Gaussian Variational Estimation in Multidimensional Item Response Theory

by

April E. Cho

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2020

Doctoral Committee:

Assistant Professor Gongjun Xu, Chair
Assistant Professor Yang Chen
Professor Naisyin Wang
Assistant Professor Zhenke Wu

April E. Cho

aprilcho@umich.edu

ORCID iD: 0000-0003-1818-0399

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my Ph.D. advisor, Professor Gongjun Xu, for his tremendous support and guidance during my graduate study. This work would never have been possible without his help. His guidance is on both academic and personal levels, from which I will benefit for my whole life. Also, I would like to thank my committee members, Professor Naisyin Wang, Professor Yang Chen, and Professor Zhenke Wu, for their advice and insights which helped me greatly improve the quality of my thesis.

My warmest thanks go to my family. I am grateful for their constant love, care and support which made this achievement possible. They are the most important people in my life and I dedicate this thesis to them.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . **ii**

LIST OF FIGURES . . . . **vii**

LIST OF TABLES . . . . **ix**

LIST OF APPENDICIES . . . . **x**

ABSTRACT . . . . **xi**

CHAPTER

I. Introduction **1**

II. Gaussian Variational EM for Multidimensional Item Response Theory **5**

II.1 Introduction 5

II.2 Gaussian Variational EM (GVEM) 9

II.3 GVEM for the M2PL Model 12

II.3.1 Algorithm Details 14

II.3.2 Theoretical Properties 17

II.4 GVEM for the M3PL Model 19

II.4.1 Algorithm Details 24

II.4.2 Stochastic Optimization of GVEM 28

II.5 Simulations 30

II.5.1 Design 30

II.5.2 Results for the M2PL model 32

II.5.3 Results for the M3PL model 39

II.5.4 Estimating the Number of Dimensions 40

II.6 Real Data Analysis 43

II.7 Discussions 45

Appendix of Chapter II 48

II.A Derivation of GVEM in the 2PL model 48

II.B Proof of Proposition II.7 52

II.C Derivation of GVEM in the 3PL model 53

II.D Proof of Theorem II.3 56

III. Regularized Variational Estimation for MIRT 61

III.1 Introduction 61

III.2 Variational Estimation for MIRT 65

III.3 Regularized Estimation of Test Structure 66

III.3.1 Computation via GVEM 70

III.4 Simulation Study 73

III.4.1 Design 73

III.4.2 Simulation Results 74

III.5 Real Data Analysis 78

III.6 Discussions 84

Appendix of Chapter III 85

III.A Derivation in M2PL 85

III.B Derivation in M3PL 86

IV. Extensions of Gaussian Variational EM 89

IV.1 Introduction 89

IV.2 Multidimensional 4PL model 92

IV.2.1 GVEM for M4PL 95

IV.2.2 Simulation studies : M4PL 98

IV.3 Differential Item Functioning (DIF) Analysis 99

IV.3.1 Algorithm Details 102

IV.3.2 Simulation Studies : DIF 104

IV.4 Discussions 107

Appendix of Chapter IV 109

IV.A Derivation of EM steps for M4PL 109

IV.B Derivation for DIF analysis 114

BIBLIOGRAPHY 117

LIST OF FIGURES

FIGURE

II.1 Parameter recovery of the between-item M2PL models : GVEM vs MHRM . 34

II.2 Parameter recovery of the within-item M2PL models : GVEM vs MHRM . . 35

II.3 Parameter recovery of the between-item M2PL models : GVEM vs JML . . 36

II.4 Parameter recovery of the within-item M2PL models : GVEM vs JML . . . 37

II.5 Average computation time 38

II.6 Parameter recovery of the between-item M3PL models : GVEM vs MHRM . 41

II.7 Parameter recovery of the within-item M3PL models : GVEM vs MHRM . . 42

II.8 Real Data: Model Selection 45

III.1 Correct estimation rates under M2PL. 75

III.2 Correct estimation rates under M3PL. 75

III.3 Comparison of Lasso and Adaptive Lasso in M2PL under Constraint 1 . . . 76

III.4 Comparison of Lasso and Adaptive Lasso in M2PL under Constraint 2 . . . 77

III.5 Comparison of Lasso and Adaptive Lasso in M3PL under Constraint 1 . . . 77

III.6 Comparison of Lasso and Adaptive Lasso in M3PL under Constraint 2 . . . 78

III.7 Description of questions in science test of NELS:88 80

IV.1 Parameter recovery of the between-item M4PL models from confirmatory factor analysis 100

IV.2 Correct Estimation Rates(%) under the between-item multidimensional structure 106

IV.3 Correct Estimation Rates(%) for the within-item multidimensional structure 106

IV.4 False Positive and False Negative rates in between-item model (first row) and within- item model (second row) 107

LIST OF TABLES

TABLE

II.1 Correct estimation rate(%) of test dimension in the M2PL model 43

II.2 Correct estimation rate(%) of test dimension in the M3PL model 43

II.3 Real Data: comparison of estimated $\hat{\Sigma}_{\boldsymbol{\theta}}$ 44

III.1 GIC comparison with Adaptive Lasso penalty 79

III.2 Estimated sparse test structure for math test in NELS:88 82

III.3 Estimated sparse test structure for science test in NELS:88 83

III.4 Estimated Correlation between latent factors 83

LIST OF APPENDICIES

APPENDIX

II.A Derivation of GVEM in the 2PL model 48

II.B Proof of Proposition II.7 52

II.C Derivation of GVEM in the 3PL model 53

II.D Proof of Theorem II.3 56

III.A Derivation in M2PL 85

III.B Derivation in M3PL 86

IV.A Derivation of EM steps for M4PL 109

IV.B Derivation for DIF analysis 114

ABSTRACT

Multidimensional Item Response Theory (MIRT) is widely used in assessment and evaluation of educational and psychological tests. It models the individual response patterns by specifying functional relationship between individuals' multiple latent traits and their responses to test items. One major challenge in parameter estimation in MIRT is that the likelihood involves intractable multidimensional integrals due to latent variable structure. Various methods have been proposed that either involve direct numerical approximations to the integrals or Monte Carlo simulations. However, these methods have some limitations in that they are computationally demanding in high dimensions and rely on sampling from a posterior distribution.

In the second chapter of the thesis, we propose a new Gaussian Variational EM (GVEM) algorithm which adopts a variational inference to approximate the intractable marginal likelihood by a computationally feasible lower bound. The optimal choice of variational lower bound allows us to derive closed-form updates in EM procedure, which makes the algorithm efficient and easily scale to high dimensions. We illustrate that the proposed algorithm can also be applied to assess the dimensionality of the latent traits in an exploratory analysis. Simulation studies and real data analysis are presented to demonstrate the computational efficiency and estimation precision of the GVEM algorithm in comparison to the popular alternative Metropolis-Hastings Robbins-Monro algorithm. In addition, theoretical guarantees are derived to establish the consistency of the estimator from the proposed GVEM algorithm.

One of the key elements in MIRT is the relationship between the items and the latent traits, so-called a test structure. The correct specification of this relationship is crucial for accurate assessment of individuals. Hence, it is of interest to study how to accurately estimate the test structure from data. In the third chapter, we propose to apply GVEM to solve a latent variable selection problem for MIRT and empirically estimate the test structure. The main idea is to impose L1-type penalty to the variational lower bound of the likelihood to recover a simple test structure in iterative procedures. Simulation studies show that the proposed method accurately estimates the test structure and is computationally efficient. A real data analysis on the large-scale assessment test called National Education Longitudinal Study of 1988 is presented.

In the last chapter, we discuss some of the interesting extensions of our proposed method. The first extension is to develop the estimation method via GVEM procedures for the Multidimensional 4-Parameter Logistic model, which is known to be more challenging than previously discussed MIRT models. The second extension is to study Differential Item Functioning (DIF) analysis in MIRT. In brief, DIF occurs when groups (such as defined by gender, ethnicity, or education) have different probabilities of responses for a given test item even though people have the same latent abilities. Our goal is to identify test items that have DIF. We formulate the DIF analysis in MIRT as the regularization problem and solve it via our proposed GVEM approach. Simulation studies are presented to show the performance of our proposed method on these topics.

ChapterI

Introduction

Educational and psychological assessment refers to a way of testing individuals on their latent abilities, characteristics and behavior using combinations of techniques. Its goal is to develop good understanding of the individuals' latent traits using the observed responses on questionnaires or assessment tests. It is widely used in various fields including education, psychology, and medicine. For example, the proper psychological assessments of individuals would potentially help prepare customized treatments to individuals with mental disorders. In addition, teachers can provide personalized feedback to students and improve the learning process of students. Another interesting application is the online recommender system. The latent preference of online consumers can be measured by analyzing their shopping or viewing history and this could help make predictions for the individualized recommendations. Hence our proposed methods and discussions could be potentially applied in various fields although we mainly focus on the setting of psychological and educational assessment in this dissertation.

The measurement of psychological properties has been a long-lasting quest that originated in the 19th century (Sijtsma & Junker, 2006). Various statistical models have been proposed for psychological assessment since then. Classical test theory (CTT, Gulliksen, 1950; Spearman, 1907, 1913) has been the most popular measurement model for most of the 20th century. Fundamental idea of CTT is that the observed test scores contain a true score plus some random error component. That is, CTT assumes that due to random error an observable test score often is not the value representative of a testee's true performance on the test. The main purpose of CTT is to determine the degree in which test scores are influenced by random error. This has lead to a multitude of methods for estimating the reliability of a test score, of which Cronbach's alpha (Cronbach, 1951) is the most famous.

CTT was the dominant statistical approach to psychological measurement until the item response theory was introduced (Rasch, 1960; Lord, 1968). Item response theory (IRT) is a general framework for specifying mathematical functions that describe the relationships between individuals' latent traits and characteristics of test items. Unlike the classical test theory, the item response theory considers the items to be heterogeneous. For example, items may differ in terms of their difficulty levels. IRT is generally regarded as being superior to classical test theory (Embretson & Reise, 2000) and has become the preferred method for developing scales in high-stake tests (e.g. Graduate Record Examination and Graduate Management Admission Test).

The early research on IRT primarily involved unidimensional IRT models that measure only a one latent trait that may represent. As an extension, several multidimensional IRT models have been proposed for modeling the individuals' response patterns driven by their multiple latent traits (e.g. McKinley & Reckase, 1982; Bock, Gibbons, & Muraki, 1988; Revuelta, 2014). The increasing availability of rich educational and psychological tests has made MIRT an attractive model to handle complex assessment data measuring multiple latent traits at the same time. However, there are some challenges in parameter estimation problem for MIRT. That is, the likelihood involves intractable multidimensional integrals due to multidimensional latent variable structure. With the advancement of computational and statistical techniques, various methods have been proposed that either involve direct numerical approximations to the integrals or Monte Carlo simulations. However, these methods still have some limitations in that they are computationally demanding in high dimensions and rely on sampling from a posterior distribution. In this thesis, we attempt to tackle the challenging estimation problems in MIRT and develop accurate and efficient estimation algorithms.

The thesis is organized as follows. In the second chapter, we propose a new Gaussian Variational EM (GVEM) algorithm which adopts a variational inference to approximate the intractable marginal likelihood by a computationally feasible lower bound. The optimal choice of variational lower bound allows us to derive closed-form updates in EM procedure, which makes the algorithm efficient and easily scale to high dimensions. We also illustrate that the proposed algorithm can also be applied to assess the dimensionality of the latent traits in an exploratory analysis. A series of simulation studies and real data analysis are presented to demonstrate the performance of the proposed GVEM method in comparison to the popular alternative Metropolis-Hastings Robbins-Monro algorithm. In essence, GVEM method produces more precise parameter estimations and is computational efficient. We also present theoretical guarantees of the estimator from the proposed GVEM algorithm to establish its consistency.

In the third chapter, we propose to apply GVEM to solve a latent variable selection problem for MIRT and empirically estimate the test structure. The test structure illustrates a relationship between the items and the latent traits being measured in a test. The correct specification of this relationship is crucial for accurate assessment of individuals and further model calibration. In practice, practitioners often use fixed test structure based on their prior knowledge for the analysis. However, wrong specification of the relationship would lead to biased estimation. Hence, we would like to study how to accurately estimate the test structure from data. The main idea is to impose L1-type penalty to the variational lower bound of the likelihood to recover a simple test structure in iterative procedures. Simulation studies show that the proposed method accurately estimates the test structure and is computationally efficient. A real data analysis on the large-scale assessment test called National Education Longitudinal Study of 1988 is presented to examine it in terms of the test design.

In the last chapter, we discuss interesting extensions of our proposed method; (1) to develop the estimation method via GVEM procedures for the Multidimensional 4-Parameter Logistic (M4PL) model and (2) to study Differential Item Functioning (DIF) analysis in MIRT. M4PL model has been less preferred so far in the field of education and psychology probably due to its challenge in parameter estimation. In most research, Bayesian approaches with MCMC sampling were used for the parameter estimation in unidimensional 4PL models. However, it gets time consuming for the high dimensional assessment data even for unidimensional latent trait. M4PL models incorporates multiple latent traits at the same time, resulting in intractable multidimensional integrals in the calculation of log-likelihood and making the parameter estimation even more challenging. We develop the variational EM method to facilitate the paraemter estimation in M4PL and discuss the performance with some simulation studies. In the second half of the last chapter, we discuss DIF anlaysis in MIRT. DIF occurs when groups (such as defined by gender, ethnicity, or education) have different probabilities of responses for a given test item even though people have the same latent abilities. Our goal is to identify biasedness in test items (i.e. that have DIF). We formulate the DIF analysis in MIRT as the regularization problem and solve it via our proposed GVEM approach. Simulation studies are presented to show the performance of our proposed method on these topics. Lastly we discuss some of the challenges remained and talk about future directions.

ChapterII

Gaussian Variational EM for Multidimensional Item Response Theory

II.1 Introduction

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

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

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

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

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

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

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

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

where $Y_i = (Y_{ij}, j = 1, \ldots, J)$ is the ith subject's response vector and J is the total number of items in the test. The ϕ denotes the K-dimensional Gaussian distribution of $\boldsymbol{\theta}$ with mean 0 and covariance $\Sigma_{\boldsymbol{\theta}}$. The maximum likelihood estimators of the model parameters are then obtained from maximizing the log-likelihood function. However, due to the latent variable structure, maximizing the log-likelihood function involves a K dimensional integrals that are usually intractable. Direct numerical approximation to the integrals have been proposed in the literature, such as the Gauss–Hermite quadrature (Bock & Aitkin, 1981) and the Laplace approximation (Lindstrom & Bates, 1988; Tierney & Kadane, 1986; Wolfinger & O'connell, 1993). However, the Gauss–Hermite quadrature approximation is known to become 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 is in skewed shape. Other numerical approximation methods based on Monte Carlo simulations have also been developed in the literature, such as the Monte Carlo expectation-maximization (McCulloch, 1997), stochastic expectation-maximization (von Davier & Sinharay, 2010), Metropolis-Hastings Robbins-Monro algorithms (Cai, 2010b, 2010a). These methods usually depends on sampling data points from a posterior distribution and would be computationally involving. Recently, S. Zhang, Chen, and Liu (2020) proposed to use the stochastic EM algorithm (Celeux & Diebolt, 1985) for the item factor analysis, where an adaptive-rejection-based Gibbs sampler is still needed for the stochastic E step. Moreover, Chen, Li, and Zhang (2019) studied the joint maximum likelihood estimation by treating the latent abilities as fixed effect parameters instead of random variables as in (II.3).

In this chapter, we propose a computationally efficient method that is based on the variational approximation to the log-likelihood. Variational approximation methods are mainstream methodology in computer science and statistical learning, and they have been applied to diverse areas including speech recognition, genetic linkage analysis, and document retrieval (Blei & Jordan, 2004; Titterington, 2004). Recently, there is an emerging interest in developing and applying variational methods in statistics (Blei, Kucukelbir, & McAuliffe, 2017; Ormerod & Wand, 2010). In particular, Gaussian variational approximation methods were developed for standard generalized linear mixed effects models (GLMM) with nested random effects (Ormerod & Wand, 2012; Hall, Ormerod, & Wand, 2011). However, the variational methods have only been slowly recognized in psychometrics and educational measurement, with the pioneer papers by Rijmen and Jeon (2013) as well as Jeon, Rijmen, and Rabe-Hesketh (2017).

In essence, variational approximations refer to a family of deterministic techniques for making approximate inference for parameters in complex statistical models (Ormerod & Wand, 2010). The key is to approximate the intractable integrals (e.g. Eq.(II.3)) with a computational feasible form, known as the variational lower bound to the original marginal likelihood. In psychometrics, Rijmen and Jeon (2013) first developed a variational algorithm for a high dimensional IRT model, but their algorithm was limited to only discrete latent variables. Recently, Jeon et al. (2017) proposed a variational maximization-maximization (VMM) algorithm for maximum likelihood estimation of GLMMs with crossed random effects. They showed that VMM outperformed Laplace approximation with small sample size. However, their study is limited in several respects: (i) They only considered the Rasch model. Although extending their algorithm to the 2PL model may be straightforward, its generalization to 3PL is unknown because 3PL does not belong to the GLMM family; (ii) The key component in their algorithm is the mean-field approximation (Parisi, 1988) that assumes independence of the latent variables given observed data. Even though it seems acceptable to assume independence of each random item effect, this independence assumption can no longer apply to the MIRT models when different dimensions are assumed to be correlated; (iii) In their first maximization step, the closed-form solution still contains a two-dimensional integration where adaptive quadrature is used; in the second maximization step, a Newton-Raphson algorithm is used. Therefore, both steps involve iterations, which may slow down the algorithm. Instead, our proposed GVEM algorithm has closed-form solutions for all parameters in both the E and M steps, and it can deal with high-dimensional MIRT models when the multiple latent traits are correlated. Moreover, the GVEM algorithm is established for both the M2PL and M3PL models. Consistency theory of the estimators from our proposed algorithm is established, and the performance of the algorithm is thoroughly evaluated via simulation studies.

