denoted as $\boldsymbol{s}_t$) with exponential decay rates β_1 and β_2, respectively. The moving averages can be seen as estimates of the first and second moments of the gradients. Then, we correct these biased exponential moving averages by $1 - \beta_1^t$ and $1 - \beta_2^t$, respectively, and update parameters using standardized gradients. The concrete steps of generic Adam are provided below, where $\boldsymbol{g}_t$ is the gradient (corresponding to that in Eqs. (12), (13) and (14), respectively) in the tth iteration:

1. $\boldsymbol{v}_t = \beta_1 \boldsymbol{v}_{t-1} + (1 - \beta_1)\boldsymbol{g}_t$ (update biased first moment estimate)
2. $\boldsymbol{r}_t = \beta_2 \boldsymbol{r}_{t-1} + (1 - \beta_2)\boldsymbol{g}_t^2$ (update biased second moment estimate)
3. $\hat{\boldsymbol{v}}_t = \boldsymbol{v}_t/(1 - \beta_1^t)$, $\hat{\boldsymbol{r}}_t = \boldsymbol{r}_t/(1 - \beta_2^t)$ (compute bias-corrected moment estimates)
4. $\hat{\boldsymbol{g}}_t = \eta\hat{\boldsymbol{v}}_t/(\sqrt{\hat{\boldsymbol{r}}_t} + \epsilon)$, where η is learning rate (update the final gradient)

With this, the proposed importance-weighted Gaussian variational EM (IW-GVEM) algorithm is summarized in Algorithm 1. For the choice of hyperparameters, we follow the suggestions in Kingma & Ba (2014) and adopt the default setting that $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Empirically in our simulation studies, for better convergence performance, we let the learning rate of Σ_{θ} to be 0.1η while the learning rate for $\boldsymbol{a}$ and $\boldsymbol{b}$ to be η, and we search for an optimal learning rate η with the maximum ELBO over a list $\{0.01, 0.05, 0.1, 0.5\}$. Lastly, we set $\epsilon = 0.001$.

In terms of convergence criteria, we evaluate the Euclidean norm of the difference between the estimated parameters of the current step and those of the previous step. When the difference is less

---

Algorithm 1: IW-GVEM for M2PL

Data: Binary response matrix $Y \in \{0, 1\}^{N \times J}$.

Run GVEM algorithm and obtain $\boldsymbol{\mu}_{i,\mathrm{GV}}$, $\boldsymbol{\Sigma}_{i,\mathrm{GV}}$, $\boldsymbol{\alpha}_{\mathrm{GV}}$, $\boldsymbol{b}_{\mathrm{GV}}$, $\boldsymbol{\Sigma}_{\theta,\mathrm{GV}}$, and ξ_{ij}. These values will serve as initial values for IW-GVEM.

Set hyperparameters S, M for importance sampling, and β_1, β_2, η, and ϵ for Adam.

Set $\boldsymbol{v}_{\boldsymbol{\alpha}_j}^{(0)} = \boldsymbol{0}$, $\boldsymbol{v}_{\boldsymbol{b}_j}^{(0)} = \boldsymbol{0}$, $\boldsymbol{v}_{\boldsymbol{\Sigma}_\theta}^{(0)} = \boldsymbol{0}$, $\boldsymbol{r}_{\boldsymbol{\alpha}_j}^{(0)} = \boldsymbol{0}$, $\boldsymbol{r}_{\boldsymbol{b}_j}^{(0)} = \boldsymbol{0}$, $\boldsymbol{r}_{\boldsymbol{\Sigma}_\theta}^{(0)} = \boldsymbol{0}$.

while *not converged* **do**

    In the t-th iteration,

    for $i \in [N]$ **do**

        draw M samples from $q_i(\boldsymbol{\theta}_i) = N(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_{i,\mathrm{GV}}, \boldsymbol{\Sigma}_{i,\mathrm{GV}})$ for S times.

    for $i \in [N]$, $s \in [S]$ *and* $m \in [M]$ **do**

        $w_i^{(s,m)} = p(\boldsymbol{Y}_i, \boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\rho}) / q_i(\boldsymbol{\theta}_i^{(s,m)})$, $\tilde{w}_i^{(s,m)} = w_i^{(s,m)} / \sum_{m'=1}^{M} w_i^{(s,m')}$.

    for $j \in [J]$ **do**

$$\boldsymbol{g}_{\boldsymbol{\alpha}_j} = \sum_{i=1}^{N} \left(\frac{1}{S} \sum_{s=1}^{S} \sum_{m=1}^{M} \tilde{w}_i^{(s,m)} \big[Y_{ij} - 1 + 1/\big(1 + \exp\{\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i^{(s,m)} - b_j\}\big) \big] \boldsymbol{\theta}_i^{(s,m)} \right),$$

$$\boldsymbol{g}_{b_j} = \sum_{i=1}^{N} \left(\frac{1}{S} \sum_{s=1}^{S} \sum_{m=1}^{M} \tilde{w}_i^{(s,m)} \big[1 - Y_{ij} - 1/\big(1 + \exp\{\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i^{(s,m)} - b_j\}\big) \big] \right).$$

$$\boldsymbol{g}_{\boldsymbol{\Sigma}_\theta} = \sum_{i=1}^{N} \left(\frac{1}{S} \sum_{s=1}^{S} \sum_{m=1}^{M} \tilde{w}_i^{(s,m)} \big[\boldsymbol{\Sigma}_\theta - \boldsymbol{\theta}_i^{(s,m)} (\boldsymbol{\theta}_i^{(s,m)})^\top \big] / 2 \right).$$

    for $j \in [J]$ **do**

$$\boldsymbol{v}_{\boldsymbol{\alpha}_j}^{(t)} = \beta_1 \boldsymbol{v}_{\boldsymbol{\alpha}_j}^{(t-1)} + (1 - \beta_1) \boldsymbol{g}_{\boldsymbol{\alpha}_j}, \quad \boldsymbol{r}_{\boldsymbol{\alpha}_j}^{(t)} = \beta_2 \boldsymbol{r}_{\boldsymbol{\alpha}_j}^{(t-1)} + (1 - \beta_2) \boldsymbol{g}_{\boldsymbol{\alpha}_j} \cdot \boldsymbol{g}_{\boldsymbol{\alpha}_j},$$

$$\boldsymbol{v}_{\boldsymbol{\alpha}_j}^{(t)} = \boldsymbol{v}_{\boldsymbol{\alpha}_j}^{(t)} / (1 - \beta_1^t), \quad \boldsymbol{r}_{\boldsymbol{\alpha}_j}^{(t)} = \boldsymbol{r}_{\boldsymbol{\alpha}_j}^{(t)} / (1 - \beta_2^t),$$

$$\boldsymbol{v}_{\boldsymbol{b}_j}^{(t)} = \beta_1 \boldsymbol{v}_{\boldsymbol{b}_j}^{(t-1)} + (1 - \beta_1) \boldsymbol{g}_{\boldsymbol{b}_j}, \quad \boldsymbol{r}_{\boldsymbol{b}_j}^{(t)} = \beta_2 \boldsymbol{v}_{\boldsymbol{b}_j}^{(t-1)} + (1 - \beta_2) \boldsymbol{g}_{\boldsymbol{b}_j} \cdot \boldsymbol{g}_{\boldsymbol{b}_j},$$

$$\boldsymbol{v}_{\boldsymbol{b}_j}^{(t)} = \boldsymbol{v}_{\boldsymbol{b}_j}^{(t)} / (1 - \beta_1^t), \quad \boldsymbol{r}_{\boldsymbol{b}_j}^{(t)} = \boldsymbol{r}_{\boldsymbol{b}_j}^{(t)} / (1 - \beta_2^t).$$

$$\boldsymbol{v}_{\boldsymbol{\Sigma}_\theta}^{(t)} = \beta_1 \boldsymbol{v}_{\boldsymbol{\Sigma}_\theta}^{(t-1)} + (1 - \beta_1) \boldsymbol{g}_{\boldsymbol{\Sigma}_\theta}, \quad \boldsymbol{r}_{\boldsymbol{\Sigma}_\theta}^{(t)} = \beta_2 \boldsymbol{r}_{\boldsymbol{\Sigma}_\theta}^{(t-1)} + (1 - \beta_2) \boldsymbol{g}_{\boldsymbol{\Sigma}_\theta} \cdot \boldsymbol{g}_{\boldsymbol{\Sigma}_\theta},$$

$$\boldsymbol{v}_{\boldsymbol{\Sigma}_\theta}^{(t)} = \boldsymbol{v}_{\boldsymbol{\Sigma}_\theta}^{(t)} / (1 - \beta_1^t), \quad \boldsymbol{r}_{\boldsymbol{\Sigma}_\theta}^{(t)} = \boldsymbol{r}_{\boldsymbol{\Sigma}_\theta}^{(t)} / (1 - \beta_2^t).$$

    for $j \in [J]$ **do**

$$\hat{\boldsymbol{g}}_{\boldsymbol{\alpha}_j} = \eta \boldsymbol{v}_{\boldsymbol{\alpha}_j}^{(t)} / \left(\sqrt{\boldsymbol{r}_{\boldsymbol{\alpha}_j}^{(t)}} + \epsilon \right), \quad \hat{\boldsymbol{\alpha}}_j^{(t)} = \hat{\boldsymbol{\alpha}}_j^{(t-1)} + \hat{\boldsymbol{g}}_{\boldsymbol{\alpha}_j},$$

$$\hat{\boldsymbol{g}}_{\boldsymbol{b}_j} = \eta \boldsymbol{v}_{\boldsymbol{b}_j}^{(t)} / \left(\sqrt{\boldsymbol{r}_{\boldsymbol{b}_j}^{(t)}} + \epsilon \right), \quad \hat{\boldsymbol{b}}_j^{(t)} = \hat{\boldsymbol{b}}_j^{(t-1)} + \hat{\boldsymbol{g}}_{\boldsymbol{b}_j}.$$

$$\hat{\boldsymbol{g}}_{\boldsymbol{\Sigma}_\theta} = \eta \boldsymbol{v}_{\boldsymbol{\Sigma}_\theta}^{(t)} / \left(\sqrt{\boldsymbol{r}_{\boldsymbol{\Sigma}_\theta}^{(t)}} + \epsilon \right), \quad \hat{\boldsymbol{\Sigma}}_\theta^{(t)} = \hat{\boldsymbol{\Sigma}}_\theta^{(t-1)} + \hat{\boldsymbol{g}}_{\boldsymbol{\Sigma}_\theta}.$$

Output: $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{\Sigma}}_\theta$.

---

than a certain tolerance value, the algorithm is stopped. For our simulation studies, in obtaining the initial model parameter using the GVEM algorithm, we reach convergence at $(l+1)$th iteration if $\|\boldsymbol{\alpha}_{GV}^{l+1} - \boldsymbol{\alpha}_{GV}^l\|_2 + \|\boldsymbol{b}_{GV}^{l+1} - \boldsymbol{b}_{GV}^l\|_2 + \|\boldsymbol{\Sigma}_{\theta,GV}^{l+1} - \boldsymbol{\Sigma}_{\theta,GV}^l\|_2 \leqslant 0.0001$. In IW-GVEM, we reach

convergence at $(t+1)$th iteration when $\max\{\|\boldsymbol{\alpha}^{t+1}-\boldsymbol{\alpha}^t\|_2, \|\boldsymbol{b}^{t+1}-\boldsymbol{b}^t\|_2, \|\boldsymbol{\Sigma}_\theta^{t+1}-\boldsymbol{\Sigma}_\theta^t\|_2\} \leqslant 0.0001$ or the iteration stops when it reaches certain maximum iteration number.

2. Simulation Studies

2.1. Design

We conducted comprehensive simulation studies to evaluate the performance of the proposed method under various manipulated conditions. We follow similar designs as in Cho et al. (2021) and consider different settings: (1) sample size: N = 200 or 500; (2) number of domains: K = 2 or 5; (3) test length: J = 30 if K = 2 or J = 55 if K = 5; (4) both within and between multidimensional structures; (5) factor correlations: low correlation $r \sim$ unif(0.1, 0.3) or high correlation $r \sim$ unif(0.5, 0.7); and (6) confirmatory or exploratory analysis.

Similar to Cho et al. (2021), for the between-item multidimensional structure, we had equal numbers of items loaded on each factor. For the within-item multidimensional structure, when K = 2, about one-third of the items were loaded onto the first, or the second, or both factors, respectively. In the cases where K = 5, there were about one-third of the items loaded onto one, two, or three factors, respectively. For the model parameters, we simulated the item discrimination parameters $\alpha_{j,k}$ from uniform distribution on [1, 2], and difficulty parameter b_j from the standard normal distribution. We generated the latent traits $\boldsymbol{\theta}_j$ from multivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma}_\theta)$, where the diagonal elements of $\boldsymbol{\Sigma}_\theta$ were all 1 and off-diagonal elements were generated from uniform distributions. Specifically, in high-correlation settings, the uniform distribution was set to be unif(0.5, 0.7), whereas in the low-correlation settings we set it to be unif(0.1, 0.3).

For evaluation, we compared the bias and root-mean-squared errors (RMSEs) of model parameters, as well as computation time between GVEM and IW-GVEM. For exploratory analysis, we did a promax rotation after model convergence, and compared the rotated parameters to the true values (Cho et al., 2022). For IW-GVEM, we first ran GVEM algorithm to get initial estimates of model parameters, and then ran several gradient descent steps using importance sampling to correct the bias. To select a proper initial learning rate for the gradient algorithm, we first sampled a set of data aside based on the GVEM estimates. After we got model parameter estimates using importance sampling, we calculated the lower bound as in our objective function based on the previously sampled data set, and chose the learning rate corresponding to the largest lower bound. In the simulation studies, we set S and M to be 10. Our empirical experiments have shown that increasing S and M did not result in significant improvements and 10 was large enough for the simulation settings. The results were averaged over 100 repetitions.

2.2. Results

Figures 1 and 2 present the bias and RMSE of confirmatory M2PL model when K = 2. Note that in confirmatory analysis, there are discrimination parameters specified to be zeros. These zero-constrained terms are excluded in the bias and RMSE computation. The two separately colored boxes represent the distribution of respective criteria across 100 replications from IW-GVEM (denoted as "IS" in the figure) and the original GVEM algorithm. As shown, GVEM already performs well by producing close to 0 bias for $\boldsymbol{b}$ and $\boldsymbol{\Sigma}_\theta$. It is the discrimination parameter, $\boldsymbol{\alpha}$, that has a non-ignorable bias. The IW-GVEM algorithm effectively corrects such bias on $\boldsymbol{\alpha}$ across all conditions without deteriorating the estimation accuracy of other parameters. And because the bias is corrected, the RMSE of $\boldsymbol{\alpha}$ is also smaller consistently compared to that from GVEM, whereas again, there is no appreciable difference between IW-GVEM and GVEM in terms of RMSE on $\boldsymbol{b}$ and $\boldsymbol{\Sigma}_\theta$. Sorting through the manipulated conditions, it is "within-item" multidimensional
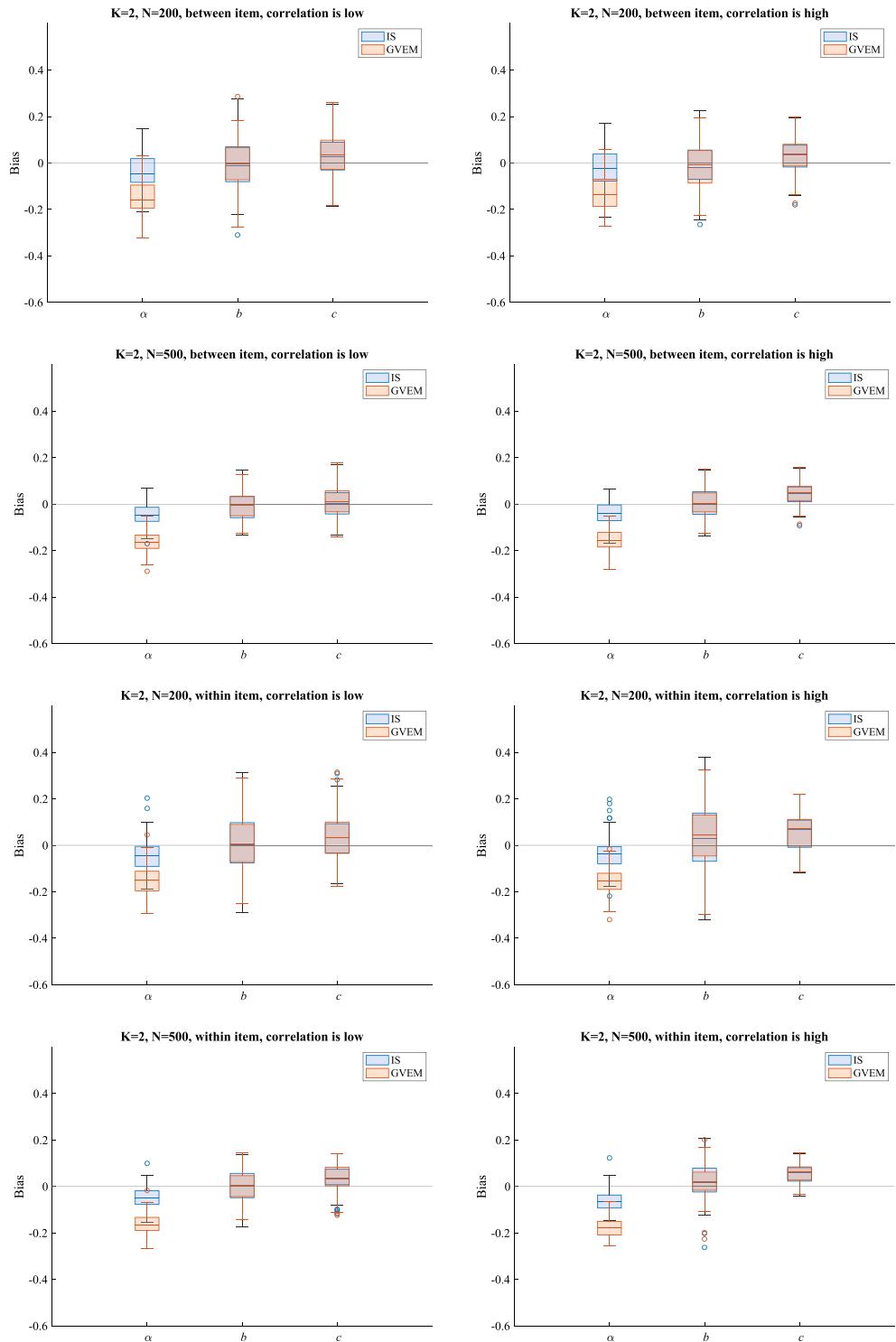
FIGURE 1.
Bias for $K = 2$ under confirmatory analysis.
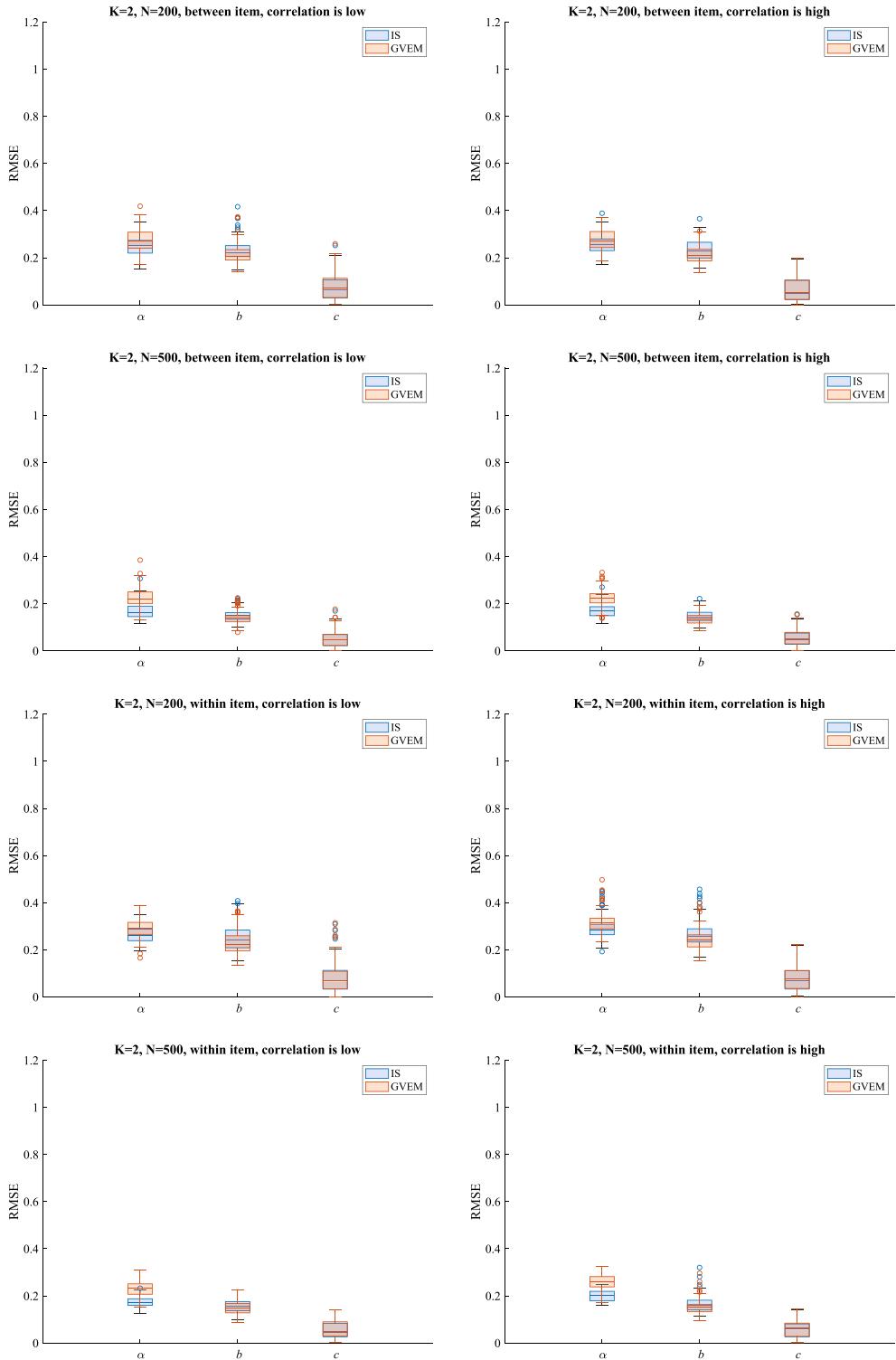
FIGURE 2.
RMSE for $K = 2$ under confirmatory analysis.

structure in combination with high factor correlation tends to yield larger RMSE for both methods and all parameters.