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

II.2 Gaussian Variational EM (GVEM)

From here onwards, for the MIRT models in (II.1) and (II.2), we denote the model parameters by $\boldsymbol{A} = \{\boldsymbol{\alpha}_j, j = 1, \ldots, J\}$, $\boldsymbol{B} = \{b_j, j = 1, \ldots, J\}$, and $\boldsymbol{C} = \{c_j, j = 1, \ldots, J\}$. As defined in Section II.1, we use the notation $M_p = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\}$ in the 3PL model and $M_p = \{\boldsymbol{A}, \boldsymbol{B}\}$ in the 2PL model for simplicity. Latent traits $\boldsymbol{\theta}$ from different dimensions are correlated, resulting in a K by K covariance matrix $\Sigma_{\boldsymbol{\theta}}$. To fix the origin and units of measurement, it is conventional to fix the mean and variance of all $\boldsymbol{\theta}$'s to be 0 or 1, respectively. To remove rotational indeterminacy in the exploratory analysis, (i.e. to ensure the model identifiability) researchers often either assume $\Sigma_{\boldsymbol{\theta}} = I_K$ or assume $\boldsymbol{A}$ contains a K by K lower triangular matrix (Reckase, 2009).

On the other hand, in the confirmatory analysis, the zero structure of the loading matrix $\boldsymbol{A}$ is completely or partially specified while the remaining nonzero elements are left unknown.

In this case, the correlation of latent traits $\boldsymbol{\theta}$ is of interest and we need to estimate the covariance matrix $\Sigma_{\boldsymbol{\theta}}$. In this chapter, we consider a general setting of $\Sigma_{\boldsymbol{\theta}}$ that works for both exploratory and confirmatory analyses.

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

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

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

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

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

for any arbitrary probability density function q_i satisfying $\int_{\boldsymbol{\theta}_i} q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i = 1$. Since $P(Y_i \mid M_p) = P(Y_i, \boldsymbol{\theta}_i \mid M_p)/P(\boldsymbol{\theta}_i \mid Y_i, M_p)$, then we can further write

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

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

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

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

The follow-up question is how to design the candidate distribution function $q_i(\boldsymbol{\theta}_i)$ that gives the best approximation of the marginal likelihood. From the above argument, the best choice is the unknown posterior distribution function $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$. Although this choice of $q_i(\boldsymbol{\theta}_i)$ is intractable, it provides a guideline to choose $q_i(\boldsymbol{\theta}_i)$ in the sense that a good choice of $q_i(\boldsymbol{\theta}_i)$ must approximate $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ well. The well-known EM algorithm follows this idea and can be interpreted as a maximization-maximization (MM) algorithm (Hunter & Lange, 2004) based on the above decomposition. In particular, the E-step chooses q_i to be a distribution that minimizes the KL distance function, which corresponds to the estimated posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, \hat{M}_p)$ with $\hat{M}_p$ from the previous step estimates. The E-step then evaluates the expectation with respect to q_i's, i.e.,

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

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

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

To avoid directly evaluating the posterior distribution of $\boldsymbol{\theta}_i$, the variational inference method uses alternative choices of the $q_i(\boldsymbol{\theta}_i)$'s to approximate the marginal likelihood function. The choices of $q_i(\boldsymbol{\theta}_i)$ not only approximate the posterior $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ well, but also are easy to compute and usually give closed form evaluations in the algorithm. In particular, from the MIRT literature, we know that as the number of items J becomes reasonably large, the posterior distribution $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ can be well approximated by a Gaussian distribution (Bishop, 2006). Motivated by this observation, we use the Gaussian approximation procedure that chooses $q_i(\boldsymbol{\theta}_i)$ from a family of Gaussian distributions such that the KL distance between $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ is minimized. The estimation is then taken as a two-step iterative procedure. In the variational E-step, we choose $q_i(\boldsymbol{\theta}_i)$ by minimizing the KL distance between $q_i(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ and evaluate the expectation of the likelihood function with respect $q_i(\boldsymbol{\theta}_i)$, which is (II.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.

II.3 GVEM for the M2PL Model

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

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

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

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

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

where $\xi_{i,j}$ is a variational parameter that is introduced to approximate the objective function $e^{x_{i,j}}/(1 + e^{x_{i,j}})$, and

$$\eta(\xi_{i,j}) = \frac{1}{2\xi_{i,j}} \Big[\frac{e^{\xi_{i,j}}}{(1 + e^{\xi_{i,j}})} - \frac{1}{2} \Big].$$

Therefore, we have the following variational lower bound on the logistic sigmoid function,

$$\frac{e^{x_{i,j}}}{(1 + e^{x_{i,j}})} \geq \frac{e^{\xi_{i,j}}}{(1 + e^{\xi_{i,j}})} \exp\left\{ \frac{(x_{i,j} - \xi_{i,j})}{2} - \eta(\xi_{i,j})(x_{i,j}^2 - \xi_{i,j}^2) \right\}. \tag{II.6}$$

We then aim to estimate the variational parameter $\xi_{i,j}$ that achieves the equality of the above display. By introducing an additional variational parameter $\xi_{i,j}$, we successfully avoid the problem of estimating the intractable integral in the E-step. The values of $\xi_{i,j}$'s will be iteratively updated in the M-step.

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

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

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

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

II.3.1 Algorithm Details

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

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

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

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

and the covariance matrix is determined by

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

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

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

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

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

$$\begin{aligned}
\boldsymbol{\alpha}_j &= \frac{1}{2}\Big[\sum_{i=1}^{N}\eta(\xi_{i,j})\Sigma_i + \eta(\xi_{i,j})\mu_i\mu_i^\top\Big]^{-1}\sum_{i=1}^{N}\Big[\Big(Y_{ij} - \frac{1}{2} + 2b_j\eta(\xi_{i,j})\Big)\mu_i^\top\Big], && \text{(II.9)}\\
b_j &= \frac{\sum_{i=1}^{N}\Big[(\frac{1}{2} - Y_{ij}) + 2\eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top\mu_i\Big]}{\sum_{i=1}^{N}2\eta(\xi_{i,j})}, && \text{(II.10)}\\
\xi_{i,j}^2 &= b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top\mu_i + \boldsymbol{\alpha}_j^\top[\Sigma_i + \mu_i\mu_i^\top]\boldsymbol{\alpha}_j. && \text{(II.11)}
\end{aligned}$$

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

$$\Sigma_{\boldsymbol{\theta}} = \frac{1}{N}\sum_{i=1}^{N}[\Sigma_i + \mu_i\mu_i^\top]. \qquad \text{(II.12)}$$

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

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

Algorithm 1 GV-EM algorithm

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

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

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

II.3.2 Theoretical Properties

In this section, we establish theoretical bounds on the estimation of the model parameters under the high-dimensional setting where both N and J go to infinity. The dimension of latent traits, K, is assumed known for this analysis and thus fixed. As defined in Section II.2, $\boldsymbol{A} = [\alpha_{jk}]_{J\times K}$ denotes a matrix of factor loadings. Additionally, let $\Theta = [\theta_{ij}]_{N\times K}$

denote a matrix of random variables following $q_i(\boldsymbol{\theta}_i)$ and let $\hat{\Theta} = [\hat{\theta}_{ij}]_{N \times K}$ denote a matrix of estimated latent abilities from data. Define $E_{\hat{\boldsymbol{\theta}} \sim \hat{q}}$ to be the expectation with respect to the estimated variational densities $\{\hat{q}_i(\hat{\boldsymbol{\theta}}_i) \sim N(\hat{\mu}_i, \hat{\Sigma}_i) : i = 1, \ldots, N\}$ from data. Lastly, a superscript $*$ denote a true parameter. For example, $\boldsymbol{\theta}_i^*$ denotes the i^{th} person's true latent ability, which is a deterministic realization from its population distribution. We assume that the true parameters Θ^* and $\boldsymbol{A}^*$ satisfy

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

Theorem II.3 derives the bound on the expected Frobenius norm of the error, $\|\hat{\Theta}\hat{\boldsymbol{A}}^\top - \Theta^*(\boldsymbol{A}^*)^\top\|_F$, where $\|M\|_F = \sqrt{\sum_{i,j} M_{ij}^2}$ denotes the Frobenius norm of a matrix M.

Theorem II.3. *Suppose that condition **(A1)** is satisfied for the true parameters Θ^* and $\boldsymbol{A}^*$. With optimally estimated variational densities $\hat{q}_i$ from data and estimated parameter matrix $\hat{\boldsymbol{A}}$ that maximizes the variational lower bound, there exists absolute constants C_1 and C_2 such that*

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

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

The proof of Theorem II.3 can be found in the Supplementary Material.

Remark II.4. *Theorem II.3 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 II.3 follows a similar argument from Davenport, Plan, Van Den Berg, and Wootters (2014) and Theorem 1 in Chen et al. (2019). However, the previous work by Chen et al. (2019) treats $\boldsymbol{\theta}_i$ as fixed effects while this work follows the conventional MIRT model setting with $\boldsymbol{\theta}_i$ random effects and following a normal population distribution.*

Remark II.5. *The Gaussian family as the candidate choice of q is reasonable according to Laplace approximation of the posterior distribution $P(\boldsymbol{\theta}_i|Y_i)$. The Laplace approximation* *of $P(\boldsymbol{\theta}_i|Y_i)$ is a normal distribution with MLE $\hat{\boldsymbol{\theta}}_i$ as mean and inverse of observed Fisher information $I^{-1}(\hat{\boldsymbol{\theta}}_i)$ as variance. Denote $\boldsymbol{\theta}_i^*$ as the true parameter. By Bernstein-von Mises Theorem, since $P(Y_i \mid \boldsymbol{\theta}_i), i = 1, \ldots, N$ have same support and $\boldsymbol{\theta}_i \to \log P(Y_i \mid \boldsymbol{\theta}_i)$ is twice continuously differentiable, then $\hat{\boldsymbol{\theta}}_i \to \boldsymbol{\theta}_i^*$ almost surely and the Laplace approximated distribution $N(\hat{\boldsymbol{\theta}}_i, I^{-1}(\hat{\boldsymbol{\theta}}_i))$ converges in distribution to the true limiting normal distribution $N(\boldsymbol{\theta}_i^*, I^{-1}(\boldsymbol{\theta}_i^*))$ as $J \to \infty$ where $I^{-1}(\boldsymbol{\theta}_i^*)$ is the inverse of expected Fisher information. This supports our choice of variational density q_i as a multivariate Gaussian distribution provides an asymptotically good approximation for the true posterior distribution of $\boldsymbol{\theta}$.*

Remark II.6. *Compared with the existing stochastic estimation algorithms, such as the Metropolis-Hastings Robbins-Monro algorithm and the stochastic EM algorithm, the proposed estimation method has the advantage that each of the estimation iterations has simple closed-form update and it does not involve the stochastic samplings from some intermediate posterior distributions as in the current stochastic estimation algorithms. As discussed in Remark II.5, even though variational distributions are used to approximate the posterior distributions in our method, the normal approximation is asymptotically valid. Simulation studies in Section II.5 further illustrate this. Moreover, the above variational EM development can be easily generalized to the M3PL model and can also be naturally combined with the idea of the stochastic EM, as illustrated in the next section.*

II.4 GVEM for the M3PL Model

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

Proposition II.7. *Given the two latent variables $\boldsymbol{\theta}_i$ and Z_{ij}, then $P(Y_{ij} \mid \boldsymbol{\theta}_i)$ under the following hierarchical model is equivalent to* (II.2) *of the 3PL model.*

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

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

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

Without loss of generality we first focus on the ith subject's likelihood function due to the independence of different subjects. Denote $\boldsymbol{Z}_i = \{Z_{i1}, Z_{i2}, \dots, Z_{iJ}\}$ and its distribution as $p(\boldsymbol{Z}_i) = \prod_{j=1}^J p(Z_{ij})$. Then the complete data likelihood of the ith subject is

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

Following the result from Proposition II.7, the hierarchical formulation of the 3PL model with the new latent variable Z_{ij} could be used to derive the GVEM algorithm for the 3PL model. Please refer to the Supplementary Material for the proof of Proposition II.7. Similar data augmentation scheme was proposed in Albert (1992) in the Bayesian framework.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appropriate choices of the variational distributions will lead to a closed form expression of the lower bound expressed in (II.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.

II.4.1 Algorithm Details

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

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

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

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

The above likelihood function implies that $q_i(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \mu_i, \Sigma_i)$ where the mean parameter of the normal distribution is

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

and the covariance matrix is determined by

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplementary Material.

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

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

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

Algorithm 2 GV-EM algorithm for M3PL

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

Remark II.8. *The theoretical property of the M3PL is more challenging to derive rigorously due to the addition of the guessing parameters c_j's. From Theorem 2 in Davenport et al. (2014) we can show that the Hellinger distance of error between estimated probability distributions and the true probability distributions is bounded above. For this discussion, we define Hellinger distance for probability distributions and matrices. Hellinger distance for two scalars $p, q \in [0,1]$ is defined as $d_H^2(p,q) := (\sqrt{p}-\sqrt{q})^2 + (\sqrt{1-p}-\sqrt{1-q})^2$. Following Davenport et al. (2014), we also allow the Hellinger distance to act on matrices by averaging Hellinger distances over their entries. For matrices $P, Q \in [0,1]^{d_1 \times d_2}$, we de-*

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

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

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

II.4.2 Stochastic Optimization of GVEM

In M3PL, the proposed GVEM algorithm may become computationally inefficient as sample size increases because of the additional variational parameters and model parameters to estimate compared to M2PL. Especially in the E step, variational parameters (i.e. $\mu_i, \Sigma_i, \xi_{i,j}, s_{ij}$) need to be optimized for every data points $i = 1, \ldots, N$. Thus, the computational burden increases with larger sample size N. To improve the computational efficiency of the GVEM algorithm, we can stochastically optimize the variational approximation in the E step (Hoffman, Blei, Wang, & Paisley, 2013). That is, at each iteration of the E step, we subsample the data to form noisy estimate of the variational lower bound and iteratively update the estimate with a decreasing step size. Then M step in Algorithm 2 follows using this stochastically estimated variational lower bound. The stochastic optimization only affects the E step, thus with minor changes to the original GVEM algorithm we can stochastically optimize the algorithm for M3PL. The noisy estimates of the variational lower bound are cheaper to compute as it only requires small subset of the data at each iteration. Also, for complicated models like M3PL, following such noisy estimates can also help the algorithm to escape local optima of complex objective functions. Specifically, the stochastic EM steps can be summarized as follows.

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

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

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

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

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

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

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

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

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

II.5 Simulations

II.5.1 Design

A series of simulation studies were conducted to evaluate the performance of the proposed GVEM algorithm in comparison to the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm implemented in the R package, 'mirt' (Chalmers, 2012)[1]. The Metropolis-Hastings sampler is used to draw missing data (which is $\boldsymbol{\theta}$ in MIRT) in the stochastic imputation step of the MH-RM algorithm (Cai, 2008, 2010a). In the 'mirt' package, "MHcand" is a vector of values used to tune the MH sampler, with larger values yielding lower acceptance rate. By default, these values are determined internally and adjusted on-the-fly, attempt-

[1] Please note that our conclusions regarding the MH-RM algorithm is based on the implementation of the algorithm in the 'mirt' package. Researchers using other packages may get slightly different results. Thoroughly evaluating the MH-RM algorithm in 'mirt' is beyond the scope of this thesis, but our preliminary check of the package revealed that the MH-RM results are credible.

ing to tune the acceptance of the draws to be between .1 and .4. In addition, the default number of Metropolis-Hastings draws at each iteration is 5, which is considered sufficient by Cai (2010a). Only the exploratory item factor analysis will be presented since it is a computationally more challenging scenario than the confirmatory analysis. That is, in the confirmatory analysis, many of the item loading parameters (or discrimination parameters) are constrained to 0 based on the pre-specified item factor loading structure. Hence, the update equation for $\boldsymbol{\alpha}$ (i.e., (II.9) for the 2PL model and (II.20) for the 3PL model) only needs minimum updates to reflect the constraints specified in the factor loading structure. In the exploratory analysis, we do not assume any constraint on the item discrimination parameter $\boldsymbol{A}$. Instead, to ensure model identifiability, we fix $\Sigma_{\boldsymbol{\theta}} = I_K$ during the estimation. A post-hoc rotation can then follow to rotate the factors and allow them to be correlated. The best-known rotation methods available in most commercial software packages are varimax (Kaiser, 1958) in orthogonal rotation or promax (Hendrickson & White, 1964) in oblique rotation. Other popular methods include, for instance, the CF-Quartimax rotation (Browne, 2001). In the simulations studies, the promax rotation was used such that the factors were allowed to be correlated. Both the M2PL and M3PL were considered in the simulation studies. The number of dimensions was fixed at 3 and test length was fixed at 45.

Additionally, we compared the performance of GVEM to the joint maximum likelihood (JML) estimator given that the JML estimator is also shown to be consistent under the same high-dimensional setting presented in Theorem II.3 and efficient (Chen et al., 2019). The JML estimator was computed using the default settings in the R package ‘mirtjml’ implemented by Chen et al. (2019). Since Chen et al. (2019) did not study M3PL, here we only compared the performances for M2PL.

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

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

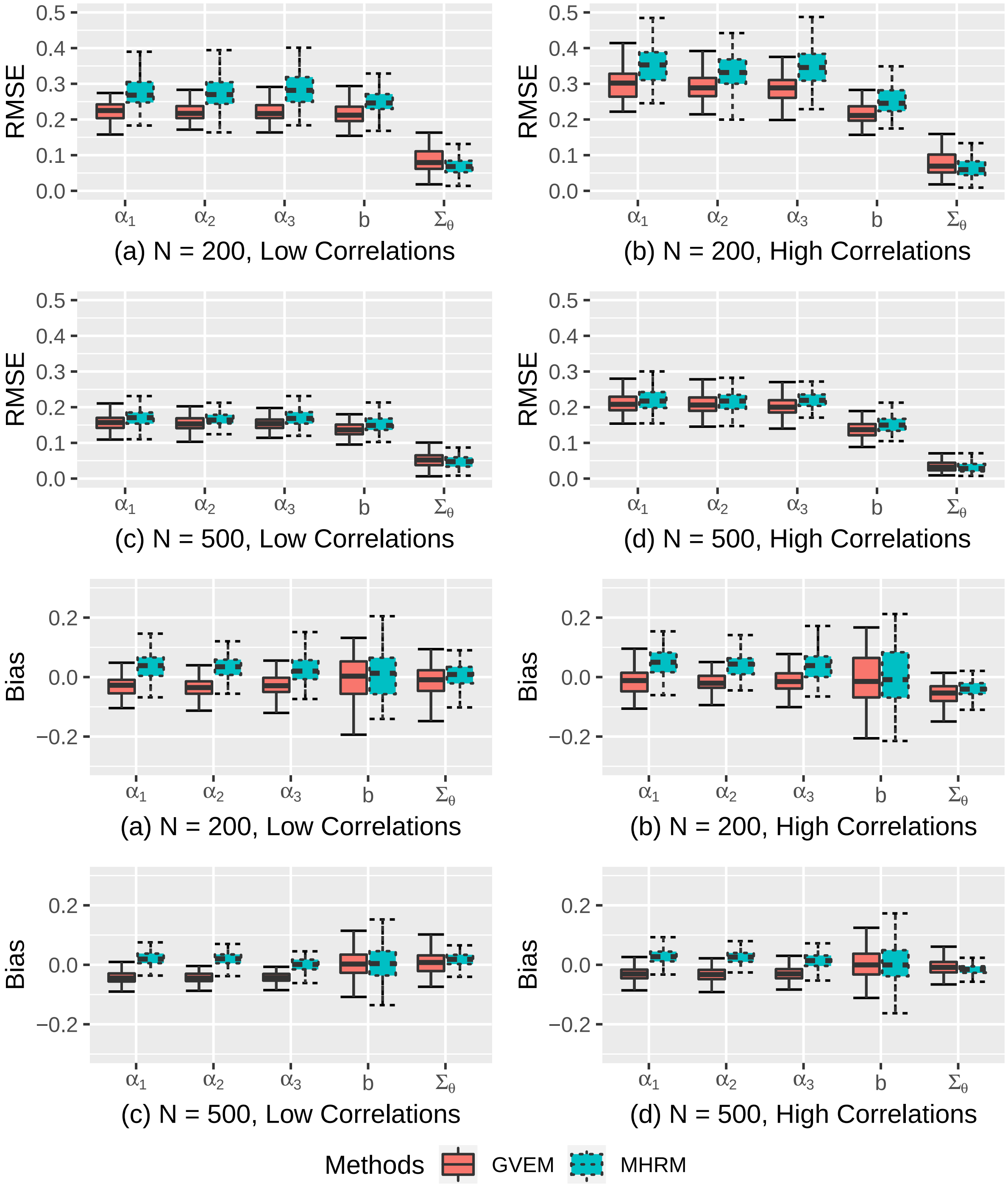
II.5.2 Results for the M2PL model

Figures II.1 and II.2 compare the distributions of bias and RMSE of the model parameters from the two methods under the four manipulated conditions for the between-item and within-item M2PL model respectively. As shown, GVEM generally produces comparable or more accurate model parameter estimates than MHRM [2] in all conditions for both between-item and within-item models. With respect to the manipulated conditions, increasing sample

[2] MHRM method is run by the R package 'mirt'

sizes helps reduce the RMSE and bias of the parameter estimates in both GVEM and MHRM. Moreover, the RMSE and bias are generally higher when the correlations among factors are higher. This may be because higher correlation introduce multicollinearity among factors, making the parameter recovery more difficult (C. Wang & Nydick, 2015). Last but not least, the parameter recovery from the between-item multidimensional model is better than the parameter recovery from the within-item multidimensional model. This is not surprising since the loading structure $\boldsymbol{A}$ is more complex in the within-item model. Figures II.3 and II.4 compare the distribution of bias and RMSE of the model parameters from GVEM and the JML method under the four manipulated conditions for the between-item and within-item M2PL models respectively. We observe that GVEM produces much lower RMSE and bias than the JML estimation under all conditions for both between-item and within-item models. The performance of the JML estimator is especially worse in small sample and high correlation settings and under more complex within-item multidimensionality structure. This could be due to the fact that the JML estimator assumes $\boldsymbol{\theta}_i$'s as fixed effects whereas GVEM models them as random effects with multivariate Gaussian distributions which account for the factor correlations. This result suggests that our proposed estimation method not only is theoretically consistent but also performs better in practice particularly under these complex simulation settings with correlated latent factors.

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



(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

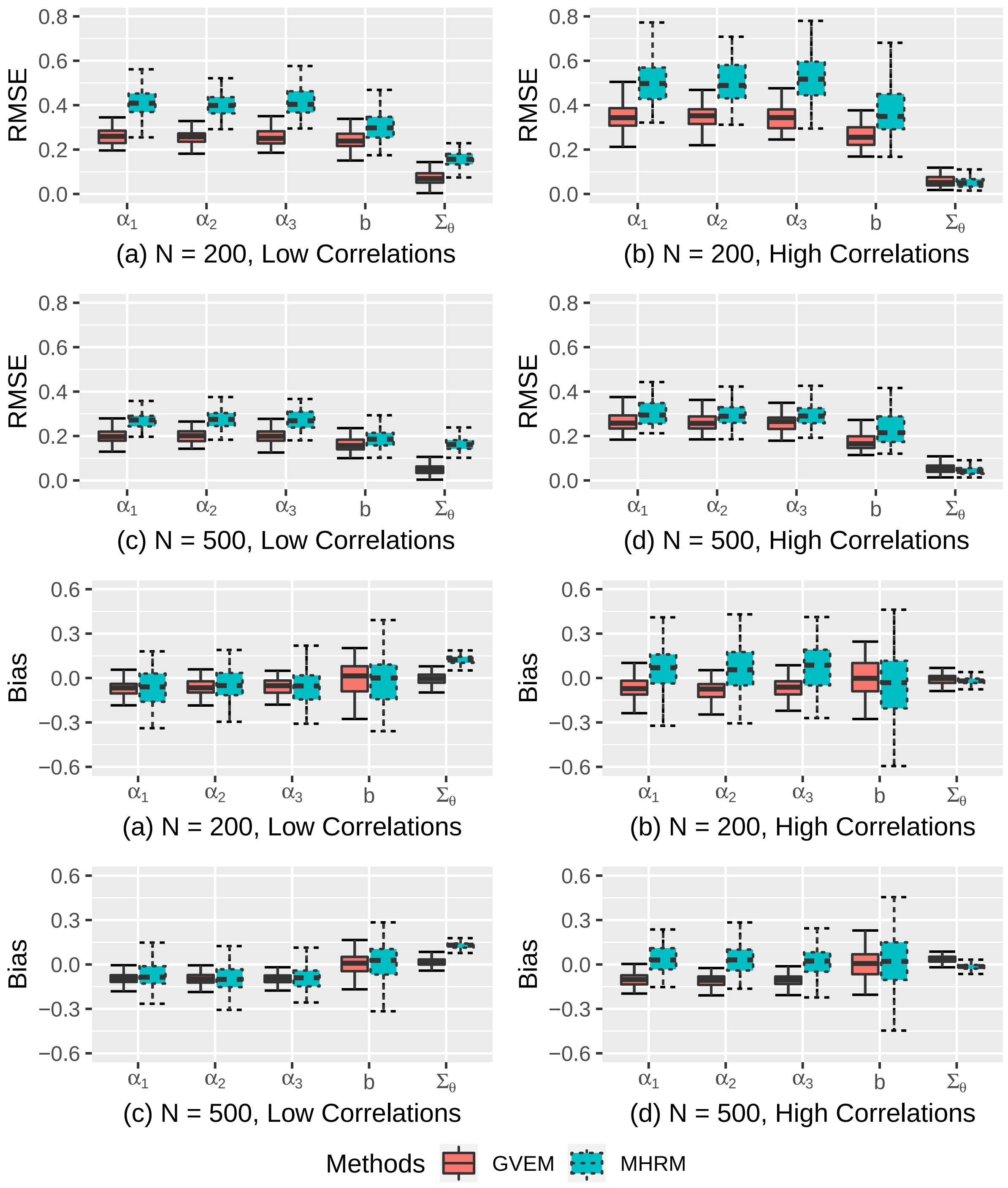
(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

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



(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

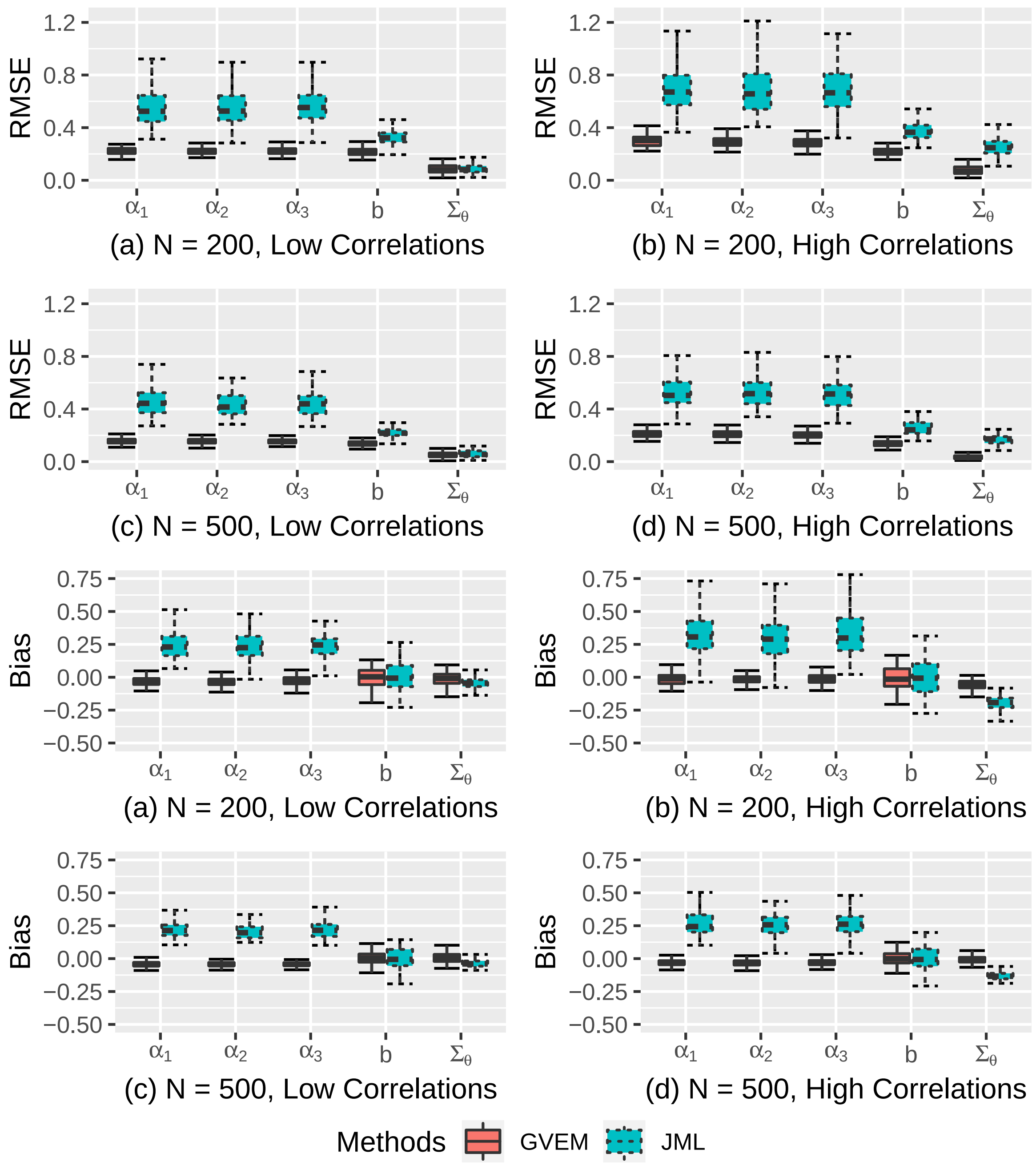
(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

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



(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

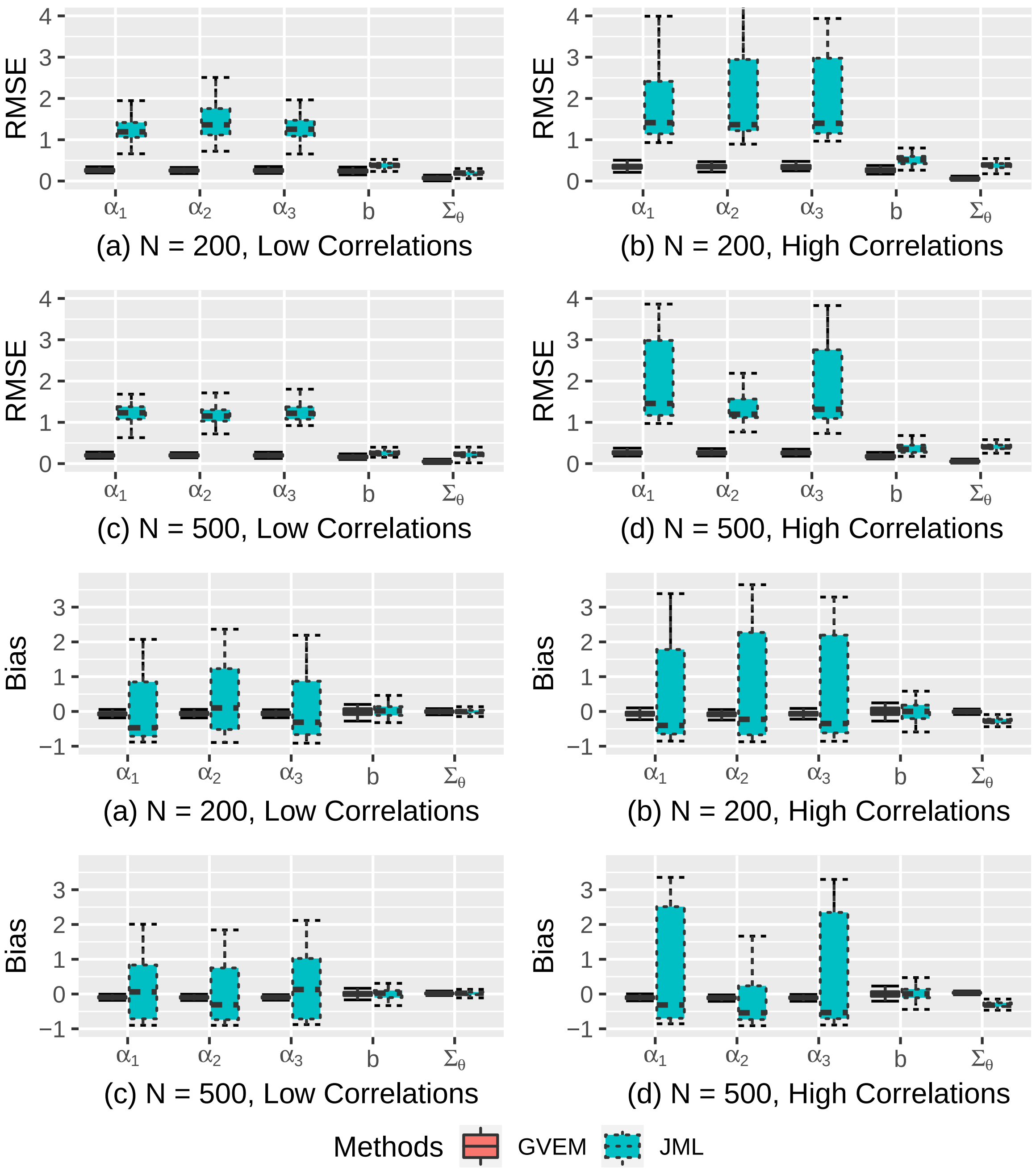
(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

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



(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

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

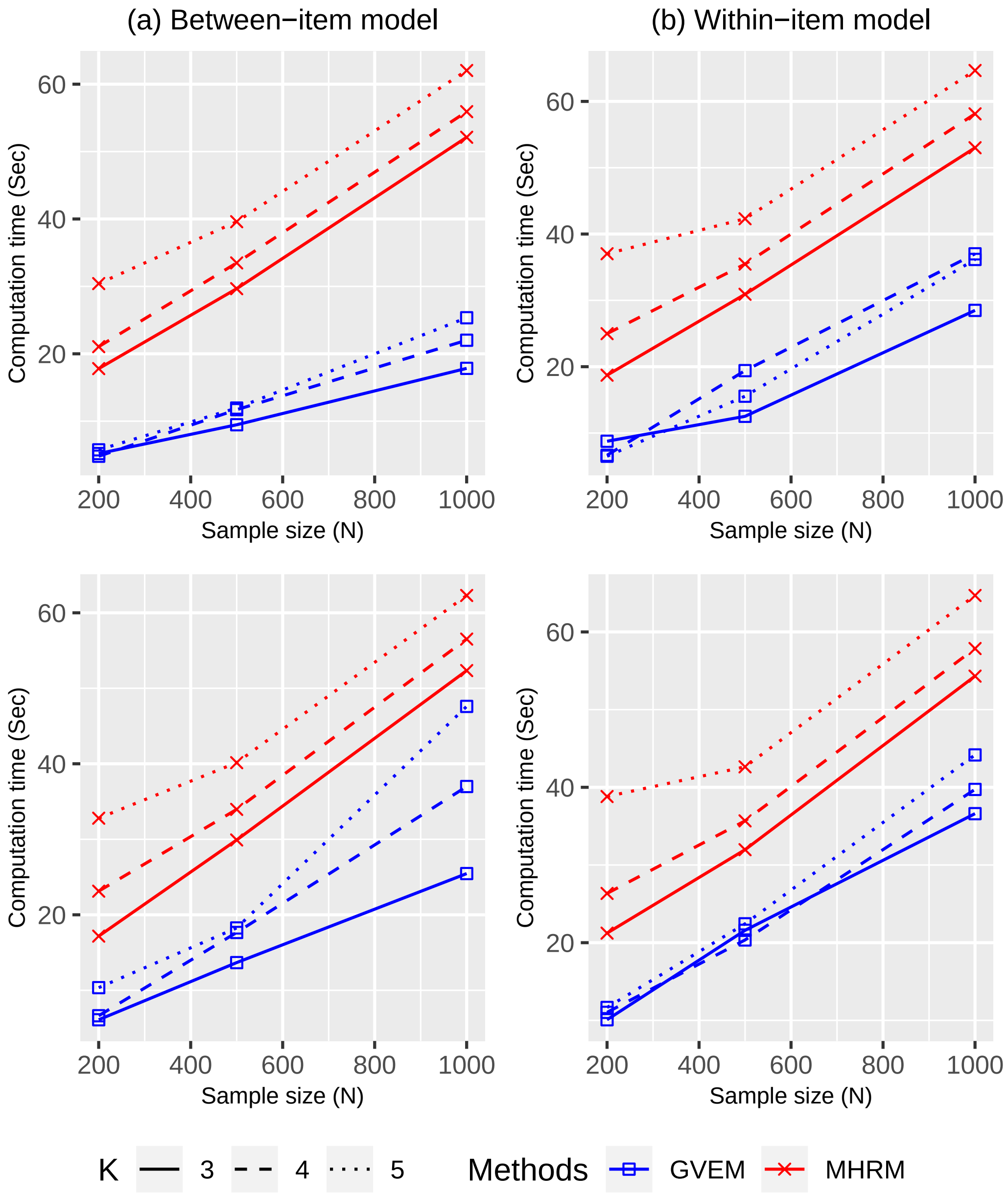


Figure II.5: Average computation time for (a) between-item model (first column) and (b) within-item model (second column) with low correlation (first row) and high correlation (second row).