Figures 3 and 4 present the bias and RMSE of confirmatory M2PL model when $K = 5$. The trend observed from the $K = 2$ condition continues to hold here. That is, IW-GVEM can correct bias on $\boldsymbol{\alpha}$ effectively and hence also brings down its RMSE, whereas bias on the other parameters is already close to 0 from both methods and their RMSEs are also comparable. Increasing the number of dimensions certainly makes the model estimation harder to converge, and the estimates are also more variable, especially for $\boldsymbol{b}$ and $\boldsymbol{\Sigma}_\theta$, as reflected by wider boxes for those parameters in Fig. 3.

Figures 5, 6, 7, and 8 presents the results from exploratory estimation condition, in the same order as before. For the exploratory M2PL model estimation, GVEM generally performs well and the bias on $\boldsymbol{\alpha}$ is already small to begin with. This is consistent with the results reported in literature (Cho et al., 2021, 2022). Even so, under all settings, the RMSEs of IW-GVEM are still smaller than or equal to that of GVEM. IW-GVEM can still further bring down the bias of $\boldsymbol{\alpha}$ to near 0 for most cases. The exceptional case when the bias of $\boldsymbol{\alpha}$ from IW-GVEM is larger than the bias from GVEM is for the "within item, correlation is high" condition. This case is the most difficult case where the items were loaded on factors via a more complicated setting and the correlations among factors are relatively high. Nonetheless, this special case has overall good estimation performance as the estimation bias from IW-GVEM is still close to the bias from GVEM, and the RMSE from IW-GVEM is lower than the RMSE from GVEM. In addition, when $K = 2$ the bias of $\boldsymbol{\Sigma}_\theta$ appears to depart from 0 and IW-GVEM does not correct for such bias, although the RMSE of $\boldsymbol{\Sigma}_\theta$ is kept small across the board. The bias of $\boldsymbol{\Sigma}_\theta$ gets closer to 0 when K increases and when the factor correlation is low. Because in the exploratory estimation mode, specific types of rotations will affect resulting factor correlations, the bias in $\boldsymbol{\Sigma}_\theta$ estimation is less of a concern. Although the increase in the number of dimensions K could lead to a more complicated model and bring challenges to parameter estimation, the increase in test length, on the other hand, improves the estimation accuracy of parameters. Specifically, at $K = 5$, we use test length $J = 55$ which is greater than $J = 30$ at $K = 2$. This increase in test length explains the results that the biases at $K = 5$ are closer to 0 than that at $K = 2$ for some cases. Overall, the results from GVEM and IW-GVEM are very close.

Table 1 presents the computation time for confirmatory M2PL estimation under both GVEM and IW-GVEM algorithms. Understandably, IW-GVEM takes longer time under all conditions because both the important sampling step and the gradient descent optimization are time-consuming compared to closed-form updates in GVEM. Unsurprisingly, both methods need longer time for larger sample sizes. It is more interesting to note that, other things being equal, when the multidimensional structure is "within-item," GVEM almost doubles (when $K = 2$) or sometimes even triples (when $K = 5$) the computation time compared to the "between-item" condition. But for IW-GVEM, the computation time is rather stable across these two multidimensional structures. Similarly, high correlation among factors is known to be more challenging, hence computation time increases by about 50% or more for GVEM from low to high-correlation conditions, but the computation time of IW-GVEM seems to be unaffected. These all suggest that IW-GVEM is better suited for more complex models. The same patterns remain for the exploratory M2PL estimation, as shown in Table 2, although exploratory analysis in general takes longer time than confirmatory analysis, simply because more parameters are needed to be updated simultaneously.