II.5.3 Results for the M3PL model

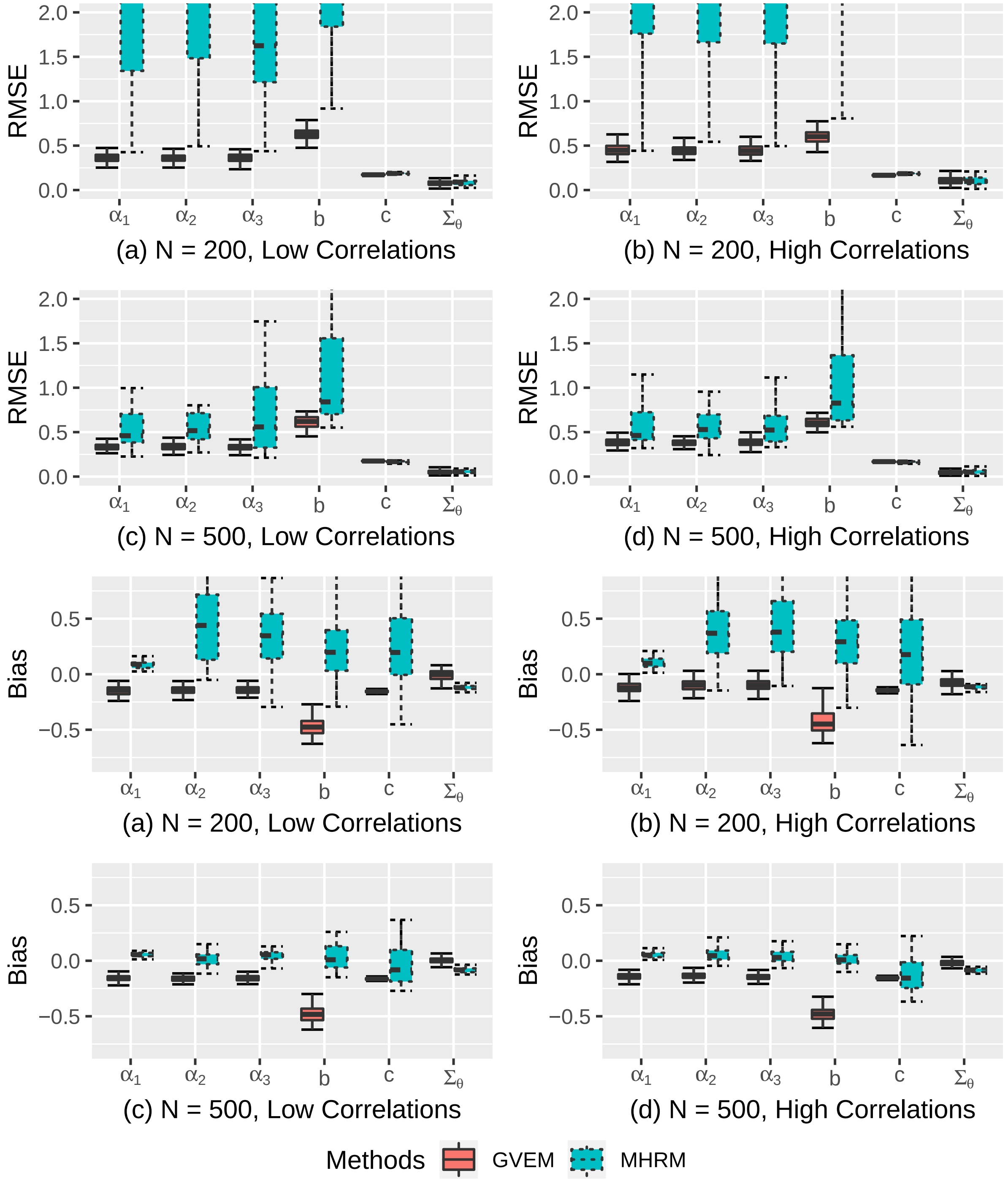
For the M3PL model, the sample size and forget rate for stochastically optimized 3PL algorithm were chosen based on pilot testing of various sample sizes and forget rates. We observed that using the whole data set for the initial estimation step helped a lot with the estimation precision. Hence the forget rate was fixed at a small value so that the information from entire data set in the first iteration was weighted more heavily in the subsequest iterations (i.e. forget the information from entire data set slowly with small forget rate). After the first iteration, only 5 data points were sampled at a time, resulting in a huge reduction in computation time.

Figures II.6 and II.7 present the distributions of bias and RMSE of the model parameters from the two methods under the four manipulated conditions for the between-item and within-item M3PL model, respectively. During simulation studies, we observed that the performance of MHRM was quite unstable and the model did not converge well in M3PL under all manipulated conditions. Specifically, model did not converge in about 30 to 45% of the total experiments in most conditions. In another 15 to 20% of the experiments, the model converged but the estimates of the model parameters exploded to surprisingly high values, which implies the instability of the parameter estimation. For MHRM method, we excluded these results from the total of 100 experiments and reported only the values that seem more meaningful. On the other hand, we report the results for all 100 experiments for the GVEM method. Precisely, in Figure II.6, 40 cases for (a), 41 for (b), 28 for (c), and 40 for (d) were reported. In Figure II.7, 48 cases for (a), 46 for (b), 54 for (c), and 47 for (d) were reported. Note again that in both Figures, we report all 100 experiments for GVEM method because they all converged successfully. Similarly as in the simulation studies for M2PL, increasing sample sizes helps reduce the RMSE and bias of the parameter estimates in both GVEM and MHRM. However, the RMSE for MHRM method is quite high with large variation under most conditions. Overall, we observe that for varying sample sizes and correlations between latent traits, GVEM performs better than MHRM, even after excluding unstable estimation results for MHRM. Given that the inclusion of guessing parameter poses model estimation challenge is well-documented in literature (e.g, Lord, 1968; Thissen & Wainer, 1982; Yen, 1987), it is not too surprising to note the large proportion of non-converged replications from MHRM. However, the stable performance of GVEM further reinforces its promise as a robust alternative method to the current status-quo, in particular when guessing parameter is included in the model. Also note that GVEM does not need much tuning for good performance, hence it is more accessible to broader audience who may not have the technical capacity to manually tune certain parameters, as may required by other algorithms.

II.5.4 Estimating the Number of Dimensions

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

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



(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

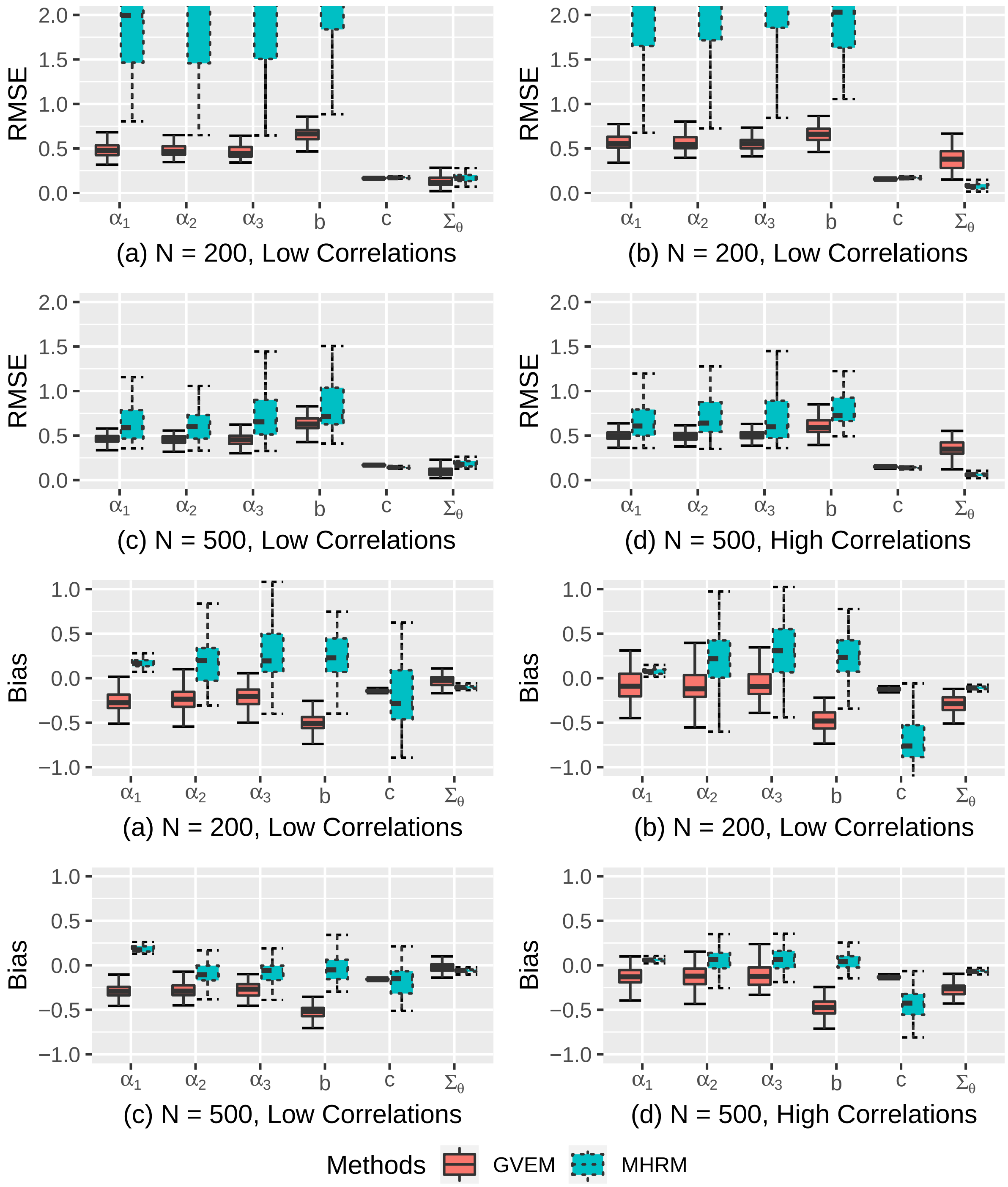
(a) N = 200, Low Correlations

(b) N = 200, High Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

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



(a) N = 200, Low Correlations

(b) N = 200, Low Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

(a) N = 200, Low Correlations

(b) N = 200, Low Correlations

(c) N = 500, Low Correlations

(d) N = 500, High Correlations

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

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

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

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

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

II.6 Real Data Analysis

In this section, the GVEM and MHRM algorithms were used to conduct an exploratory item factor analysis on the National Education Longitudinal Study of 1988 (NELS:88) data. In this data set, a nationally representative sample of approximately 24,500 students were tracked via multidimensional cognitive batteries from 8th to 12th grade (the first three studies) in years 1988, 1990, and 1992. In this study, we focused on the science and mathematics test data where the multidimensional factorial structure has been previously investigated (e.g, Kupermintz & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997). For the science subject, there are 25 items and four factors emerged from the data collected in 1988: “Elementary science (ES)”, “Chemistry knowledge (CK)”, “Scientific reasoning (SR)” and “Reasoning with knowledge (RK)”. For the math subject, there are 40 items in 1988 and two factors emerged, they are “Mathematical reasoning (MR)” and “Mathematical knowledge (MK)”. We pooled together data from both domains, resulting in 65 items and a complete sample size of N=13,488. Because the factor structure was analyzed using normal theory factor analysis more than two decades ago, we plan to reanalyze the data using the proposed new methods. In addition, pooling together both math and science domains result in potentially high dimensional data. First, both GVEM and MHRM were conducted assuming the number of factors were 6. The focus is on the recovery of the correlation matrix $\Sigma_{\boldsymbol{\theta}}$ and its comparison between two methods. Since the exploratory item factor analysis was conducted, in both GVEM and MHRM we assumed that $\Sigma_{\boldsymbol{\theta}} = I_K$ during GVEM estimation and later performed the same promax rotation to estimate the correlation matrix $\hat{\Sigma}_{\boldsymbol{\theta}}$. Second, GVEM was used to explore the dimension of latent traits from the data.

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

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

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

To further explore the optimal number of factors from data, we applied the GVEM algorithm with the information criteria for dimension selection. Figure II.8 presented the results of latent dimension selection under M2PL and M3PL models. By fitting the M2PL model to the data, the optimal dimensionality of the latent traits was estimated to be six by

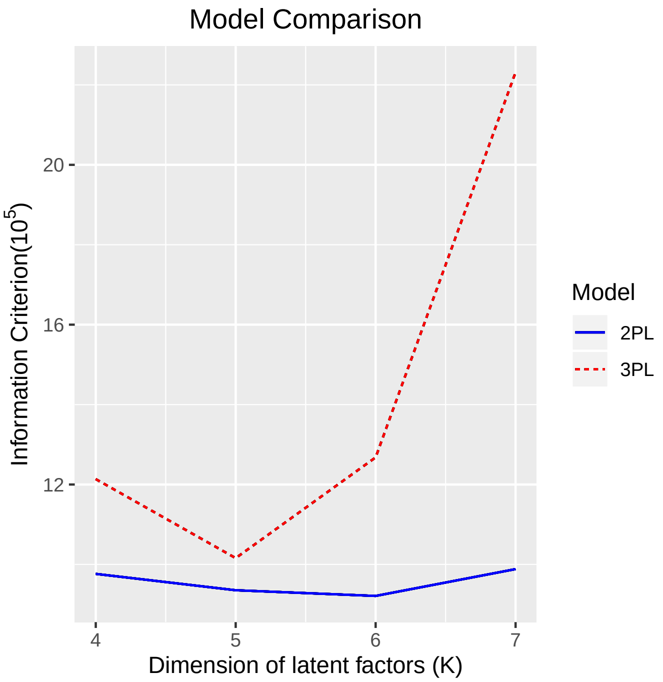


Figure II.8: Real Data: Model Selection ($BIC^\star$ for both M2PL and M3PL. $AIC^\star$ shows the same trend).

both $AIC^\star$ and $BIC^\star$ as shown in Figure II.8. This corresponds to the number of latent traits identified in prior research. However, the dimensionality of the latent traits was estimated to be five under the M3PL model. This result implies that some of the six latent traits may be highly correlated under the M3PL model that they are merged. Comparing the information criteria values across both M2PL and M3PL, it appears that $AIC^\star$ and $BIC^\star$ were smallest for the M2PL model with six factors. Hence, our results further validate the number of latent factors that could be extracted from the NELS:88 data. In addition, it suggests that the guessing didn't play a significant role in students' performance on the math and science cognitive test data.

II.7 Discussions

Variational methods are first introduced in psychometrics by Rijmen and Jeon (2013) for high dimensional IRT model with discrete latent traits, and later by Jeon et al (2017) in a form of a variational maximization-maximization algorithm for GLMMs with crossed random effects. Although their findings demonstrate great promise of variational methods as they apply in psychometrics, their methods are not ready for calibrating high-dimensional MIRT models with correlated latent factors and guessing parameters. In this chapter, a new method based on variational approximation is proposed for the parameter estimation in the M2PL and M3PL models. Compared to the existing methods, it has the advantage of avoiding the calculation of intractable log-likelihood by approximating the lower bound to the log-likelihood. It also greatly reduces the computation complexity by deriving the closed-form updates in the every EM step. Moreover, the efficiency of the algorithm is further improved in the stochastic version. Simulation studies demonstrate that the proposed methods show better performance in terms of parameter recovery and computation time in both M2PL and M3PL compared to the widely used MHRM method. Theoretical results are provided on the convergence rate, which shows that the estimation error goes to 0 as both the sample size and number of test items go to infinity. As byproducts of the GVEM algorithm, both $AIC^\star$ and $BIC^\star$ could be used to help identify the optimal number of latent factors from data, as reflected by the simulation results.

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

Taking one step further, the GVEM algorithm could be coupled with latent variable selection (Sun, Chen, Liu, Ying, & Xin, 2016). Traditional approaches for identifying item factor loading structure proceeds in two steps: (i) allowing all item factor loadings to be freely estimated, and (ii) conducting a post-hoc rotation (Browne, 2001). While these rotation methods intend to produce a near-simple structure for the ease of interpretation, an arbitrary cut-off for the rotated factor loadings is often needed. In contrast, the latent variable selection avoids setting subjective cut-offs. The principle idea is to estimate the non-zero elements in the $\boldsymbol{A}$ matrix. Specifically, a penalty will be added to elements in $\boldsymbol{A}$ and when a factor is not associated with an item, the corresponding element in $\boldsymbol{A}$ will shrink to 0. Hence, this is a one-step approach where model calibration and factor selection are completed simultaneously. This idea was first proposed by Sun et al. (2016), but they still used a traditional EM algorithm that can hardly be generalized to higher dimensions due to the computation burden. The GVEM algorithm proposed in this study is a good candidate for such one-step latent variable selection, and future studies could explore this direction.

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

Appendix of Chapter II

In this section, we present the derivations and proofs for the GVEM algorithms and theorems. Appendix A and C illustrate the derivation of the GVEM algorithm for M2PL and M3PL models, respectively. In Appendix B, we present the proof for the Proposition II.7. Finally in Appendix D, we prove the consistency result in Theorem II.3.

II.A Derivation of GVEM in the 2PL model

Here, we provide a step by step derivation of the GVEM algorithm in 2PL model. Especially, EM steps are described in detail.

E-Step In E step, we evaluate the lower bound of the expected log-likelihood with respect to the variational distributions $q_i(\boldsymbol{\theta}_i)$'s. Recall that we can mathematically write the lower bound as

$$E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}) := \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} l(Y_i, \boldsymbol{\theta}_i, \boldsymbol{\xi}_i \mid \boldsymbol{A}, \boldsymbol{B}) \times q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \tag{II.A.1}$$

Our main interest is to evaluate the integral in (II.A.1) to derive a closed-form expression of the variational lower bound, $E(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi})$. In each E step, we iteratively update the lower bound until convergence.