3. Discussion

In this note, we proposed an importance-weighted version of GVEM to correct its bias on the $\boldsymbol{\alpha}$ estimates in the confirmatory M2PL models. Because the evidence lower bound (ELBO),
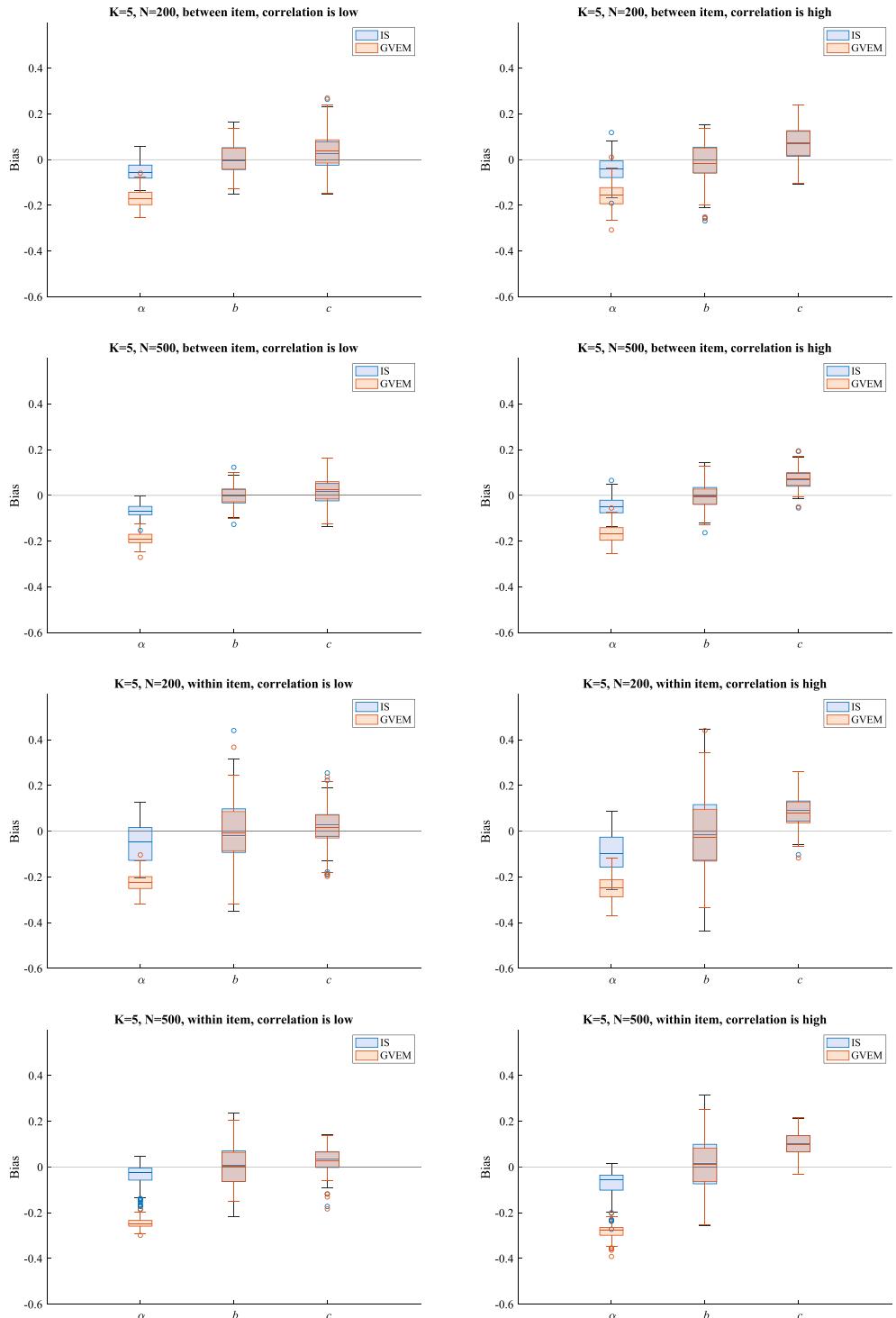
FIGURE 3.
Bias for $K = 5$ under confirmatory analysis.
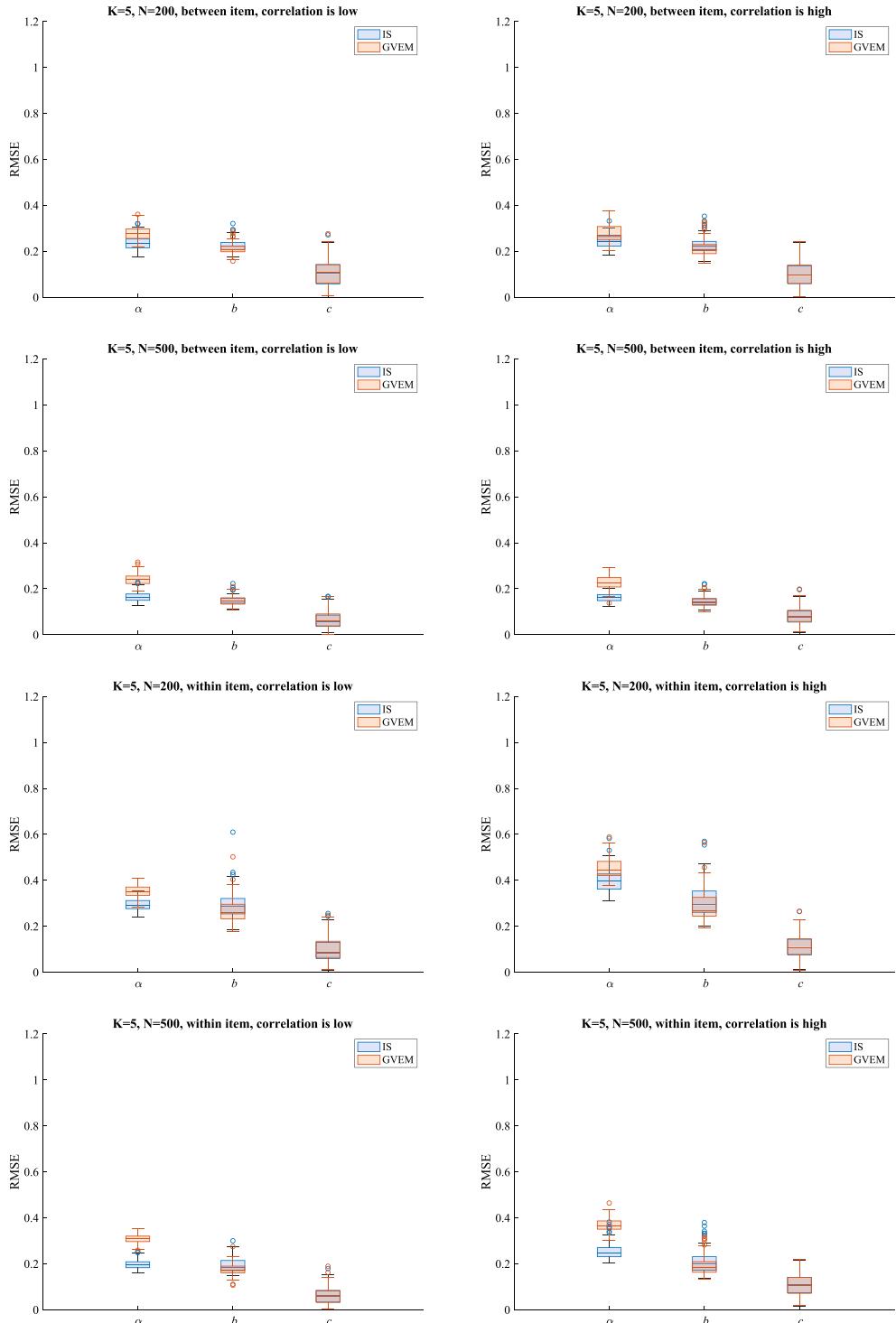
FIGURE 4.
RMSE for $K = 5$ under confirmatory analysis.
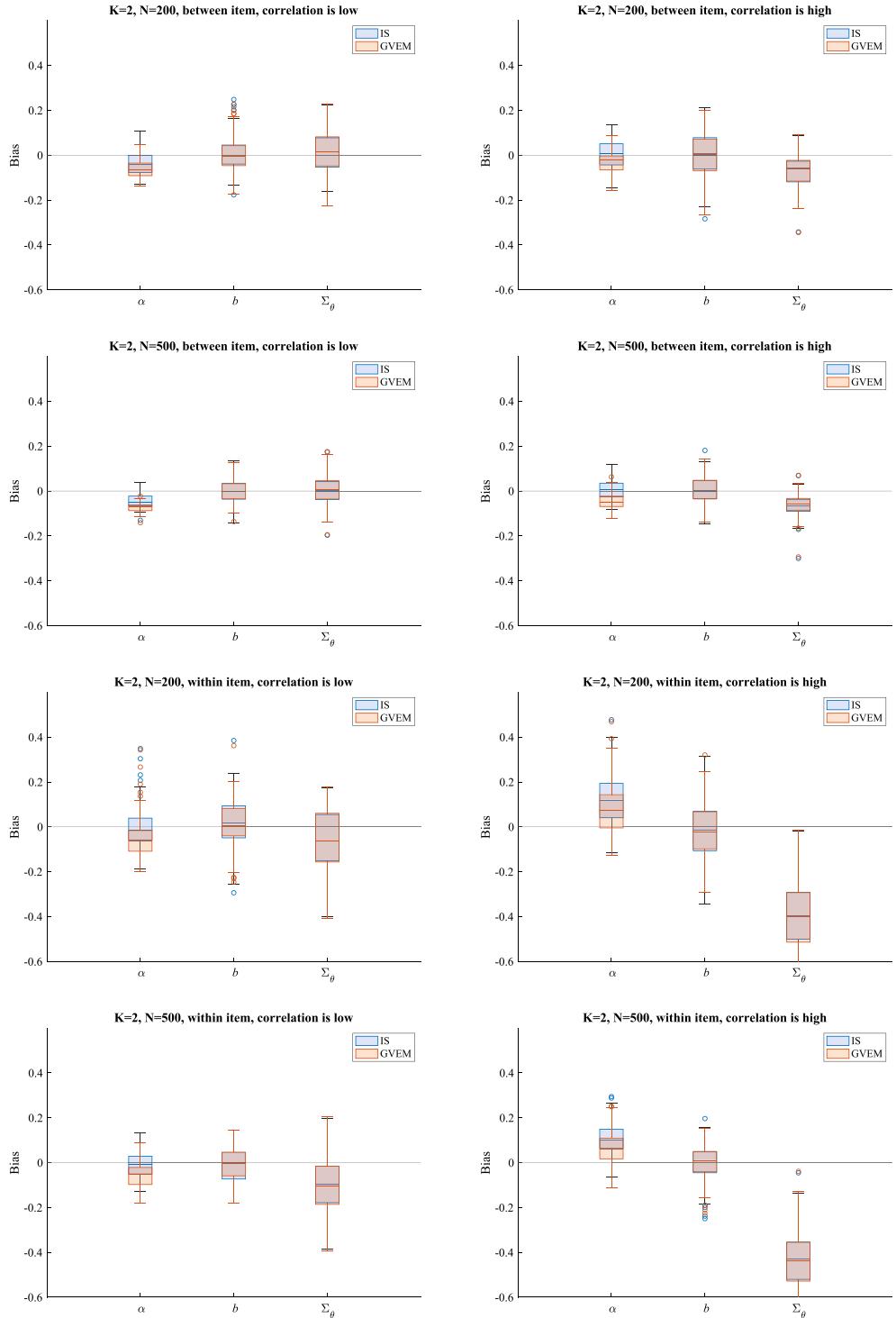
FIGURE 5.
Bias for $K = 2$ under exploratory analysis.
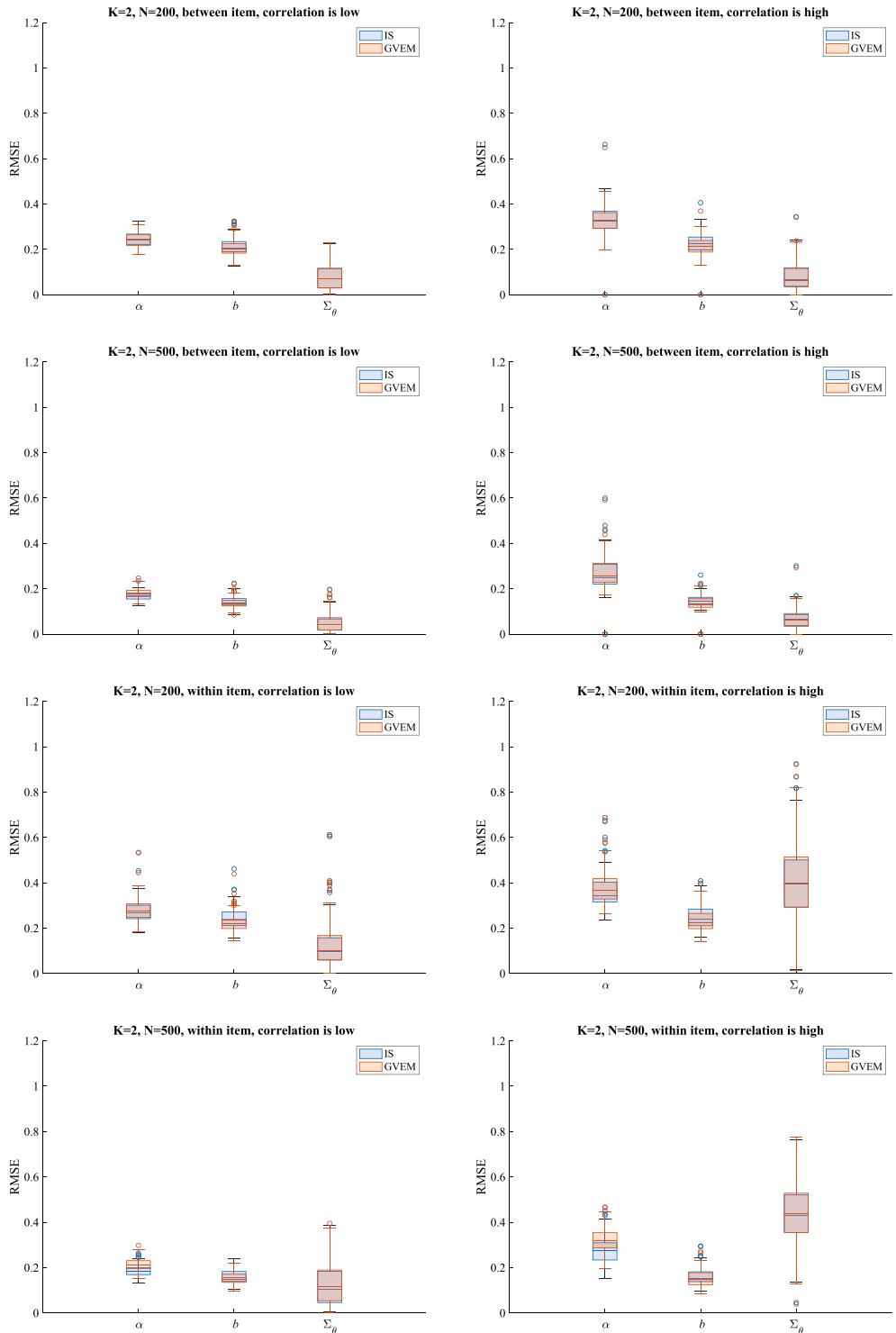
FIGURE 6.
RMSE for $K = 2$ under exploratory analysis.
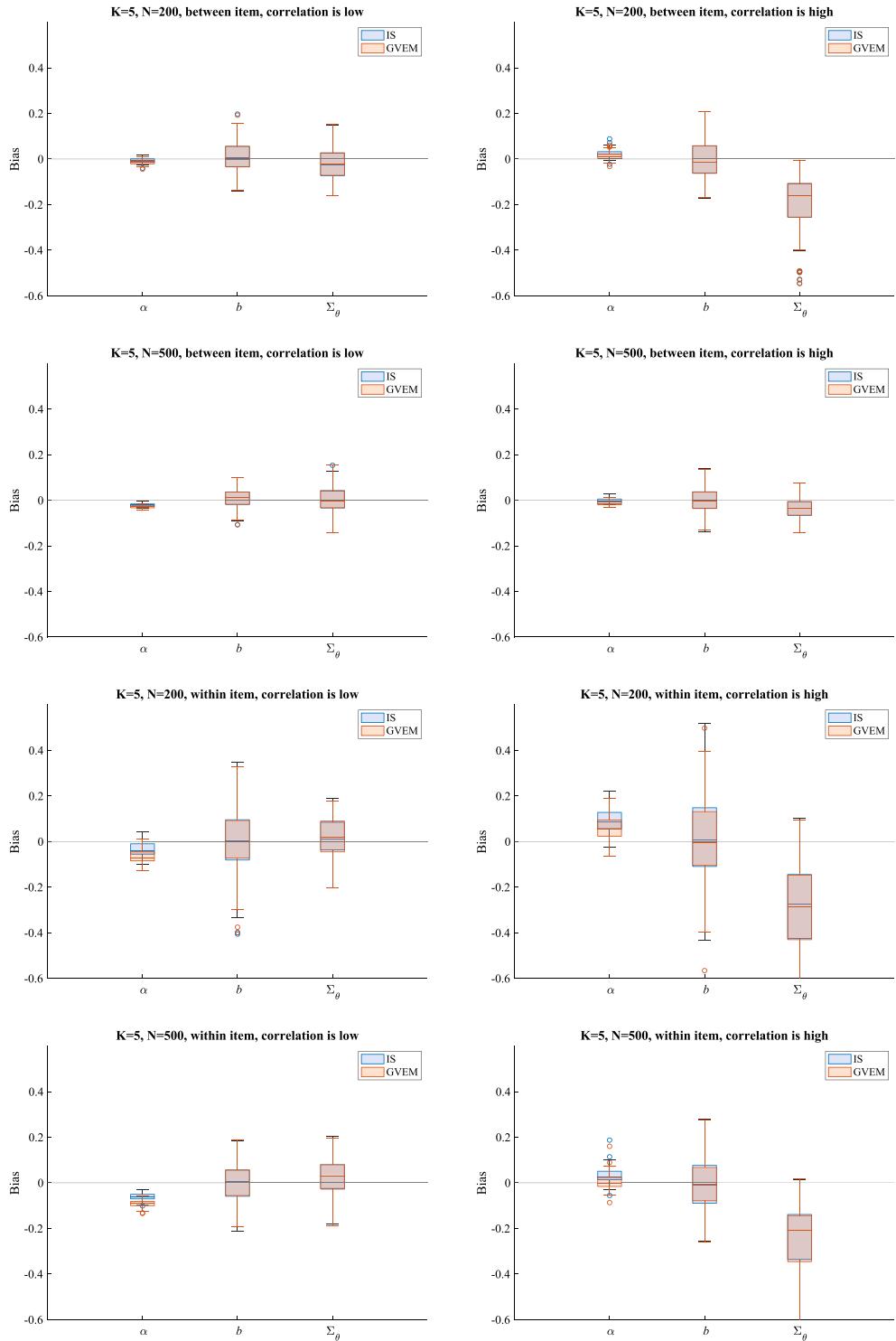
FIGURE 7.
Bias for $K = 5$ under exploratory analysis.
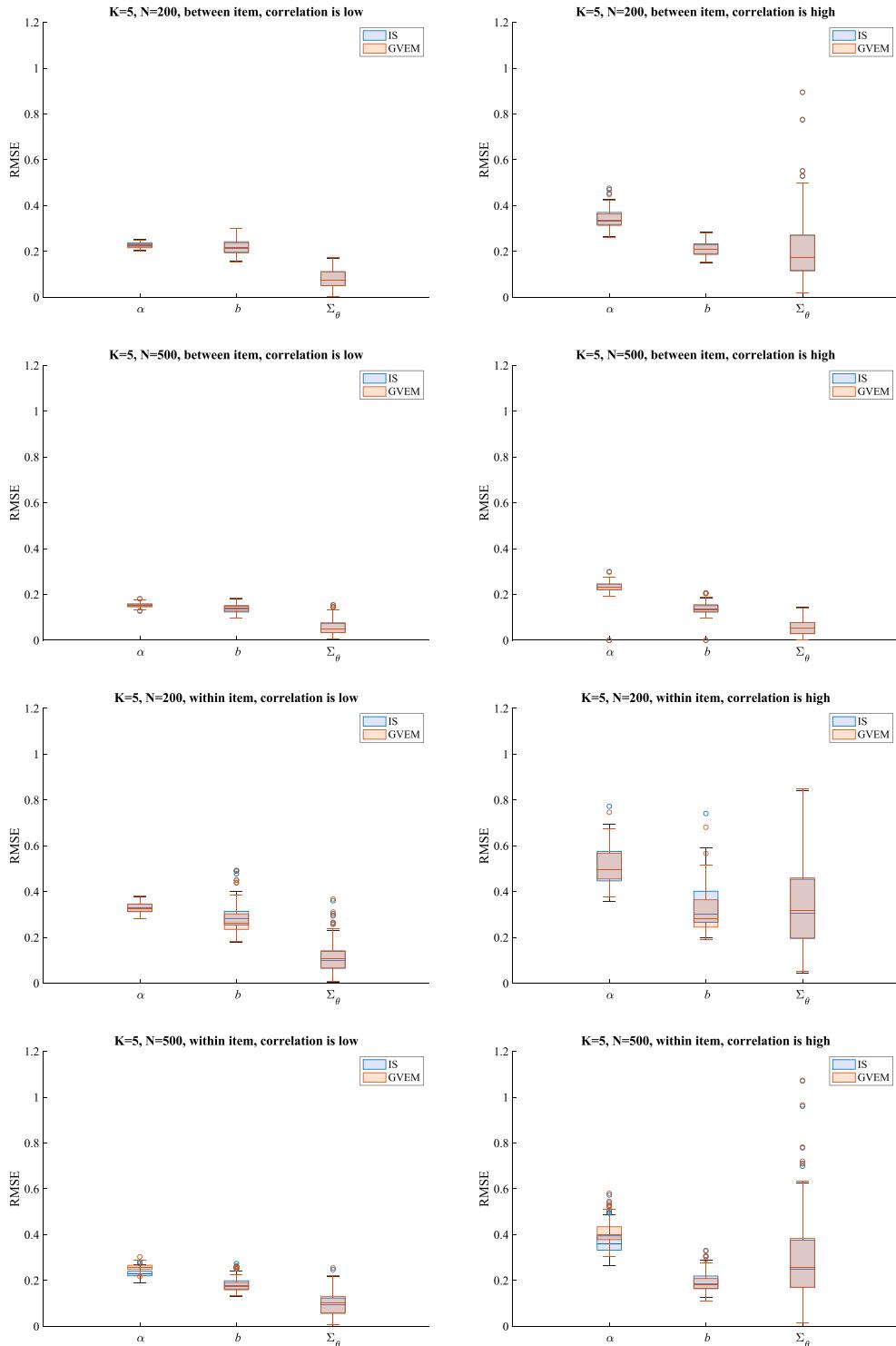
FIGURE 8.
RMSE for $K = 5$ under exploratory analysis.

TABLE 1.
Computation time (seconds) for the confirmatory M2PL estimation.

N	r	Model	$K = 2$		$K = 5$	
			GVEM	IW-GVEM	GVEM	IW-GVEM
200	Low	Between	0.68	2.88	1.15	11.31
		Within	1.17	2.89	4.96	11.59
	High	Between	1.06	2.98	2.33	12.81
		Within	1.52	2.89	11.61	15.44
500	Low	Between	1.52	6.88	2.29	33.98
		Within	2.51	6.95	10.80	35.52
	High	Between	1.90	7.04	3.70	33.49
		Within	3.35	6.95	21.62	34.45

TABLE 2.
Computation time (seconds) for the exploratory M2PL estimation.

N	r	Model	$K = 2$		$K = 5$	
			GVEM	IW-GVEM	GVEM	IW-GVEM
200	Low	Between	0.93	2.00	6.17	25.67
		Within	1.10	2.01	11.18	25.75
	High	Between	1.13	2.39	12.53	26.27
		Within	1.38	2.03	20.55	26.18
500	Low	Between	2.10	5.02	13.42	68.10
		Within	2.51	4.98	24.16	68.04
	High	Between	2.66	5.91	21.60	69.43
		Within	3.39	5.40	42.27	68.69

a key component of variational inference, is derived based on Jensen's inequality, the ELBO will approximate the log-marginal distribution (i.e., $\log P(X)$) more closely when $R \equiv P(X, Z)/q(Z)$ is more concentrated around its mean $P(X)$. Hence, the primary idea of IW-GVEM is to replace R with its sample mean by drawing i.i.d. samples from variational distribution $q(z)$. In so doing, we achieve a tighter bound of Jensen's inequality, but at the slight cost of computational efficiency. The added computation time is mainly due to sampling in the E-step and gradient descent in the M-step. From our simulation results, the bias correction is effective for confirmatory models and the extra computation time is acceptable because even with additional computational cost, the total time is still short. In fact, the time increase from GVEM to IW-GVEM is at a slow rate in that the time ratio between the two methods is smaller for more complex models (i.e., $K = 5$, within-item multidimensional structure, and high correlations). Note that for exploratory M2PL models, the original GVEM is still recommended because it already produces almost unbiased results and hence importance sampling seems unnecessary, although it does not introduce any undesirable bias either. Theoretically, Cho et al. (2021) proved that the estimated factor loading matrix and estimated latent factor from the GVEM algorithm is consistent as $N \to \infty$ and $J \to \infty$. The proposed IW-GVEM algorithm is based on the GVEM estimation, hence with consistent initial GVEM estimators, the final estimators from the IW-GVEM algorithm also have the theoretical guarantee to be consistent in the high-dimensional setting. Moreover, compared to ELBO in GVEM, the importance-weighted ELBOs are greatly improved after importance sampling. In finite-sample simulations, importance-weighted ELBOs at $M = 5$, 10, 50, and 100 are all larger than ELBO from GVEM and converge as M increases (See Appendix B).

In IW-GVEM, we propose to use the adaptive moment estimation method to automatically update the learning rate on the fly. Our preliminary results showed that the Adam algorithm performs better than fixed learning rate. Further, we also evaluated the effect of Monte Carlo sample size (i.e., $S = 10, 50, 100$) and sample size for the importance sampling step (i.e., $M = 10, 50$) and noted essentially the same results. Hence, we set $S = 10$ and $M = 10$ in our simulation study, which explains the only modest increase in computation time.

Aside from GVEM, another recently proposed fast algorithm for high-dimensional IRT estimation is the joint maximum likelihood estimation (Chen et al., 2019). This method treats the latent abilities as fixed effect parameters instead of random variables. Although this approach is innovative and their algorithm appears to produce accurate parameter estimates efficiently, the interpretation of person parameters is different such that caution needs to be exercised when one intends to generalize findings to a certain population. Plus, treating each individual as a separate fixed effect is, at the conceptual level, hard to justify when generalizing M2PL to a multiple-group MIRT model. This is because the goal of a multiple-group extension is to allow for unbiased marginal estimation of group-specific population distributions.

Instead, the GVEM method can be generalized to multiple-group MIRT in a more straightforward fashion. Our other study exploring multiple-group GVEM for differential item functioning detection (DIF) reveals that it can very well detect uniform DIF, but the power of detecting DIF on discrimination parameter is low. This is likely due to the estimation bias on $\boldsymbol{\alpha}$ from GVEM in the confirmatory model estimation, and hence the IW-GVEM will likely improve detection of the non-uniform DIF, in particular the DIF on discrimination parameters. Our study can also be extended in other directions. For instance, like in Cho et al. (2021), the IW-GVEM can be extended to M3PL models. Moreover, the current IW-GVEM algorithm does not automatically output standard error of item parameter estimates, and hence future studies may consider combining it with the supplemented EM algorithm (Cai, 2008; Chen & Wang, 2021) to produce accurate SE estimates. In addition to MIRT, the proposed method may also shed light on improving the performance of the variational estimation for other psychometric models, such as generalized linear mixed models (Jeon et al., 2017) and cognitive diagnosis models (Yamaguchi & Okada, 2020, 2020a).