The optimal variational density $q_i(\boldsymbol{\theta}_i)$ is chosen as a Gaussian distribution with mean and covariance determined by

$$\begin{aligned} \mu_i &= \Sigma_i \times \sum_{j=1}^{J} \Big\{ 2\eta(\xi_{i,j}) b_j + Y_{ij} - \frac{1}{2} \Big\} (\boldsymbol{\alpha}_j)^\top, && \text{(II.A.2)} \\ \Sigma_i^{-1} &= (\Sigma_{\boldsymbol{\theta}})^{-1} + 2 \sum_{j=1}^{J} \eta(\xi_{i,j}) \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^\top. && \text{(II.A.3)} \end{aligned}$$

Let $q_i^{(t)}(\boldsymbol{\theta}_i) = q_i(\boldsymbol{\theta}_i \mid \boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \Sigma_{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\xi}_i^{(t)})$ denote the tth iteration's variational density $q_i(\boldsymbol{\theta}_i)$ with all recent updates of the model parameters $(\boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \Sigma_{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\xi}_i^{(t)})$. Also, let $E_{q_i}^{(t)}$ denote the expectation with respect to the distribution $q_i^{(t)}(\boldsymbol{\theta}_i)$. Then, we can write the tth iteration's variational lower bound as

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

Note that the expectation $E_{q_i}^{(t)}$ can be expressed with $\mu_i^{(t)}$ and $\Sigma_i^{(t)}$ since

$$E_{q_i}^{(t)}[\boldsymbol{\theta}_i] = \mu_i^{(t)}, \quad E_{q_i}^{(t)}[(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2] = b_j^2 - 2 b_j \boldsymbol{\alpha}_j^\top \mu_i^{(t)} + (\boldsymbol{\alpha}_j^{(t)})^\top [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top] \boldsymbol{\alpha}_j^{(t)},$$

and

$$E_{q_i}^{(t)}[\boldsymbol{\theta}_i^\top \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\theta}_i] = E_{q_i}^{(t)}[Tr(\Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\theta}_i \boldsymbol{\theta}_i^\top)] = Tr(\Sigma_{\boldsymbol{\theta}}^{-1} E_{q_i}^{(t)}[\boldsymbol{\theta}_i \boldsymbol{\theta}_i^\top]) = Tr(\Sigma_{\boldsymbol{\theta}}^{-1} [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]).$$

Therefore, by plugging in we have the following equivalent form.

$$
\begin{aligned}
E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi}) &= \sum_{i=1}^{N}\sum_{j=1}^{J}\Bigg(\log\frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + Y_{ij}(\boldsymbol{\alpha}_j^\top \mu_i^{(t)} - b_j) + \frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top \mu_i^{(t)} - \xi_{i,j}) \\
&\quad -\eta(\xi_{i,j})\{b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top \mu_i^{(t)} + (\boldsymbol{\alpha}_j^{(t)})^\top[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]\boldsymbol{\alpha}_j^{(t)} - \xi_{i,j}^2\}\Bigg) \\
&\quad +\frac{N}{2}\log|\Sigma_{\boldsymbol{\theta}}^{-1}| - \sum_{i=1}^{N}\frac{1}{2}Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{J}\Bigg(\log\frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + (\frac{1}{2} - Y_{ij})b_j + (Y_{ij} - \frac{1}{2})\boldsymbol{\alpha}_j^\top \mu_i^{(t)} - \frac{1}{2}\xi_{i,j} \\
&\quad -\eta(\xi_{i,j})\{b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top \mu_i^{(t)} + \boldsymbol{\alpha}_j^\top[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]\boldsymbol{\alpha}_j - \xi_{i,j}^2\}\Bigg) \\
&\quad +\frac{N}{2}\log|\Sigma_{\boldsymbol{\theta}}^{-1}| - \sum_{i=1}^{N}\frac{1}{2}Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]). \qquad \text{(II.A.4)}
\end{aligned}
$$

This gives a closed form expression of the expectation function $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi})$, i.e the tth iteration's variational lower bound of the marginal likelihood. In every E step, we iteratively update $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi})$ (i.e. (II.A.4)) with all recently updated model parameters for $t \geq 1$ steps.

M-Step In tth iteration's M step, we maximize the $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi})$ to estimate the parameters $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi})$. This is achieved by setting the derivative of $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi})$ with respect to $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi})$ to be zero.

First, consider the $\boldsymbol{\alpha}_j$. Setting the derivative with respect to $\boldsymbol{\alpha}_j$ equal to zero, we have

$$
\frac{\partial E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\xi})}{\partial \boldsymbol{\alpha}_j} = \sum_{i=1}^{N}(Y_{ij} - \frac{1}{2})(\mu_i^{(t)})^\top + 2b_j\eta(\xi_{i,j})(\mu_i^{(t)})^\top - 2\eta(\xi_{i,j})[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]\boldsymbol{\alpha}_j = 0
$$

which implies that $\boldsymbol{\alpha}_j^{(t+1)}$ is updated according to

$$
\boldsymbol{\alpha}_j = \frac{1}{2}\Big[\sum_{i=1}^{N}\eta(\xi_{i,j})\Sigma_i^{(t)} + \eta(\xi_{i,j})(\mu_i^{(t)})(\mu_i^{(t)})^\top\Big]^{-1}\sum_{i=1}^{N}\Big[\Big(Y_{ij} - \frac{1}{2} + 2b_j\eta(\xi_{i,j})\Big)(\mu_i^{(t)})^\top\Big]. \qquad \text{(II.A.5)}
$$

Similarly, set the derivative with respect to b_j equal to zero and we have

$$\frac{\partial E^{(t)}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi})}{\partial b_j} = \sum_{i=1}^{N} (\frac{1}{2} - Y_{ij}) - 2\eta(\xi_{i,j}) b_j + 2\eta(\xi_{i,j}) \boldsymbol{\alpha}_j^\top \mu_i^{(t)} = 0$$

which implies that $b_j^{(t+1)}$ is updated according to

$$b_j = \frac{\sum_{i=1}^{N} \left[(\frac{1}{2} - Y_{ij}) + 2\eta(\xi_{i,j}) \boldsymbol{\alpha}_j^\top \mu_i^{(t)} \right]}{\sum_{i=1}^{N} 2\eta(\xi_{i,j})}. \tag{II.A.6}$$

Setting the derivative with respect to $\xi_{i,j}$ equal to zero, we have

$$\begin{aligned} 0 = \frac{\partial E^{(t)}(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi})}{\partial \xi_{i,j}} &= \frac{1}{(1 + e^{\xi_{i,j}})} - \frac{1}{2} - \eta'(\xi_{i,j}) \{ E_{q_i}^{(t)} [(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2] - \xi_{i,j}^2 \} + 2\eta(\xi_{i,j}) \xi_{i,j} \\ &= -\eta'(\xi_{i,j}) \{ E_{q_i}^{(t)} [(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2] - \xi_{i,j}^2 \}. \end{aligned}$$

This implies that $\xi_{i,j}^{(t+1)}$ is updated according to the equation

$$\xi_{i,j}^2 = E_{q_i}^{(t)} [(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2] = b_j^2 - 2 b_j \boldsymbol{\alpha}_j^\top \mu_i^{(t)} + \boldsymbol{\alpha}_j^\top [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top] \boldsymbol{\alpha}_j. \tag{II.A.7}$$

When there is no constraint for Σ, i.e., all parameters of Σ are free, we set the derivative with respect to $\Sigma_{\boldsymbol{\theta}}^{-1}$ to be 0 and we obtain

$$0 = \frac{N}{2} \frac{\partial \log |\Sigma_{\boldsymbol{\theta}}^{-1}|}{\partial \Sigma_{\boldsymbol{\theta}}^{-1}} - \frac{1}{2} \sum_{i=1}^{N} \frac{\partial Tr(\Sigma_{\boldsymbol{\theta}}^{-1} [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top])}{\partial \Sigma_{\boldsymbol{\theta}}^{-1}} = \frac{N}{2} \Sigma_{\boldsymbol{\theta}} - \frac{1}{2} \sum_{i=1}^{N} [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top],$$

which gives the update of $\Sigma_{\boldsymbol{\theta}}^{(t+1)}$ as

$$\Sigma_{\boldsymbol{\theta}} = \frac{1}{N} \sum_{i=1}^{N} [\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]. \tag{II.A.8}$$

Hence, we update $\Sigma_{\boldsymbol{\theta}}$ by (II.A.8) in confirmatory factor analysis. However, in exploratory factor analysis we keep $\Sigma_{\boldsymbol{\theta}} = I_K$ and ignore the step (II.A.8). Note that if the $\Sigma_{\boldsymbol{\theta}}$ is assumed to be the correlation matrix with diagonals being 1, then we can standardize the estimated $\Sigma_{\boldsymbol{\theta}}$ to get correlation matrix.

II.B Proof of Proposition II.7

Proposition II.7 states that the hierarchical formulation of the 3PL model with new latent variable Z_{ij} is equivalent to the general IRT formulation of the 3PL model. This can be proved by showing that the two approaches yield the same distribution, i.e. $P(Y_i \mid \boldsymbol{\theta}_i, M_p)$.

We first start from the hierarchical formulation of the 3PL model. The conditional distribution of Y_{ij} given Z_{ij} and $\boldsymbol{\theta}_i$ can be equivalently written as

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

Then, the joint distribution of a response Y_{ij} and a latent variable Z_{ij} is

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

By summing the joint distribution over the domain of Z_{ij}, we recover the general IRT formulation of the 3PL model.

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

Hence, the result of Proposition II.7 allows us to use the hierarchical formulation of the 3PL model instead of the general IRT formulation for the derivation of GVEM algorithm in the case of M3PL.

II.C Derivation of GVEM in the 3PL model

In 3PL model, the EM steps are derived in the similar fashion as in 2PL model. We again start with the derivation of the E step.

E-Step As in the M2PL model, we estimate the variational parameters first and then compute the variational lower bound on the expected log-likelihood. As previously discussed in Section II.4.1, the choice of the optimal variational density for the first latent variable $\boldsymbol{\theta}_i$ is a Gaussian distribution $q_i(\boldsymbol{\theta}_i)$ with mean and covariance determined by

$$\begin{aligned} \mu_i &= \Sigma_i \times \sum_{j=1}^{J} \Big\{ 2\eta(\xi_{i,j}) b_j + Y_{ij} - \frac{1}{2} \Big\} (1 - Y_{ij} + s_{ij} Y_{ij}) (\boldsymbol{\alpha}_j)^\top, && \text{(II.C.1)} \\ \Sigma_i^{-1} &= (\Sigma_{\boldsymbol{\theta}})^{-1} + 2 \sum_{j=1}^{J} \eta(\xi_{i,j}) (1 - Y_{ij} + s_{ij} Y_{ij}) \boldsymbol{\alpha}_j \boldsymbol{\alpha}_j^\top. && \text{(II.C.2)} \end{aligned}$$

The optimal variational density of the second latent variable Z_{ij} is a Bernoulli distribution $r_{ij}(Z_{ij})$ with the success probability s_{ij} determined by

$$\begin{aligned} s_{ij}^{-1} &= 1 + \frac{c_j}{1 - c_j} \frac{1 + e^{\xi_{i,j}}}{e^{\xi_{i,j}}} \exp\Big\{ - Y_{ij} (\boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i] - b_j) + \\ &\quad \frac{1}{2} (b_j - \boldsymbol{\alpha}_j^\top E_{q_i}[\boldsymbol{\theta}_i] - \xi_{i,j}) - \eta(\xi_{i,j}) \{ E_{q_i}[(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2] - \xi_{i,j}^2 \} \Big\}. && \text{(II.C.3)} \end{aligned}$$

Let the tth iteration's variational densities for the latent variables $\boldsymbol{\theta}_i$ and Z_{ij} be denoted as $q_i^{(t)}(\boldsymbol{\theta}_i) = q_i(\boldsymbol{\theta}_i \mid \boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \boldsymbol{C}^{(t)}, \Sigma_{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\xi}_i^{(t)})$ and $r_{ij}^{(t)}(Z_{ij}) = r_{ij}(Z_{ij} \mid \boldsymbol{A}^{(t)}, \boldsymbol{B}^{(t)}, \boldsymbol{C}^{(t)}, \Sigma_{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\xi}_i^{(t)})$ with all recent updates of the model parameters, respectively. Then, the tth iteration's variational lower bound of the expected log-likelihood is

$$E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi}) := \sum_{i=1}^{N}\int_{\boldsymbol{\theta}_i}\left[\sum_{\boldsymbol{Z}_i} l(Y_i,\boldsymbol{\theta}_i,\boldsymbol{Z}_i,\boldsymbol{\xi}_i \mid \boldsymbol{A},\boldsymbol{B},\boldsymbol{C}) \times r_i^{(t)}(\boldsymbol{Z}_i)\right] \times q_i^{(t)}(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i. \quad \text{(II.C.4)}$$

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

With the variational parameters discussed above (i.e. (II.C.1), (II.C.2), (II.C.3)), we can derive a closed form expression of the variational lower bound. Consistent with the previously defined notations, $E_r^{(t)}$ denotes the expectation with respect to the variational distribution $r_{ij}^{(t)}$'s. Now we evaluate the integrals in expectation function $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ with respect to $q_i(\boldsymbol{\theta}_i)$'s and $r_{ij}(Z_{ij})$'s.

$$\begin{aligned}
& E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi}) \\
= \; & \sum_{i=1}^{N} E_{q_i}^{(t)}\Bigg[\sum_{j=1}^{J}(1-Y_{ij}+E_r^{(t)}[Z_{ij}]Y_{ij})\Bigg(\log\frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + Y_{ij}(\boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i - b_j) \\
& +\frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i - \xi_{i,j}) - \eta(\xi_{i,j})\{(b_j-\boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i)^2 - \xi_{i,j}^2\}\Bigg) + \\
& \sum_{j=1}^{J} Y_{ij}(1-E_z[Z_{ij}])\log I(Y_{ij}=1) + \log\phi(\boldsymbol{\theta}_i) + \sum_{j=1}^{J} E_r^{(t)}[\log p(Z'_{ij})]\Bigg] \\
= \; & \sum_{i=1}^{N}\sum_{j=1}^{J}(1-Y_{ij}+s_{ij}Y_{ij})\Bigg(\log\frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + (\frac{1}{2}-Y_{ij})(b_j - \boldsymbol{\alpha}_j^\top\mu_i^{(t)}) - \frac{1}{2}\xi_{i,j} \\
& -\eta(\xi_{i,j})\{b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top\mu_i^{(t)} + (\boldsymbol{\alpha}_j)^\top[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]\boldsymbol{\alpha}_j - \xi_{i,j}^2\}\Bigg) \\
& +\sum_{i=1}^{N}\sum_{j=1}^{J} Y_{ij}(1-s_{ij})\log I(Y_{ij}=1) + \frac{N}{2}\log|\Sigma_{\boldsymbol{\theta}}^{-1}| - \sum_{i=1}^{N}\frac{1}{2}Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]) \\
& +\sum_{i=1}^{N}\sum_{j=1}^{J}\{(1-Y_{ij}+s_{ij}Y_{ij})log(1-c_j) + Y_{ij}(1-s_{ij})log(c_j)\}, \qquad \text{(II.C.5)}
\end{aligned}$$

since $E_r[Z_{ij}] = s_{ij}$ for the Bernoulli distribution r_{ij}. The equation (II.C.5) is the closed form expression of the tth iteration's variational lower bound $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$.

M-Step As in 2PL case, in tth iteration's M step we maximize the $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ to update the model parameters $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$. This is again achieved by setting the derivative of $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ with respect to $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ to be zero.

For $\boldsymbol{\alpha}_j$, setting the derivative with respect to $\boldsymbol{\alpha}_j$ equal to zero, we have

$$\frac{\partial E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})}{\partial \boldsymbol{\alpha}_j} = \sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})\Big(Y_{ij}-\frac{1}{2}+2b_j\eta(\xi_{i,j})\Big)(\mu_i^{(t)})^\top \\ - 2\eta(\xi_{i,j})(1-Y_{ij}+s_{ij}Y_{ij})[\Sigma_i^{(t)}+(\mu_i^{(t)})(\mu_i^{(t)})^\top]\boldsymbol{\alpha}_j = 0.$$

This implies that $\boldsymbol{\alpha}_j^{(t+1)}$ is updated by

$$\boldsymbol{\alpha}_j = \frac{1}{2}\Big[\sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})\eta(\xi_{i,j})[\Sigma_i^{(t)}+(\mu_i^{(t)})(\mu_i^{(t)})^\top]\Big]^{-1}\times \\ \sum_{i=1}^{N}\Big[(1-Y_{ij}+s_{ij}Y_{ij})\Big(Y_{ij}-\frac{1}{2}+2b_j\eta(\xi_{i,j})\Big)(\mu_i^{(t)})^\top\Big]. \quad \text{(II.C.6)}$$

Similarly for b_j, the derivative of the variational lower bound with respect to b_j is

$$\frac{\partial E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})}{\partial b_j} = \sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})\Big[(\frac{1}{2}-Y_{ij})-2\eta(\xi_{i,j})b_j+2\eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top\mu_i^{(t)}\Big]=0$$

Setting it equal to 0, we obtain the updating equation for $b_j^{(t+1)}$ as

$$b_j = \frac{\sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})\Big[(\frac{1}{2}-Y_{ij})+2\eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top\mu_i^{(t)}\Big]}{\sum_{i=1}^{N}2(1-Y_{ij}+s_{ij}Y_{ij})\eta(\xi_{i,j})}. \quad \text{(II.C.7)}$$

Finally for a guessing parameter c_j, we again take derivate of $E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})$ with respect to c_j and set it equal to zero.

$$\frac{\partial E^{(t)}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{\xi})}{\partial c_j} \;=\; \sum_{i=1}^{N}[(1-Y_{ij}+s_{ij}Y_{ij})\frac{-1}{1-c_j}+Y_{ij}(1-s_{ij})\frac{1}{c_j}]=0.$$

This implies that $c_j^{(t+1)}$ is updated according to

$$c_j = \frac{\sum_{i=1}^{N}(Y_{ij}-s_{ij}Y_{ij})}{\sum_{i=1}^{N}(1-Y_{ij}+s_{ij}Y_{ij})+\sum_{i=1}^{N}(Y_{ij}-s_{ij}Y_{ij})} = \frac{\sum_{i=1}^{N}Y_{ij}(1-s_{ij})}{N}. \quad \text{(II.C.8)}$$

Following the same procedure, it is easy to check that $\boldsymbol{\xi}$ and $\Sigma_{\boldsymbol{\theta}}$ are updated with the same updating rule as in 2PL model (i.e. (II.11), (II.A.8)). Hence this completes the derivation of the M step for the 3PL model. In every M step, we iteratively update the tth iteration's estimate of the model parameters by (II.11), (II.A.8), (II.C.6), (II.C.7), and (II.C.8) until convergence, for $t \geq 1$ steps.