Appendix A: Additional Comparative Studies

A.1: Comparing IW-GVEM with Importance-Weighted Variational Bayesian Method

In recent literature, researchers also proposed importance-weighted variational Bayesian (IW-VB) methods for the estimation of MIRT models. In particular, Urban and Bauer (2021) and Liu et al. (2022) proposed to use importance-weighted variational autoencoder (IW-VAE) for exploratory factor analysis. This method is a deep learning-based variational method and is computationally fast in large data sets. Although IW-VB methods handle large-scale data with high computational efficiency, their performances at relatively small-sized and medium-sized data are not competitive. While MCMC could be an alternative method for small samples, in situations with small to medium-sample sizes, our variational method is faster and more competitive than MCMC.

In this section, we provide additional finite-sample simulation results to show that our method outperforms the IW-VB methods in small to medium samples. To illustrate it, we compare our proposed IW-GVEM method and IW-VB method by Liu et al. (2022) at $N = 200$, $N = 500$ and $N = 1000$. Because their method focuses only on exploratory MIRT, we will compare the performance of our method (denoted as "IS" in the figure) to IW-VB for exploratory analysis. The simulation settings follow the same settings as in Sect. 2.1. The results are presented in Figures 9, 10, 11, 12, 13, 14, 15, and 16. From the results, we see the biases of IW-GVEM are closer to 0
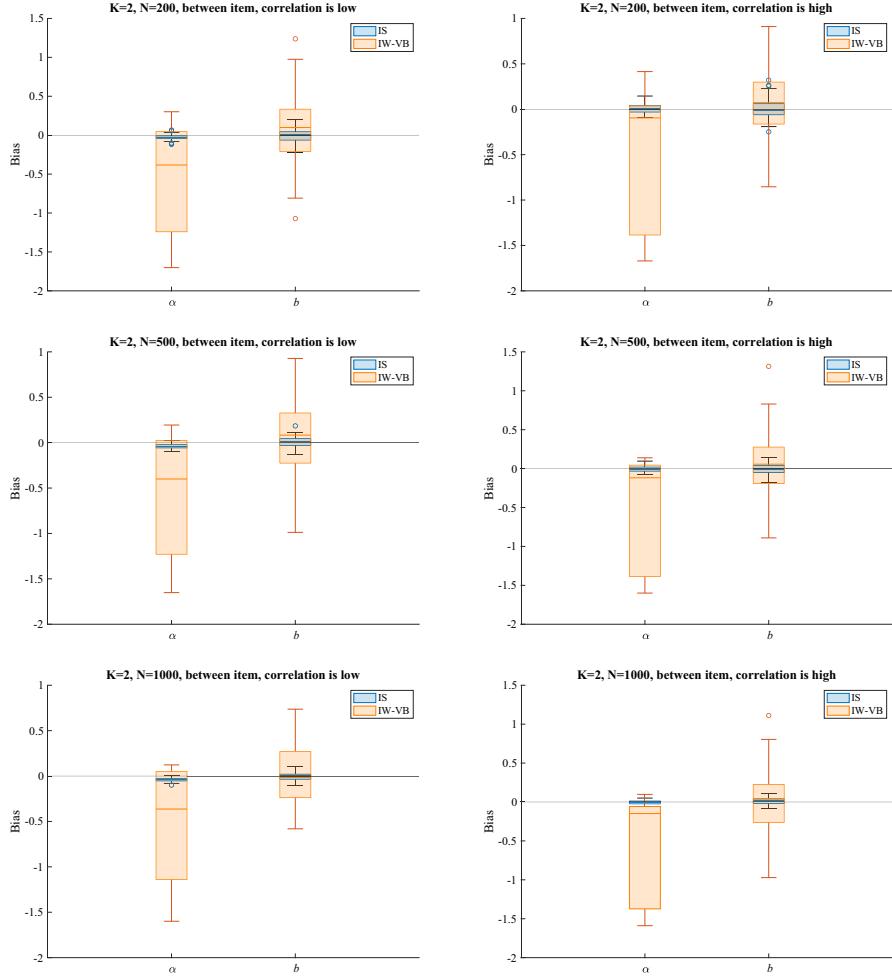
FIGURE 9.
Bias for $K = 2$ between item under exploratory analysis.

than the IW-VB method under all simulation settings. The RMSEs of our proposed method are substantially smaller than the IW-VB in Liu et al. (2022).

A.2: Comparing IW-GVEM with Joint Maximum Likelihood Method

The joint maximum likelihood (JML) estimator is a computationally efficient estimator with theoretical consistency established. It is proved in Chen et al. (2019) that JML estimator is consistent under high-dimensional settings and it outperforms the marginal maximum likelihood approaches in terms of computational costs. However, different from our IW-GVEM method, the latent abilities are treated as fixed effect parameters instead of random variables in JML method, which may constrain its performances in settings where latent factors are correlated. The JML estimation is also inconsistent in the setting when the number of items is fixed and the sample size grows to infinity. Because the number of parameters in the joint likelihood function grows to infinity, the standard theory for the maximum likelihood method cannot directly apply and the
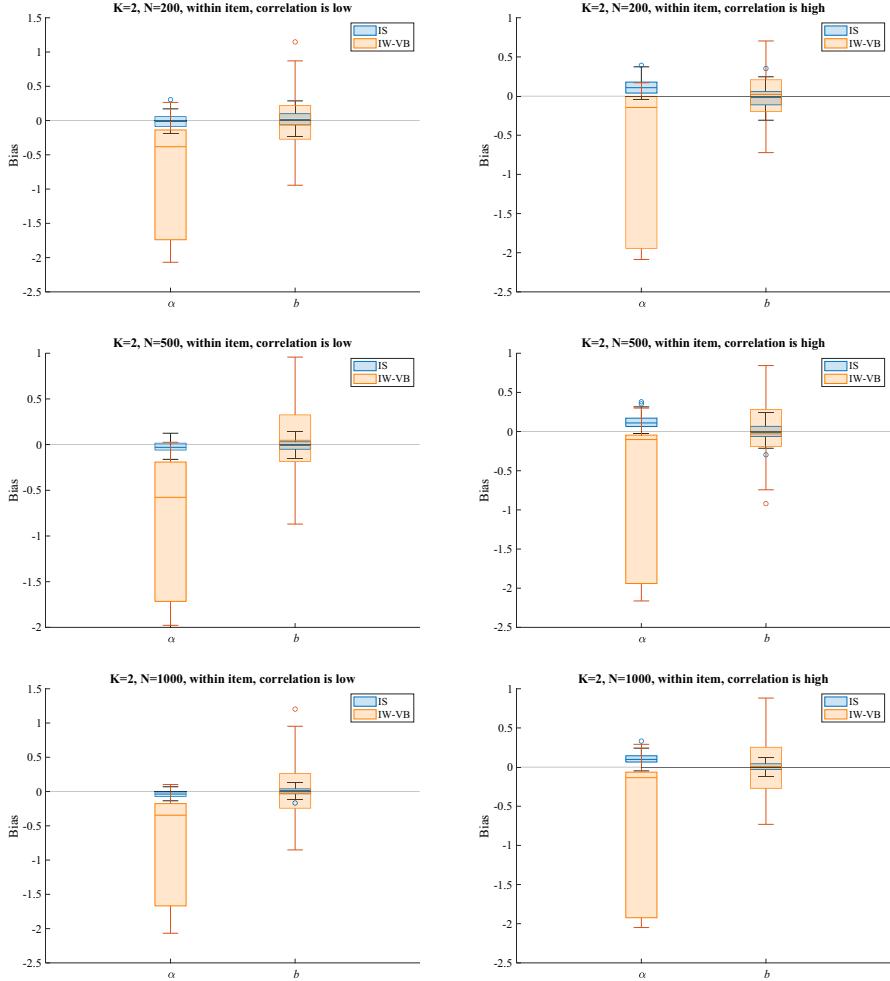
FIGURE 10.
Bias for $K = 2$ within item under exploratory analysis.

point estimation consistency for each item cannot be attained, which is known as Neyman–Scott phenomenon (Neyman & Scott, 1948).

Extensive simulation studies were conducted in Cho et al. (2021) to compare GVEM to JMLE method under the same simulation settings (sample sizes, within or between multidimensional structures, factor correlations, etc.) and using the same evaluation criteria (bias and RMSE) as in Sect. 2.1. Specifically, Figures 3 and 4 of Cho et al. (2021) compared the bias and RMSE of GVEM and JML and showed that GVEM has much lower bias and RMSE than JML across all settings. At certain challenging cases such as "within item, correlation is high", JML estimator has even worse performances. This could be explained by that latent factors are fixed effects in JMLE, whereas GVEM treats them as random effects with multivariate Gaussian distributions accounting for the correlations among factors.

As an improvement of GVEM method, our IW-GVEM method outperforms GVEM in confirmatory factor analysis and has overall comparable performances as GVEM in exploratory factor analysis, across all simulation settings. For a detailed comparison of the simulation results of IW-GVEM and GVEM, please refer to Sect. 2.2. As our IW-GVEM is comparable to, if not better
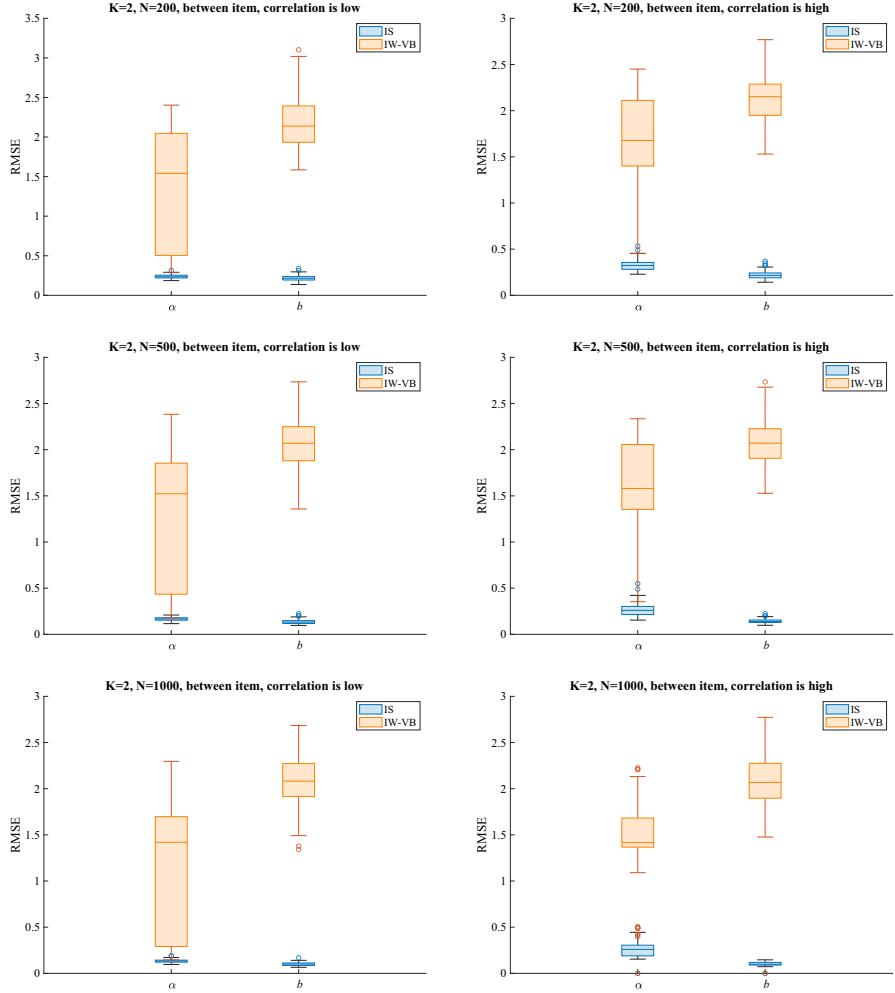
FIGURE 11.
RMSE for $K = 2$ between item under exploratory analysis.

than, GVEM, the performance of our IW-GVEM is also better than JML under our simulation settings.