II.D Proof of Theorem II.3

In this section, we provide theoretical bounds on the estimation of the model parameters. We follow the proof of Theorem 1 in Davenport et al. (2014) and Theorem 1 in Chen et al. (2019). Define a matrix $M = [M_{ij}] = [\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i - b_j]$ and define $f(x)$ to be a logistic sigmoid function. For simplicity, we use the notation $\sum_{ij} = \sum_{i=1}^N \sum_{j=1}^J$ for the following proof. Then the variational lower bound to the marginal log-likelihood is as follows.

$$
\begin{aligned}
L_E(M) &= \sum_{i=1}^N \int_{\boldsymbol{\theta}_i} \log P(Y_i, \boldsymbol{\theta}_i | \boldsymbol{A}, \boldsymbol{B}) q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\
&= \sum_{i=1}^N \int_{\boldsymbol{\theta}_i} \Bigg[\sum_{j=1}^J Y_{ij} \log(\frac{\exp(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i - b_j)}{1 + \exp(a_j^T \boldsymbol{\theta}_i - b_j)}) \\
&\quad + (1 - Y_{ij}) \log(\frac{1}{1 + \exp(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i - b_j)}) + \log \phi(\boldsymbol{\theta}_i) \Bigg] q_i(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\
&= \sum_{i=1}^N \sum_{j=1}^J Y_{ij} E_{q_i} \Bigg[\log(\frac{\exp(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i - b_j)}{1 + \exp(a_j^T \boldsymbol{\theta}_i - b_j)}) \Bigg] \\
&\quad + (1 - Y_{ij}) E_{q_i} \Bigg[\log(\frac{1}{1 + \exp(\boldsymbol{\alpha}_j^T \boldsymbol{\theta}_i - b_j)}) \Bigg] + \sum_{i=1}^N E_{q_i} \Bigg[\log \phi(\boldsymbol{\theta}_i) \Bigg] \\
&= \sum_{ij} Y_{ij} E_{q_i} \Big[\log(f(M_{ij})) \Big] + (1 - Y_{ij}) E_{q_i} \Big[\log(1 - f(M_{ij})) \Big] + \sum_{i=1}^N E_{q_i} \Big[\log \phi(\boldsymbol{\theta}_i) \Big],
\end{aligned}
$$

where E_{q_i} denotes the expectation with respect to the distribution function $q_i(\boldsymbol{\theta}_i)$.

Define $\bar{L}_E(M) = L_E(M) - L_E(\mathbf{0})$ where $\mathbf{0}$ is a zero matrix with the same dimension as

M. Then,

$$\bar{L}_E(M) \quad = \quad \sum_{ij} Y_{ij} E_{q_i}\left[\log \frac{f(M_{ij})}{f(0)}\right] + (1 - Y_{ij}) E_{q_i}\left[\log \frac{1 - f(M_{ij})}{1 - f(0)}\right].$$

By the Mean Value Theorem of integrals, we can express $E_{q_i}[\log f(M_{ij})] = \log f(\bar{M}_{ij})$ and $E_{q_i}[\log(1 - f(M_{ij}))] = \log(1 - f(\tilde{M}_{ij}))$ for some $\bar{M}_{ij}$ and $\tilde{M}_{ij}$. Since we only observe either $Y_{ij} = 0$ or 1 for each data point, we then can rewrite $\bar{L}_E(M)$ as

$$\begin{aligned}
\bar{L}_E(M) \quad &= \quad \sum_{ij} Y_{ij}\left[\log \frac{f(\bar{M}_{ij})}{f(0)}\right] + (1 - Y_{ij})\left[\log \frac{1 - f(\tilde{M}_{ij})}{1 - f(0)}\right] \\
&= \quad \sum_{ij} I_{\{Y_{ij}=1\}}\left[\log \frac{f(\bar{M}_{ij})}{f(0)}\right] + \sum_{ij} I_{\{Y_{ij}=0\}}\left[\log \frac{1 - f(\tilde{M}_{ij})}{1 - f(0)}\right] \\
&=: \quad \bar{L}_{E_1}(\bar{M}) + \bar{L}_{E_0}(\tilde{M}). \qquad \text{(II.D.1)}
\end{aligned}$$

Define $G = \{M \in \mathbb{R}^{N \times J} : \|M\|_* \leq C\sqrt{KNJ}\} \subset \mathbb{R}^{N \times J}$ for $C \geq 0$, where $\|M\|_*$ is defined as a nuclear norm of a matrix M. Define

$$H = \{M = [M_{ij}]_{1 \leq i \leq N, 1 \leq j \leq J} : M_{ij} = \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j \text{ st } \|\boldsymbol{\theta}_i\|^2 \leq C \text{ and } \|\boldsymbol{\alpha}_j\|^2 \leq C \text{ for all } i, j\},$$

which is the set that satisfies the boundedness assumption $(A1)$. As shown in Chen et al. (2019), if $M \in H$ then $M \in G$ since

$$\|M\|_* \leq \sqrt{NJ}\sqrt{\text{rank}(M)}\|M\|_\infty \leq C\sqrt{KNJ}.$$

Note that $\text{rank}(M) = K$ in this proof as we assume that the number of latent traits K is fixed and known. For the following proof, we define $C_b = C_0 C\sqrt{K}\sqrt{NJ(N + J) + NJ\log(NJ)}$

with an absolute constant C_0 for simplicity. Hence we have

$$\begin{aligned}
&\mathbb{P}\bigg(\sup_{M\in H,\mu_i,\Sigma_i}\Big|\bar{L}_E(M)-E[\bar{L}_E(M)]\Big|\geq C_b\bigg)\\
\leq\ &\mathbb{P}\bigg(\sup_{\bar{M}\in H,\tilde{M}\in H}\Big|\bar{L}_{E_1}(\bar{M})-E[\bar{L}_{E_1}(\bar{M})]+\bar{L}_{E_0}(\tilde{M})-E[\bar{L}_{E_0}(\tilde{M})]\Big|\geq C_b\bigg)\\
\leq\ &\mathbb{P}\bigg(\sup_{M\in G}\Big|\bar{L}_E(M)-E[\bar{L}_E(M)]\Big|\geq C_b\bigg) \qquad \text{(II.D.2)}
\end{aligned}$$

where the first inequality follows from (II.D.1). The last expression in (II.D.2) satisfies conditions of Lemma A.1 of Davenport et al. (2014). Hence, we achieve the following result in lemma II.9 for GVEM.

Lemma II.9. *For absolute constants C_0 and C_1,*

$$\mathbb{P}\bigg(\sup_{M\in H,\mu_i,\Sigma_i}\Big|\bar{L}_E(M)-E[\bar{L}_E(M)]\Big|\geq C_0C\sqrt{K}\sqrt{NJ(N+J)+NJ\log(NJ)}\bigg)\leq\frac{C_1}{N+J}.$$

We can show with slight modification to page 210 of Davenport et al. (2014) that for any choice of M and $M^{'}\in H$,

$$\begin{aligned}
&E[\bar{L}_E(M^{'})]-E[\bar{L}_E(M)]\\
=\ &E[L_E(M^{'})-L_E(M)]\\
=\ &E\bigg[\sum_{ij}Y_{ij}E_{q_i}\bigg[\log\bigg(\frac{f(M^{'}_{ij})}{f(M_{ij})}\bigg)\bigg]+(1-Y_{ij})E_{q_i}\bigg[\log\bigg(\frac{1-f(M^{'}_{ij})}{1-f(M_{ij})}\bigg)\bigg]\bigg]\\
=\ &\sum_{ij}f(M_{ij})E_{q_i}\bigg[\log\bigg(\frac{f(M^{'}_{ij})}{f(M_{ij})}\bigg)\bigg]+(1-f(M_{ij}))E_{q_i}\bigg[\log\bigg(\frac{1-f(M^{'}_{ij})}{1-f(M_{ij})}\bigg)\bigg]\\
=\ &-NJE_q[D(f(M)||f(M^{'}))] \qquad \text{(II.D.3)}
\end{aligned}$$

where $D(P\|Q)=\frac{1}{NJ}\sum_{ij}KL(P_{ij}\|Q_{ij})$ for $P,Q\in[0,1]^{N\times J}$ is the KL divergence defined on the matrices of scalar inputs $P_{ij},Q_{ij}\in[0,1]$ for all i,j as defined in Davenport et al. (2014).

Now, define $\hat{L}_E(\hat{M})=\bar{L}_E|_{\hat{\mu_i},\hat{\boldsymbol{A}},\hat{\boldsymbol{B}},\hat{\Sigma_i}}$ which is estimated lower bound evaluated at the estimates from GVEM. It can be written as

$$\begin{aligned}
\hat{L}_E(\hat{M}) &= \sum_{ij} Y_{ij} E_{\hat{q}_i}\left[\log\left(\frac{f(\hat{M}_{ij})}{f(0)}\right)\right] + (1-Y_{ij})E_{\hat{q}_i}\left[\log\left(\frac{1-f(\hat{M}_{ij})}{1-f(0)}\right)\right] \\
&= \sum_{ij} Y_{ij} E_{\hat{q}_i}\left[\log\left(\frac{f(\hat{\boldsymbol{\alpha}}_i^\top \boldsymbol{\theta}_i - \hat{b}_j)}{f(0)}\right)\right] + (1-Y_{ij})E_{\hat{q}_i}\left[\log\left(\frac{1-f(\hat{\boldsymbol{\alpha}}_i^\top \boldsymbol{\theta}_i - \hat{b}_j)}{1-f(0)}\right)\right].
\end{aligned}$$

Then, we have

$$\begin{aligned}
& \hat{L}_E(\hat{M}) - \bar{L}_E(M) \\
= \ & E[\hat{L}_E(\hat{M}) - \bar{L}_E(M)] + \hat{L}_E(\hat{M}) - E[\hat{L}_E(\hat{M})] - \Big(\bar{L}_E(M) - E[\bar{L}_E(M)]\Big) \\
\leq \ & E[\hat{L}_E(\hat{M}) - \bar{L}_E(M)] + 2\sup_{M\in H,\mu_i,\Sigma_i}\Big|\bar{L}_E(M) - E[\bar{L}_E(M)]\Big| \\
= \ & -NJ \times E_{\hat{q}}[D(f(M)||f(\hat{M}))] + 2\sup_{M\in H,\mu_i,\Sigma_i}\Big|\bar{L}_E(M) - E[\bar{L}_E(M)]\Big| \qquad \text{(II.D.4)}
\end{aligned}$$

where (II.D.4) follows from (II.D.3). Since the estimates $\hat{M}$ from GVEM should satisfy $\hat{L}_E(\hat{M}) \geq \bar{L}_E(M^*)$ for the true parameter matrix $M^* \in H$,

$$-NJE_{\hat{q}}[D(f(M^*)||f(\hat{M}))] + 2\sup_{M\in H,\mu_i,\Sigma_i}\Big|\bar{L}_E(M) - E[\bar{L}_E(M)]\Big| \geq 0. \qquad \text{(II.D.5)}$$

By Lemma II.9, with probability $1 - \frac{C_1}{N+J}$

$$\sup_{M\in H,\mu_i,\Sigma_i}\Big|\bar{L}_E(M) - E[\bar{L}_E(M)]\Big| \leq C_0 C\sqrt{K}\sqrt{NJ(N+J) + NJ\log(N+J)}. \qquad \text{(II.D.6)}$$

Combining (II.D.5) and (II.D.6),

$$E_{\hat{q}}[D(f(M^*)||f(\hat{M}))] \leq \frac{2C_0 C\sqrt{K}}{\sqrt{NJ}}\sqrt{(N+J) + \log(N+J)}. \qquad \text{(II.D.7)}$$

Note that the KL divergence can be bounded below by the Hellinger distance; $d_H^2(p,q) \leq$

$D(p||q)$. Using this fact with Lemma A.2 from Davenport et al. (2014), we have for $C > 0$

$$||M^* - \hat{M}||_F^2 \leq \frac{8(1+e^C)^2}{e^C} NJ \times D(f(M)||f(\hat{M})) \leq 32e^C NJ \times D(f(M)||f(\hat{M})).\text{(II.D.8)}$$

From (II.D.7) and (II.D.8),

$$\frac{1}{NJ} E_{\hat{q}}[||M^* - \hat{M}||_F^2] \leq 64e^C C_0 C\sqrt{K}\sqrt{\frac{N+J}{NJ}}\sqrt{1+\frac{\log(N+J)}{N+J}}. \qquad \text{(II.D.9)}$$

This completes the proof of Theorem II.3.

ChapterIII

Regularized Variational Estimation for MIRT

III.1 Introduction

In large-scale educational and psychological tests, dichotomouse or polytomouse responses are often collected to investigate respondents' underlying latent abilities or traits. It is not uncommon that a single test is designed to examine multiple latent abilities at the same time. Multidimensional Item Response Theory (MIRT) is useful in this scenario as it models multiple latent abilities simultaneously to account for different mixtures of the abilities required by each test item. The MIRT models contain two or more parameters to describe the interaction between the latent traits and the responses to test items (Reckase, 2009). In this chapter, we focus on the logistic model with dichotomous responses but the proposed method can be adapted for other types of responses as well. In the multidimensional 2-Parameter Logistic (M2PL) model, the item response function of the ith individual to the jth item is modeled by

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

where there are N subjects who respond to J test items independently with binary response variables Y_{ij}, for $i = 1, \ldots, N$ and $j = 1, \ldots, J$. $\boldsymbol{\alpha}_j$ denotes a K-dimensional vector of item discrimination parameters for the jth item and b_j denotes the corresponding item difficulty parameter. $\boldsymbol{\theta}_i$ denotes the K-dimensional vector of latent ability for student i. For the multidimensional 3-Parameter Logistic (M3PL) model, there is an additional parameter c_j, which denotes the guessing probability of the jth test item. The item response function is expressed as

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

The maximum likelihood estimators of the model parameters are then obtained from maximizing the log-likelihood function. However, due to the latent variable structure in MIRT, maximizing the log-likelihood function involves a K dimensional integrals that are usually intractable where K is the dimension of latent factor. In the literature, direct numerical approximation to the integrals have been proposed, 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 is in skewed shape. Other numerical approximation methods based on Monte Carlo simulations have also been developed in the literature, such as the Monte Carlo expectation-maximization (McCulloch, 1997), stochastic expectation-maximization (von Davier & Sinharay, 2010), and Metropolis-Hastings Robbins-Monro algorithms (Cai, 2010b, 2010a). These methods usually depends on sampling data points from a posterior distribution and would be computationally involving. More recently, variational approximation to the integrals has been proposed, such as Gausssian Variational EM (GVEM) (Cho, Zhang, Wang, & Xu, 2020). GVEM adopts a variational lower bound of the intractable likelihood for the Expectation-Maximization procedure. This allows us to derive closed-form updates in the iterative EM steps, which makes the algorithm computationally efficient. Even when iterative parameter estimation under the GVEM framework get computationally intensive as both the number of subject N and number of test item J grows, we can stochastically optimize the variational estimation to reduce the computational burden. The advantage of having simple closed-form updates and stochastic optimization combined, the GVEM estimation can be computationally efficient in high dimensional MIRT models. Additionally, it was shown that GVEM works well in complex M3PL models compared to the existing methods. Hence variational approximated based method seems like a good alternative to the existing methods that uses direct numerical approximations to the integrals.

One of the primary goal in MIRT is to investigate a relationship between test items and multiple latent traits, a.k.a. a test structure or a factor loading structure. It shows a set of latent traits associated with each test item. Traditionally, identifying the factor loading structure proceeds in two steps. First, all item factor loadings are allowed to be freely estimated, and then a post-hoc rotation is conducted (Browne, 2001). While these rotation methods intend to produce a near-simple structure, an arbitrary cut-off for the rotated factor loadings is often needed to achieve a simple enough structure for interpretability. Instead, recent work has formulated the problem of estimating a test structure in MIRT as a latent variable selection problem (Sun et al., 2016). That is, for each test item, a set of latent traits influencing the distribution of the responses has been selected by the L_1-regularized regression. The basic idea is to penalize the factor loadings towards zero if the corresponding latent traits are not associated with a test item. This leads to correctly estimating an optimal non-zero factor loading structure, instead of setting subjective cut-offs. This approach also has the advantage over the model selection methods based on information criterion in terms of the computational cost as it simultaneously performs model estimation and selection of the sparse test structure. However, the computation is still quite challenging in MIRT model due to its intractable likelihood function. For parameter estimation, Sun et al. (2016) uses direct numerical approximation of the likelihood in the iterative EM procedure, which can be computationally inefficient especially in high dimensions. Specifically, they showed that the computation time for the latent variable selection is 30 minutes for the first penalization parameter λ and additional 10 minutes for the subsequent λs. Considering that multiple λs have to be used for the latent variable selection via regularization, it can take a few hours to estimate a test structure for a single dataset in high dimensions. Also, they could not tackle the challenging estimation problem in multidimensional 3-Parameter Logistic (M3PL) model. Hence, it is of our interest to develop computationally efficient estimation method for the item-trait structure that is flexible enough to work well in both M2PL and M3PL and in high dimensions.

In this chapter, we propose to apply the Gaussian Variational EM (GVEM) to facilitate the estimation of the item-trait relationship in MIRT. GVEM was proposed as a flexible and efficient estimation algorithm for the parameter estimation in MIRT models. It avoids direct calculation of the intractable likelihood function by approximating the variational lower bound to the likelihood. This leads to closed-form updates in the iterative EM algorithm, which greatly reduces the computational complexity. Moreover, the GVEM can be stochastically optimized to future improve the efficiency of the algorithm. We will apply these aspects of the GVEM to develop a flexible estimation algorithm of the item-trait relationship that extends well to more challenging scenarios involving complex M3PL model and high dimensional data. The performance of the proposed algorithm is thorougly studied with simulation studies.