Appendix B: Additional Simulation Study

In this section, we present finite-sample simulation studies to show that our proposed IW-GVEM greatly improves the ELBO from GVEM. For the purpose of illustration, we consider the four settings under $N = 200$ and $J = 30$: (1) within-item and low factor correlation; (2) between-item and low factor correlation; (3) within-item and high factor correlation; (4) between-item and high factor correlation. For each setting, we generate the ELBOs from the GVEM algorithm and importance-weighted ELBOs for different sample sizes $M = 5, 10, 50,$ and 100 at the importance sampling step over 100 replications. The calculated ELBOs are presented in Fig. 17. From Fig. 17, we see that the importance sampling step leads to a tighter importance-weighted ELBO ($M =$
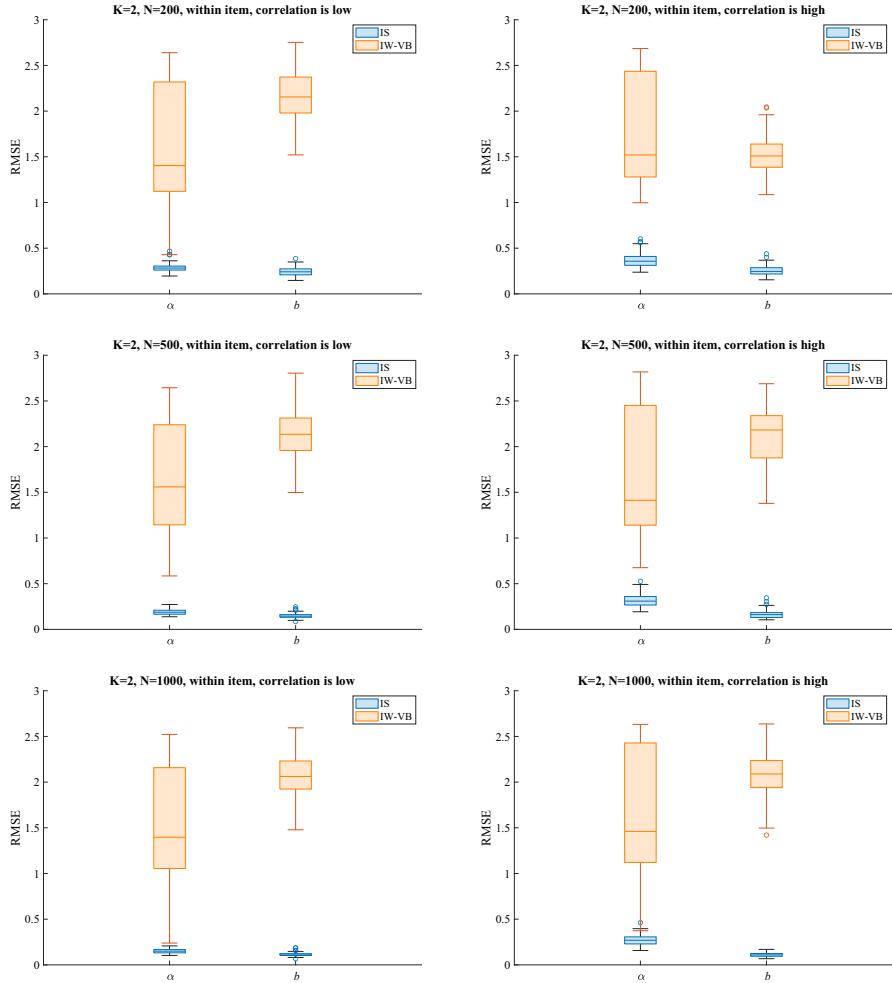
FIGURE 12.
RMSE for $K = 2$ within item under exploratory analysis.