The rest of the chapter is organized as follows. Section III.2 introduces a framework of the Gaussian Variational method in MIRT models. In Section III.3, we presents the general regularized variational algorithm. Section III.4 and section III.5 illustrate the performance of the proposed method with simulation studies and on real data, respectively. The chapter is concluded with Section III.6 to discuss future studies. Supplementary Material includes the detailed mathematical derivations of the estimation procedures presented in Section III.3

and additional results for the real data analysis.

III.2 Variational Estimation for MIRT

In this section, we will briefly illustrate the key idea of variational approximation discussed in Cho et al. (2020). We will provide general idea under M3PL model but it can be easily simplied to M2PL model. For conciseness, let us denote the model parameters for the MIRT models in Eqn (II.2) by $\boldsymbol{A} = \{\boldsymbol{\alpha}_j, j = 1, \ldots, J\}$, $\boldsymbol{B} = \{b_j, j = 1, \ldots, J\}$, and $\boldsymbol{C} = \{c_j, j = 1, \ldots, J\}$. Also, denote the responses $\boldsymbol{Y} = \{Y_i, i = 1, \ldots, N\}$ where $Y_i = \{Y_{ij}, j = 1, \ldots, J\}$ is the ith subject's response vector. Given the typical local independence assumption in IRT, the marginal log-likelihood of $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ in M3PL model given the responses $\mathbf{Y}$ is

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

where N is the total number of respondents and J is the total number of items in the test. Similarly this holds for M2PL model with model parameters $\boldsymbol{A}$ and $\boldsymbol{B}$. Here, ϕ denotes the K-dimensional Gaussian distribution of $\boldsymbol{\theta}$ with mean 0 and covariance $\Sigma_{\boldsymbol{\theta}}$. The maximum likelihood estimators of the model parameters are then obtained from maximizing the log-likelihood function, which is often intractable under MIRT. Hence we obtain variational approximation of (III.3) as follows. Further denote all model parameters as M_p to be general. Then the log-likelihood function $l(M_p; \mathbf{Y})$ can be equivalently rewritten as

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

for any arbitrary probability density function q_i satisfying $\int_{\boldsymbol{\theta}_i} q_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i = 1$. Since $P(Y_i \mid M_p) = P(Y_i, \boldsymbol{\theta}_i \mid M_p)/P(\boldsymbol{\theta}_i \mid Y_i, M_p)$, then we can further write

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

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

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

and the equality holds when $q_i(\boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i \mid Y_i, M_p)$ for $i = 1, \ldots, N$. With the optimal choice of the variational densities q_i's, one can derive a closed-form variational lowerbound as close as possible to the intractable log-likelihood. As a result, all parameter updates in EM procedure is done in closed-form, which makes the estimation computationally efficient.

III.3 Regularized Estimation of Test Structure

In this chapter, our main interest is to estimate a sparse test structure, denoted as $Q_{\mathbf{A}} = (q_{jk})$ where $q_{jk} = I(\boldsymbol{\alpha}_{jk} \neq 0)$. We follow the approach of formulating this as a latent variable selection problem in MIRT as presented in Sun et al. (2016). That is, we formulate the problem of estimating sparse test structure as a latent variable selection problem and solve it using the idea of regularized regression via L_1–type penalization. Our main contribution is to apply variational approach to avoid directly calculating intractable log-likelihood in MIRT models in solving the regularization problem.

Although Lasso regularization is a popular technique for simultaneous model estimation and efficient variable selection, there has been some arguments against the lasso oracle statement. For instance, Zou (2006) argued that there exist nontrivial conditions for the lasso variable selection to be consistent and thus Lasso rarely enjoys oracle properties. Although the computational efficiency of Lasso is appealing for the estimation problems in high-dimensional MIRT models, the bias of the Lasso may prevent consistent variable selection and model estimation. On the other hand, Adaptive Lasso is shown to enjoy oracle properties if the regularization parameters are chosen to be data-dependent (Zou, 2006). Since it is a convex optimization problem, its global optimizer can be efficiently solved. Additionally Adaptive Lasso is a simple extension of Lasso, which makes it easy to implement with the existing algorithm for the lasso and is computationally efficient as well. Hence, Adaptive Lasso is a good candidate as a penalization method for the latent variable selection problem in MIRT. In this chapter we focucs on regularized estimation via adaptive lasso penalization and compare its performance with Lasso. Simulation studies confirms that Adaptive Lasso is computationally efficient and also estimates model parameters more accurately than Lasso penalization under various conditions. See Section III.4 for more detail.

Specifically for parameter estimation, we solve the following optimization problem;

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\mathbf{C}}_\lambda) = argmax_{\mathbf{A},\mathbf{B},\mathbf{C}} l(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{Y}) - \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{w}_{jk} |\alpha_{jk}| \tag{III.5}$$

where $\hat{w}_{jk} = \frac{1}{|\hat{\alpha}_{jk}|^\gamma}$. In the adaptive lasso penalization, we use adaptive penalization weights for each parameter α_{jk}, instead of a constant penalization parameter, λ. The penalization weight for α_{jk} is $\lambda \hat{w}_{jk} = \frac{\lambda}{|\hat{\alpha}_{jk}|^\gamma}$ for some $\gamma > 0$. Thus, $\alpha_{jk} < 1$ will get penalized more than the bigger values such as $\alpha_{jk} > 1$. The weight is chosen to be dependent on data to satisfy the regulatory conditions discussed in Zou (2006).

To choose the constant sparsity parameter λ, we apply three different information criteria; AIC, BIC and GIC. We estimate the information criteria by substituting the log-likelihood with the variational lower bound from GVEM algorithm. The sparsity parameter that minimizes these information criteria will be chosen as optimal.

$$\begin{aligned}
AIC^{\star} &= max_{Q_A=Q^*,b} 2(||\mathbf{A}||_0 + ||\mathbf{B}||_0 + ||\mathbf{C}||_0) - 2E(\mathbf{A},\mathbf{B},\mathbf{C},\xi) \\
BIC^{\star} &= max_{Q_A=Q^*,b} ln(N)(||\mathbf{A}||_0 + ||\mathbf{B}||_0 + ||\mathbf{C}||_0) - 2E(\mathbf{A},\mathbf{B},\mathbf{C},\xi) \\
GIC^{\star} &= max_{Q_A=Q^*,b} ln(ln(N))(||\mathbf{A}||_0 + ||\mathbf{B}||_0 + ||\mathbf{C}||_0) - 2E(\mathbf{A},\mathbf{B},\mathbf{C},\xi)
\end{aligned}$$

where $||A||_0 = \sum_{j,k} I(\boldsymbol{\alpha}_{jk} \neq 0)$ denotes the L_0 norm of matrix A.

To ensure identifiability, we impose certain constraints on the a $K \times K$ sub-matrix of Q_A. For the remaining part of the A matrix, we don't assume any pre-specified zero structure and instead the appropriate penalization was imposed to shrink $\boldsymbol{\alpha}_{jk}$'s to recover the true zero structure, Q_A^*. The $\Sigma_{\boldsymbol{\theta}}$ is considered unknown throughout the estimation, thus it requires the estimation during GVEM steps. Below we describe two different constraints on A matrix which satisfy the identifiability conditions. Due to more flexible constraint in constraint 2, it is more challenging simulation setting than constraint 1. We will compare the performance between two constraint settings in the simulation study.

Constraint 1 To ensure identifiability, we designate one item for each latent factor and this item is associated with only that factor. That is, we set a $K \times K$ sub-matrix of Q_A to be identity matrix, I_K. The $\Sigma_{\boldsymbol{\theta}}$ is unknown and thus estimated.

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

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

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\mathbf{C}}_\lambda) = argmax_{\mathbf{A},\mathbf{B},\mathbf{C}} \; l(\mathbf{A}, \mathbf{B}, \mathbf{C}; \mathbf{Y}) - P_\lambda(\mathbf{A}) + P_\lambda(\mathbf{B}) + P_\lambda(\mathbf{C}) \tag{III.6}$$

where $P_\lambda(\mathbf{A}) = \lambda||A||_1 = \lambda \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{w}_{jk}|\alpha_{jk}|$, $P_\lambda(\mathbf{B}) = \sum_{j=1}^{J} \log N(b_j|\mu_b, \sigma_b^2)$, and $P_\lambda(\mathbf{C}) = \sum_{j=1}^{J} \log Beta(c_j|\alpha_c, \beta_c)$ for some distribution parameters μ_b,σ_b^2, α_c, and β_c. These penalty functions were chosen to satisfy the ranges of values on which the parameters are defined. For instance, since the guessing parameters $\mathbf{C}$ naturally satisfy the constraint $\{0 < c_j < 1; j = 1, \ldots, J\}$ as they are defined as probabilities of guessing, we can assume the distribution of $c_j \sim Beta(\alpha_c, \beta_c)$. The penalty on b_j and c_j are essentially a L_2-type and Laplace penalization, respectively. By imposing these additional penalties on model parameters $\mathbf{B}$ and $\mathbf{C}$, the number of variational parameters to estimate during iterative EM update greatly decreases, leading to better estimation results especially in high dimensional M3PL models.

The approach of imposing additional penalty on model parameters with the chosen distributions is similar to the Bayes Modal estimation presented by Tierney and Kadane (1986). That is, an augmented optimization objective is employed that includes the likelihood and some prior beliefs on the item parameters. These priors can be used to prevent deviant parameter estimates and help the algorithm to produce more accurate estimation in complex M3PL models. Essentially, Bayes Modal estimation can be seen as a regularization on maximum likelihood estimation where maximum likelihood estimation is a special case of Bayes Model estimation that assumes uniform prior distributions.

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

III.3.1 Computation via GVEM

To update A matrix, we use coordinate descent algorithm developed by Friedman et al. (2010). The general estimation procedure as follows; The regularization parameter is chosen as follows: For each λ we first obtain the estimate $(\hat{A}_\lambda, \hat{B}_\lambda, \hat{C}_\lambda)$ via (III.6). Then we obtain from a zero structure, $\hat{Q}_\lambda$ matrix from $\hat{A}_\lambda$. We fit a MIRT model with $\hat{Q}_\lambda$ as zero structure for A using GVEM algorithm without penalty. Hence, the final estimate $\hat{A}$ satisfies $Q(\hat{A}) = \hat{Q}_\lambda$. We compute the information criteria using estimates from a MIRT model without penalization. The regularization parameter is chosen to be the one admitting the smallest information criterion values.

For each item j, there are one difficulty parameter b_j and K discrimination parameters $\boldsymbol{\alpha}_j$. The coordinate descent algorithm update each of the $K+1$ variables iteratively according to the following updating rule. See appendix for detailed derivation of the updating rule. Define a function S to be a soft threshold operator such that $\mathrm{S}(\delta, \lambda) = sign(\delta)(|\delta| - \lambda)_+$

we update model parameters b_j and c_j following (III.7) and (III.8), respectively. For M3PL, we penalize $\boldsymbol{\alpha}_j$'s with adaptive lasso penalty as well. Then $\hat{\boldsymbol{\alpha}}_{jk}$ is updated according to the (III.9). See Appendix for the detailed derivation.

$$\begin{aligned} b_j^{(t+1)} &= \frac{\sum_{i=1}^N (1 - Y_{ij} + s_{ij} Y_{ij}) \left[\frac{1}{2} - Y_{ij} + 2\eta(\xi_{i,j}) \boldsymbol{\alpha}_j^\top \mu_i\right] + \frac{\mu_b}{\sigma_b^2}}{2 \sum_{i=1}^N (1 - Y_{ij} + s_{ij} Y_{ij}) \eta(\xi_{i,j}) + \frac{1}{\sigma_b^2}}, & \text{(III.7)} \\ c_j^{(t+1)} &= \frac{\sum_{i=1}^N Y_{ij}(1 - s_{ij}) + \alpha - 1}{N + \alpha + \beta - 2}. & \text{(III.8)} \end{aligned}$$

and

$$\boldsymbol{\alpha}_{jk}^{(t+1)} = \left[\sum_{i=1}^N (1 - Y_{ij} + s_{ij} Y_{ij}) \left(2\eta(\xi_{i,j})[\Sigma_i + (\mu_i)(\mu_i)^\top]_{k,k}\right)\right]^{-1} \times S\Bigg(\sum_{i=1}^N (1 - Y_{ij} + s_{ij} Y_{ij}) \Big\{$$
$$(Y_{ij} - \frac{1}{2})\mu_{i,k} + 2b_j \eta(\xi_{i,j})\mu_{i,k} - 2\eta(\xi_{i,j}) \sum_{l \neq k} \boldsymbol{\alpha}_{jl} [\Sigma_i + (\mu_i)(\mu_i)^\top]_{l,k} \Big\}, \lambda\Bigg) \qquad \text{(III.9)}$$

where λ is the sparsity parameter of choice.

The detailed algorithm of the regularized estimation of the test structure via Adaptive Lasso penalization is illustrated below in Algorithm 3.

Algorithm 3 Regularization with Adaptive Lasso Penalization

```
Set a range of $\lambda$. Choose $\gamma > 0$.
Initialize model parameters $A_0$, $B_0$, $\Sigma_0$ such that the identifiability condition holds.
Run confirmatory factor analysis (CFA) to obtain $\hat{A}_w := [\hat{\boldsymbol{\alpha}}_{jk}]_{J \times K}$
for each $\lambda$ starting from smallest do
    Set the adaptive penalization weights, $\frac{\lambda}{|\hat{A}_w(j,k)|^\gamma}$ for $j = 1, \dots, J$ and $k = 1, \dots, K$.
    Update $\boldsymbol{A}$ according to (III.9) using the adaptive weight as a sparsity parameter. Update $\boldsymbol{B}, \boldsymbol{C}$ according to (III.7), (III.8). Update $\Sigma_{\boldsymbol{\theta}}$ as in regular GVEM estimation.
    Estimate $A\hat{I}C^\star$, $B\hat{I}C^\star$, and $G\hat{I}C^\star$ with recent updates.
    Set $\hat{A}_\lambda$,$\hat{B}_\lambda$ as the initial values for next step.
end for
Find $\lambda^*$ that minimizes the information criteria. Calculate the correct estimation rate of $\hat{A}_{\lambda^*}$.
```

Remark III.1. *Here, we present the Algorithm 4 that summarizes the detailed regularized estimation procedure with Lasso regularization. It is to illustrate the slight difference between Lasso and Adaptive Lasso penalization approaches. After regularization, we re-estimate parameters in the Algorithm 4 step 7 to correct for the biased estimation of Lasso. Although this re-estimation might help correct the biasedness, it does not guarantee the consistent estimation after all. In addition, Lasso estimation requires more computation time due to this additional step. However, this difference in the estimation procedure is minor and we can easily check that the Adaptive Lasso is a simple extension of Lasso. Hence, it is easy to implement the adaptive lasso penalization using the existing numerical algorithms for Lasso.*

Algorithm 4 Regularization with Lasso Penalization

```
1: Set a range of $\lambda$.
2: Initialize model parameters $A_0$, $B_0$, $\Sigma_0$ such that the identifiability condition holds.
3: for each $\lambda$ starting from smallest do
4:     Update $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ according to (III.9), (III.7), (III.8). Update $\Sigma_{\boldsymbol{\theta}}$ as in regular GVEM
       estimation.
5:     Re-estimate $A$, $B$, and $\Sigma_{\boldsymbol{\theta}}$ according to confirmatory factor analysis with most recent
       updates (i.e. $\hat{A}_\lambda$, $\hat{B}_\lambda$) as initial values.
6:     With re-estimated $A$, $B$, and $\Sigma_{\boldsymbol{\theta}}$, estimate $A\hat{I}C^\star$, $B\hat{I}C^\star$, and $G\hat{I}C^\star$.
7:     Set $\hat{A}_\lambda$,$\hat{B}_\lambda$ as the initial values for next step.
8: end for
9: Find $\lambda^*$ that minimizes the information criteria. Calculate the correct estimation rate
   of $\hat{A}_{\lambda^*}$.
```

Remark III.2. *In addition to our choice of Adaptive Lasso for* $P_\lambda(\mathbf{A})$ *in Eqn* (III.6)*, there are generally other methods of penalizations. For instance, Fan and Li (2001) showed that the Lasso penalization problem is suboptimal to their proposed method called smoothly clipped absolute deviation (SCAD) penalty as Lasso produces biased estimates for the large coefficients. They showed that the SCAD penalization enjoys asymptotic normality and oracle properties with proper choice of regularization parameters. Due to its solid theoretical properties, SCAD has been widely applied in variable selection problems (T. Wang, Xu, & Zhu, 2012; Liu, Yao, & Li, 2016; Breheny & Huang, 2011). Additionally, Minimax Concave Penalty (MCP) has* *been presented as a fast, continuous and nearly unbiased method of penalization and hence claimed to be a good alternative to Lasso (C. H. Zhang, 2010). Truncated lasso is also another popular penalization method; however penalty function for these methods are non-convex and it makes local solutions to be nonunique in general. The nonconvex optimization problem is computationally challenging to solve as well. On the other hand, Adaptive Lasso is a convex problem and it is also computationally efficient, which makes it a good candidate for regularization problem under complex MIRT models. Hence, we chose Adaptive Lasso for solving our regularized problem.*

III.4 Simulation Study

III.4.1 Design

We present the simulation results for M2PL and M3PL models from 50 independent samples generated with number of subjects $N = 2000$, the dimension of test structure $K = 3$, and the test length $J = 45$. For the between-item multidimensional structure, there were 15 items loaded onto each factor with loading values set to $1, 1.5, 2$ respectively for each factor. For the within-item multidimensional structure, about 60% of the items are loaded onto one factor, about 25% are loaded onto two factors and the rest are loaded onto all three factors so that the data contains some single-, double- and triple-attribute test items. Their loading values are randomly drawn from $Unif(1, 2)$. For both models, the latent traits $\boldsymbol{\theta}$ are simulated from $MVN(0, \Sigma_{\boldsymbol{\theta}})$ with variance 1 and a common inter-factor correlation $\rho = 0.1$. The difficulty parameters b_j^* are set to zeros. Additionally in M3PL, we fixed true c^* to be 0.15 for all j's and c_0's were initialized from $Unif[0.05, 0.3]$. For prior-penalty, we used $Beta(\alpha, \beta) = Beta(2, 5)$ for both c_j and d_j so that the mode is around 0.2. For b_j's, we used $N(\mu_b, \sigma_b^2) = N(0, 1)$. The model parameters are estimated with tuning parameter λ chosen by the information criteria, which will be compared in terms of their estimation accuracy in the simulation.