5, 10, 50, 100) than that of GVEM. As the sample M in the importance sampling step increases, the ELBOs converge, which is consistent with theoretical results in Proposition 1.
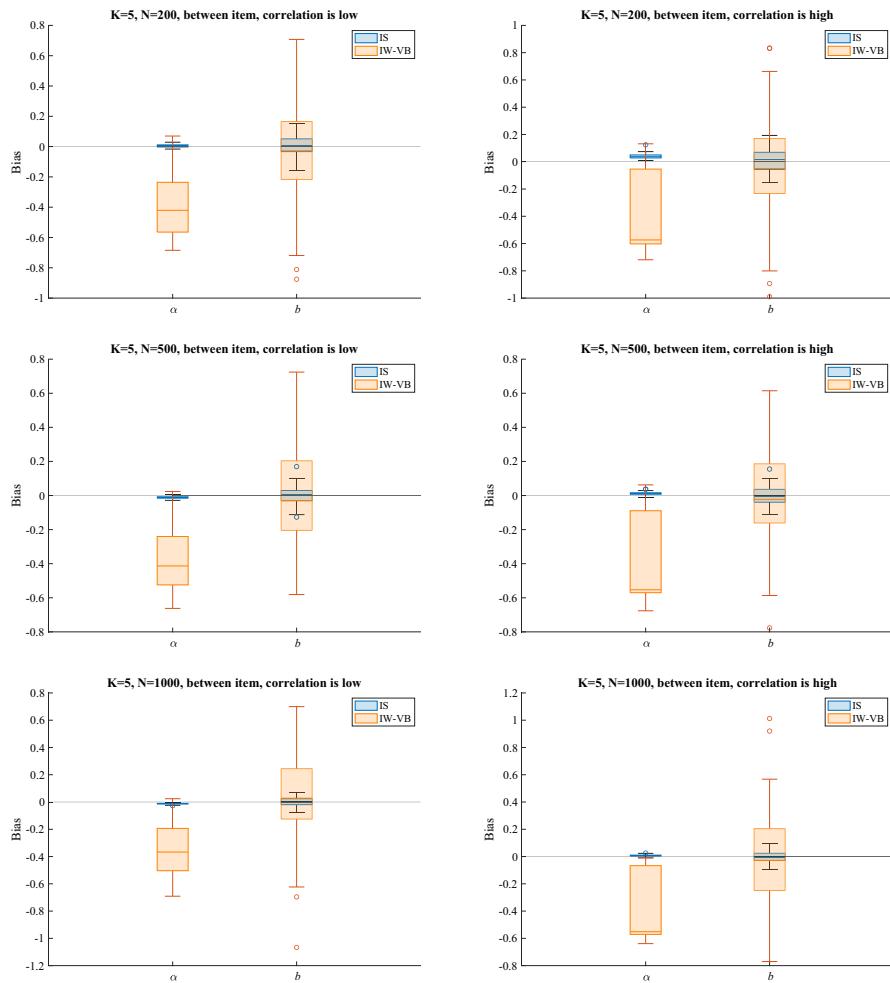
FIGURE 13.
Bias for $K = 5$ between item under exploratory analysis.
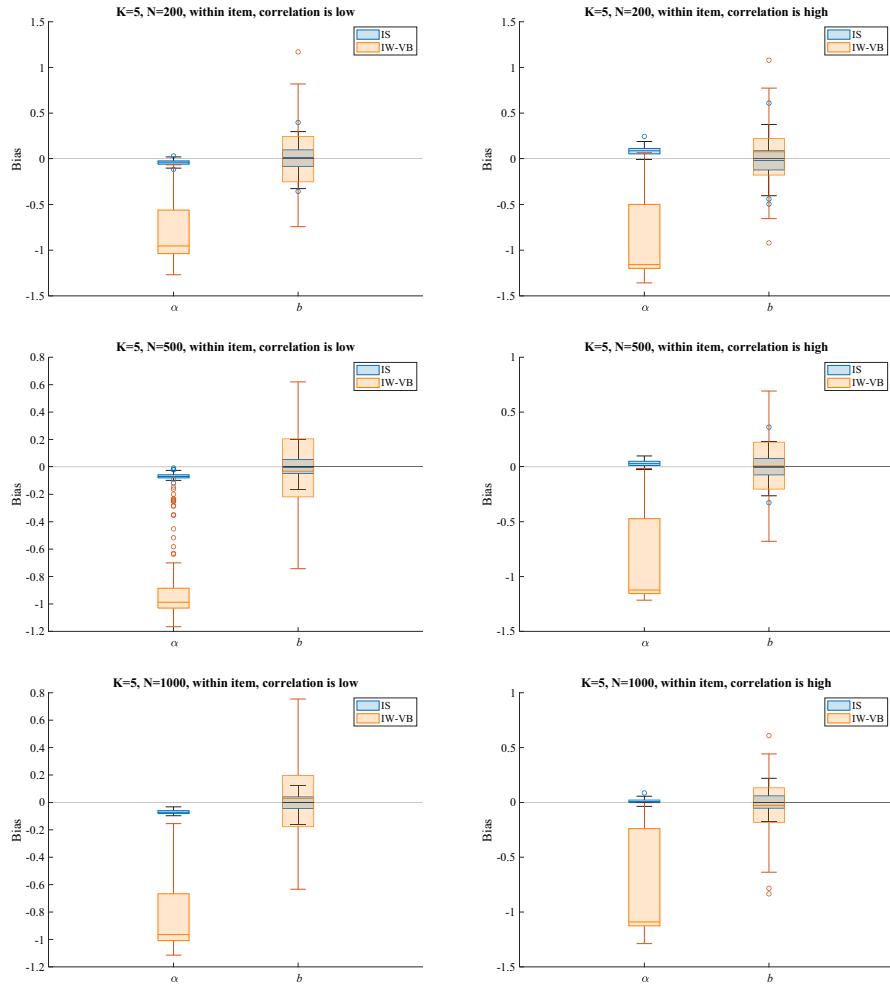
FIGURE 14.
Bias for $K = 5$ within item under exploratory analysis.
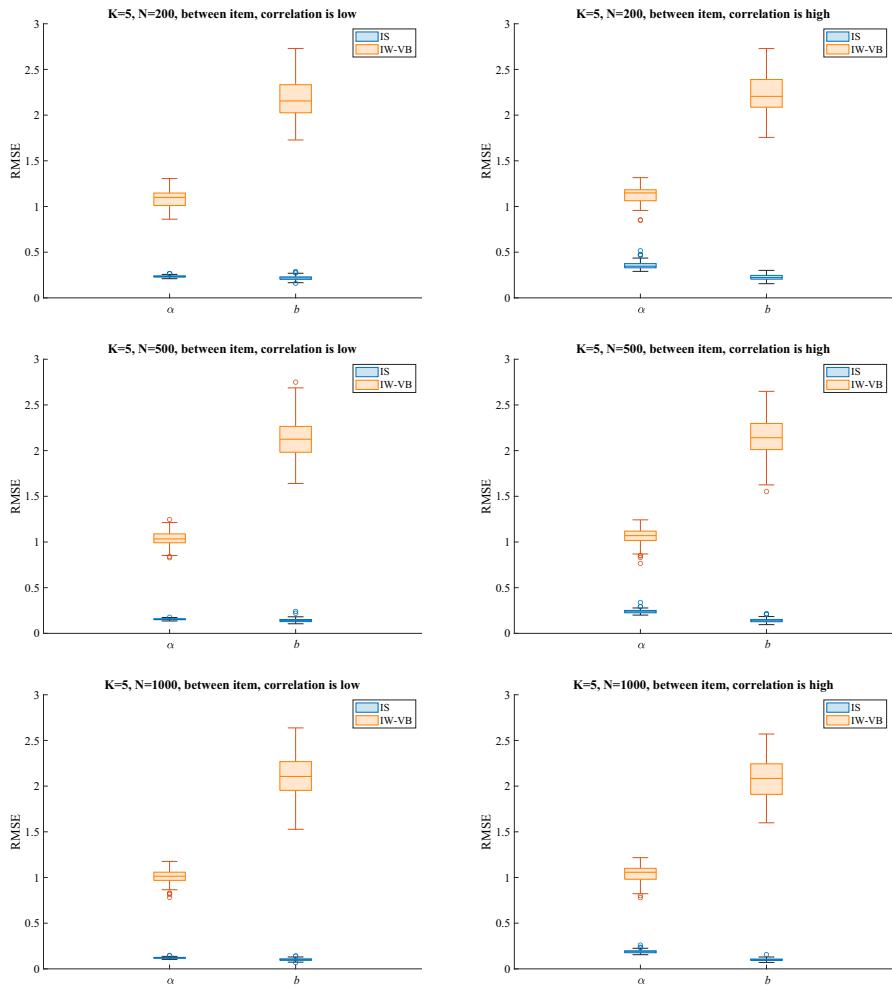
FIGURE 15.
RMSE for $K = 5$ between item under exploratory analysis.
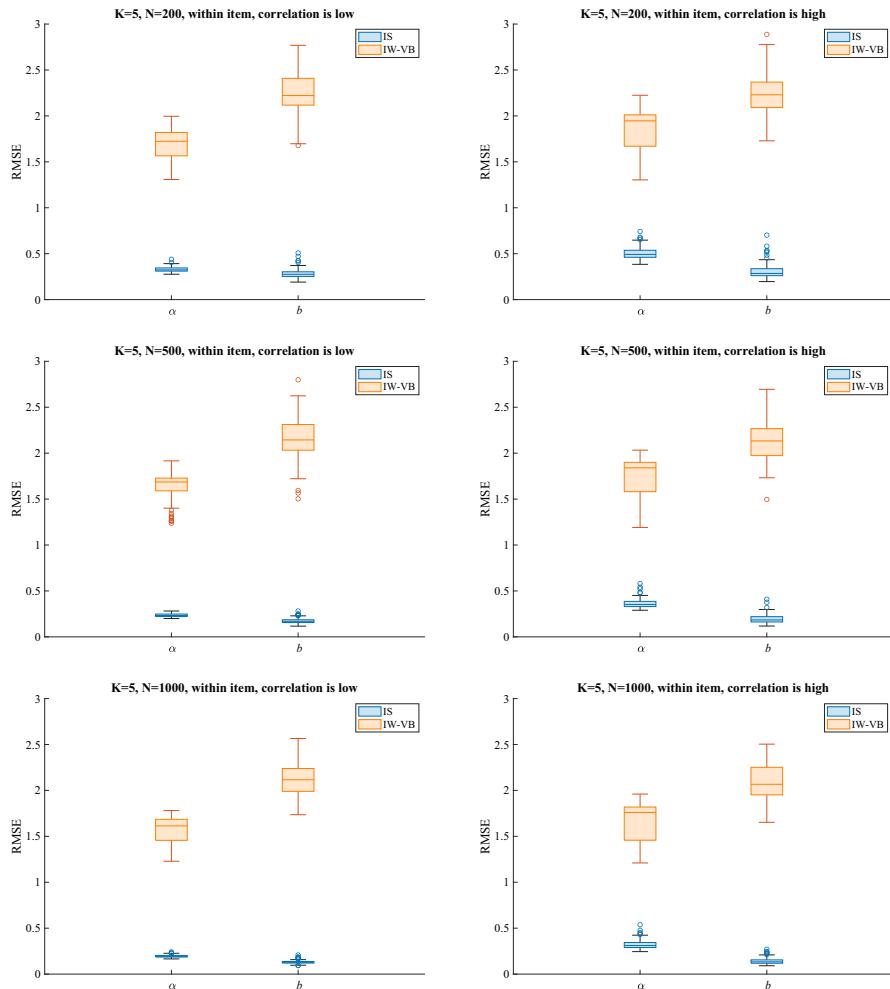
FIGURE 16.
RMSE for $K = 5$ within item under exploratory analysis.
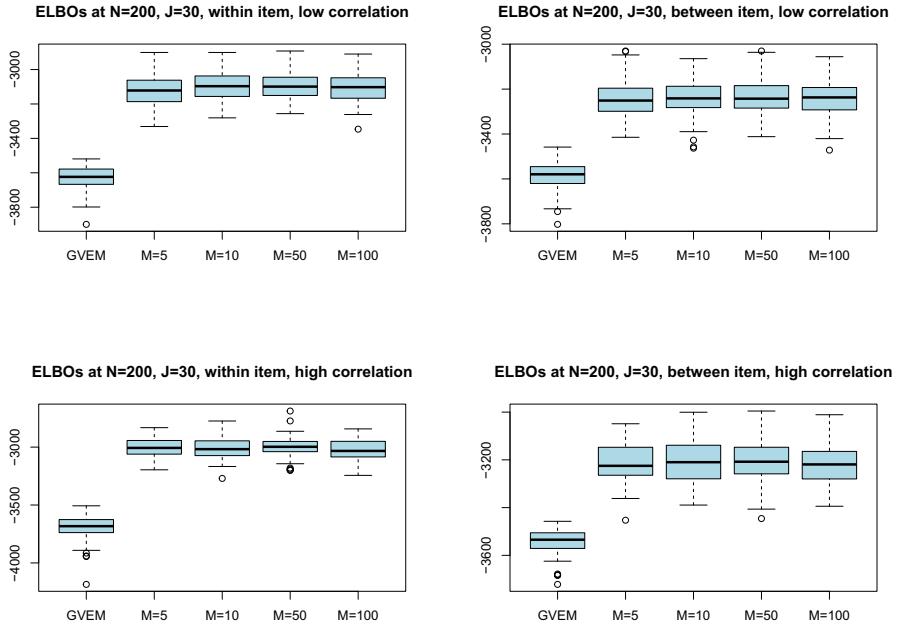
FIGURE 17.
Importance-weighted ELBO at $N = 200$, $J = 30$

Acknowledgments

Declarations

Data Availability The simulation code and datasets generated during the current study are available at https://github.com/jingoystat/A-Note-on-Improving-Variational-Estimation-for-Multidimensional-Item-Response-Theory.

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint arXiv:1406.5823.

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

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.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models.

Burda, Y., Grosse, R., & Salakhutdinov, R. (2015). *Importance weighted autoencoders*. arXiv preprint arXiv:1509.00519.

Cai, L. (2008). Sem of another flavor: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*, 309–329.

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

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

Cai, L., & Hansen, M. (2018). Improving educational assessment: Multivariate statistical methods. *Policy Insights from the Behavioral and Brain Sciences, 5*(1), 19–24.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological methods, 16*(3), 221.

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

Chen, P., & Wang, C. (2021). Using EM algorithm for finite mixtures and reformed supplemented EM for MIRT calibration. *Psychometrika, 86*, 299–326.

Cho, A. E., Xiao, J., Wang, C., & Xu, G. (2022). Regularized variational estimation for exploratory item response theory. *Psychometrika*, pp. 1–29.

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology, 74*, 52–85.

CRESST (2017). *English language proficiency assessment for the 21st century: Item analysis and calibration*.

Curi, M., Converse, G. A., Hajewski, J., & Oliveira, S. (2019). Interpretable variational autoencoders for cognitive models. In *2019 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE.

Domke, J., & Sheldon, D. R. (2018). Importance weighting and variational inference. *Advances in Neural Information Processing Systems, 31*.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436.

Hamilton, L. S., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J. I., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: Ii. nels: 88 science achievement. *American Educational Research Journal, 32*(3), 555–581.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2–3), 57–63.

Hui, F. K., Warton, D. I., Ormerod, J. T., Haapaniemi, V., & Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics, 26*(1), 35–43.

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. (2004). Graphical models. *Statistical science, 19*(1), 140–155.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kupermintz, H., Ennis, M. M., Hamilton, L. S., Talbert, J. E., & Snow, R. E. (1995). In dedication: Leigh burstein: Enhancing the validity and usefulness of large-scale educational assessments: I. nels: 88 mathematics achievement. *American Educational Research Journal, 32*(3), 525–554.

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.

Liu, T., Wang, C., & Xu, G. (2022). Estimating three- and four-parameter MIRT models with importance-weighted sampling enhanced variational auto-encoder. *Frontiers in Psychology, 13*.

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

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational bayes. *Frontiers in Psychology, 7*, 1422.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, 1–32.

OECD, N. (2003). *The pisa 2003 assessment framework: Mathematics, reading, science and problem solving knowledge and skills*.

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

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of educational and behavioral statistics, 24*(4), 342–366.

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

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory*, pp. 79–112. 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.

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73*(2), 167–182.

Thissen, D. (2013). Using the testlet response model as a shortcut to multidimensional item response theory subscore computation. In *New developments in quantitative psychology*, pp. 29–40. Springer.

Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika, 86*(1), 1–29.

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.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press.

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

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., & Goodman, N. (2020). *Variational item response theory: Fast, accurate, and expressive*. arXiv preprint arXiv:2002.00276.

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika, 85*(4), 973–995.

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics, 45*(5), 569–597.

Zhang, H., Chen, Y., & Li, X. (2020). A note on exploratory item factor analysis by singular value decomposition. *Psychometrika, 85*, 358–372.

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