As the main objective of this section is to estimate relationship between test items and latent traits, we use the correct estimation rate of A matrix (eq. (III.10)). It measures how well the sparsity of the A matrix is estimated by the regularized estimation. Notice that we only calculate correct rate for entries excluding the first K by K sub-matrix since we fix this part to have identity matrix as a zero structure to ensure identifiability.

$$CR = \frac{1}{K(J-K)} \sum_{K+1 \le j < J, 1 \le k \le K} I(\hat{Q}_{jk} = Q_{jk}^{true}) \tag{III.10}$$

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

III.4.2 Simulation Results

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

As shown in Figure III.1, Lasso penalization performs well in the Between-item model under both constraint settings. However, its performance under within-item model structure is much worse than that of the Adaptive Lasso as you can observe in Figure III.1 (b). Essentially, Adaptive Lasso shows close to 100% performance under all simulation settings in M2PL. We can observe from Figure III.2 that under M3PL both Lasso and Adaptive

Lasso penalization generally perform worse than in M2PL. However, overall Adaptive Lasso still performs well with mostly above 80% correct rates even in more complicated simulation conditions under more flexible identifiability constraint.

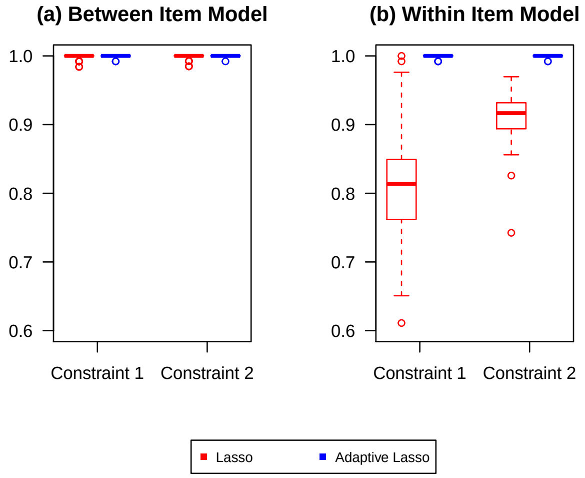


Figure III.1: Correct estimation rates under M2PL.

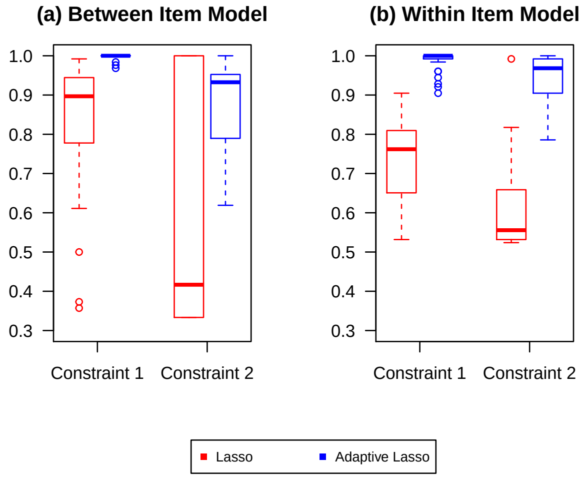


Figure III.2: Correct estimation rates under M3PL.

In addition to correct estimation rates, we compared the performance of Lasso and Adaptive lasso in terms of the sensitivity and specificity under various simulation settings. Figures III.3 and III.4 show the distributions of sensitivities and specificities for M2PL under Constraint 1 and Constraint 2, respectively. We can observe the Lasso performs as well as the

Adaptive Lasso in between item model. However its performance is much worse in more complex within item model. Both sensitivities and specificities are lower than those for Adaptive Lasso.

In Figures III.5 and III.6, we observe the distributions of sensitivities and specificities for M3PL under Constraint 1 and Constraint 2, respectively. Adaptive Lasso generally outperforms the Lasso penalization method under all model conditions. Considering that the parameter estimation is much more challenging in complex M3PL, Adaptive Lasso performs especially well for M3PL under Constraint 1. The specificities gets quite low for both Lasso and Adaptive Lasso under Constraint 2. However the Adaptive Lasso still performs quite well on average and much better than Lasso penalization.

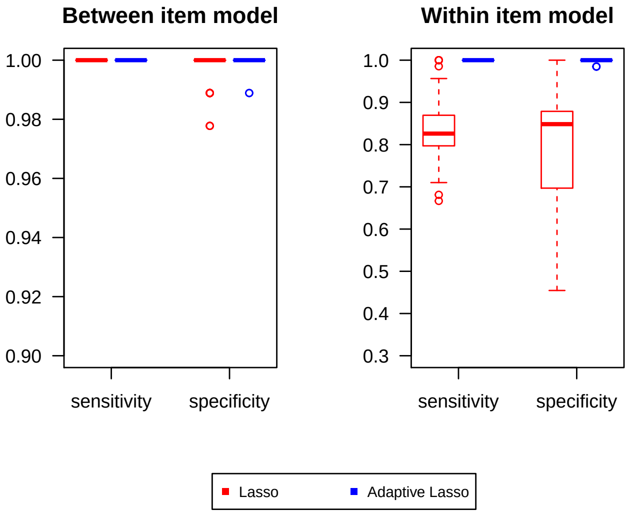


Figure III.3: Comparison of Lasso and Adaptive Lasso in M2PL under Constraint 1

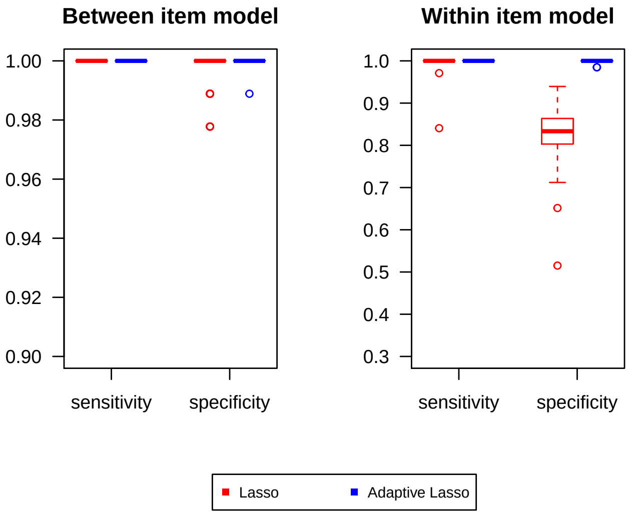


Figure III.4: Comparison of Lasso and Adaptive Lasso in M2PL under Constraint 2

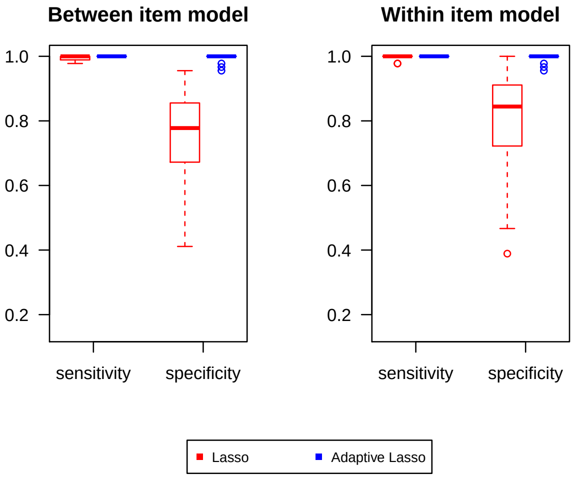


Figure III.5: Comparison of Lasso and Adaptive Lasso in M3PL under Constraint 1

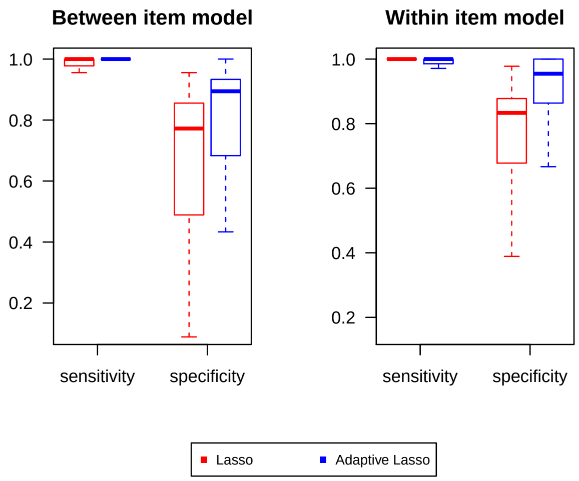


Figure III.6: Comparison of Lasso and Adaptive Lasso in M3PL under Constraint 2

III.5 Real Data Analysis

In this section, we apply our proposed regularization method on the National Education Longitudinal Study of 1988 (NELS:88) data. To revisit the data, a nationally representative sample of approximately 24,500 students were tracked via multidimensional cognitive batteries from 8th to 12th grade (the first three studies) in years 1988, 1990, and 1992. In this study, we focused on the science and mathematics test data where the multidimensional factorial structure has been previously investigated (e.g, Kupermintz & Snow, 1997; Nussbaum et al., 1997). Figure III.7 shows the example of the content of the questions in science test. For the science subject, there are 25 items and four factors emerged from the data collected in 1988: "Elementary science (ES)", "Chemistry knowledge (CK)", "Scientific reasoning (SR)" and "Reasoning with knowledge (RK)". For the math subject, there are 40 items in 1988 and two factors emerged. They are "Mathematical reasoning (MR)" and "Mathematical knowledge (MK)". We pooled together data from both domains, resulting in 65 items and a complete sample size of $N = 13,488$.

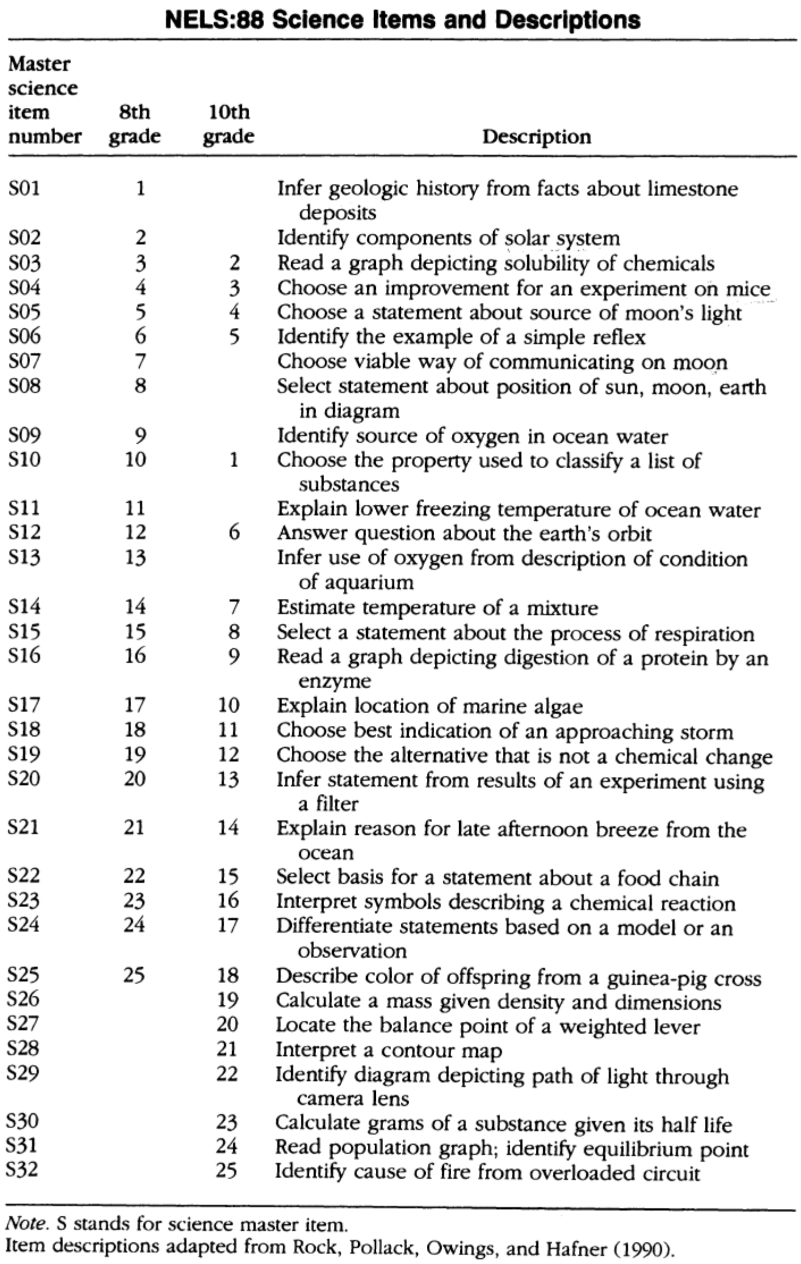
In the previous analysis, we used GVEM to empirically estimated the optimal number of latent traits from the data. The result suggested there exists six latent traits being tested by NELS:88. This finding also correspond to what the previous literature suggests as the number of latent factors being measured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

First, we compared the GIC values for different models as shown in Table III.1. For this analysis, we focused on adaptive lasso penalization as it was shown to perform better than lasso penalty under most model conditions in the simulation studies. Table III.1 shows that the M2PL with penalization has the lowest GIC value and thus is chosen as the model that best fit the NELS:88 test data. The fact that M2PL is chosen instead of M3PL as the optimal model implies that guessing does not play a big role on the performance in NELS:88 math and science assessment test. This result is again consistent with our previous analysis of NELS:88.

M2PL		M3PL	
no penalty	w/ penalty	no penalty	w/ penalty
8.46E5	8.31E5	1.28E06	1.81E06

Table III.1: GIC comparison with Adaptive Lasso penalty

As a second part of the analysis, we apply our proposed method of regularized estimation of test structure on NELS:88 data. Following the model selection result, we report the estimated sparse Q matrix for M2PL under constraint 1. The NELS:88 data is potentially a high dimensional measurement data with sample size of 13,488 and 65 test items. Hence we used stochastic optimization of the GVEM procedure to reduce the computational burden. Specifically we used a stochastic sampling of 200 at each iteration and initially sampled 3000 for more stable convergence. Table III.2 and Table III.3 illustrate the estimated sparse test



NELS:88 Science Items and Descriptions

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

Note. S stands for science master item.
Item descriptions adapted from Rock, Pollack, Owings, and Hafner (1990).

Figure III.7: Description of questions in science test of NELS:88

structure from math and science test, respectively. After penalization, total of 6 factors remained in the estimated test structure. We observe sparser structure between mathematrical abilities and science questions, and between science skills and math questions. This suggests that our regularized estimation procedure accurately shrunk the entries towards zero when the associated latent factors are not required to answer given test items. The sparse pattern is more apparant for the math test as shown in Table III.2. This could be because more test items are available in the math test to measure smaller number of latent factors (i.e. MR and MK). Although the estimated test structure in Table III.3 is less sparser overall, we do observe bigger coefficients for the entries associated with science skills.

The sparse test structure may be challenging to obtain due to the correlatedness of the latent traits. Although the NELS:88 test was designed so that each test item only requires subset of the six latent skills, the correlation between facors are quite high as shown in Table III.4. The correlations between science abilities are over 90%. Additionally some latent abilities are tested by only few test items according to the design of NELS:88. For example, "Reasoning with Knowledge(RK)" were intended to be measured by only $S15$ and $S22$ according to the NELS:88 test design. This makes it even more challenging to get accurate sparse test structure from the science test.

Factor	MR	MK	ES	SR	CK	RK
M1	0	1.1882	0	0	0	0
M2	0	1.1460	0	0	0	0
M3	1.3400	0	0	0	0	0
M4	0	1.4999	0	0	0	0
M5	0	1.5282	0	0	0	0
M6	1.2909	0	0	0	0	0
M7	1.2938	0	0	0	0	0
M8	0	1.0713	0	0	0	0
M9	1.4850	0	0	0	0	0
M10	0	1.3691	0	0	0	0
M11	1.0457	0	0	0	0	0
M12	0	1.5105	0	0	0	0
M13	1.3216	0	0	0	0	0
M14	0	1.1564	0	0	0	0
M15	0.7200	0	0.0902	0	0	0
M16	1.6220	0	0	0	0	0
M17	0	1.2101	0	0	0	0
M18	0.8592	0	0	0	0	0
M19	1.1810	0	0	0	0	0
M20	1.0592	0	0	0	0	0
M21	1.5735	0	0	0	0	0
M22	0.9947	0	0	0	0	0
M23	0.3162	0.4328	0.2156	0	0	0
M24	0.2515	0.4673	0.1739	0.0302	0.0310	0.0323
M25	1.4601	0	0	0	0	0
M26	1.3114	0	0	0	0	0
M27	0	0.5693	0	0	0	0
M28	0.4440	0.6379	0	0	0	0
M29	0	0.9575	0	0	0	0
M30	1.6277	0	0	0	0	0
M31	0	1.5418	0	0	0	0
M32	1.3457	0	0	0	0	0
M33	0.3465	0.2134	0	0.7503	0.7717	0.8065
M34	0	1.3468	0	0	0	0
M35	0	0.8882	0	0	0	0
M36	1.8644	0	0	0	0	0
M37	0	1.6834	0	0	0	0
M38	0	1.8049	0	0	0	0
M39	0	0.6994	0	0	0	0
M40	0.8942	1.2048	0	0	0	0

Table III.2: Estimated sparse test structure for math test in NELS:88

Factor	MR	MK	ES	SR	CK	RK
S1	0	0	0	0	1.0249	0
S2	0	0	0.8427	0	0	0
S3	0.3739	0	0.5413	0	0	0
S4	0	0.1888	0.5024	0.1793	0.1842	0.1924
S5	0	0	1.4329	0	0	0
S6	0	0	1.3383	0	0	0
S7	0	0	0.9671	0.0968	0.0997	0.1042
S8	0	0.1733	0.6884	0.2005	0.2061	0.2154
S9	0	0	0.8239	0	0	0
S10	0.1339	0.4530	0.4113	0.4632	0.4762	0.4974
S11	0	0	0	0	0	0
S12	0	1.1572	0	0	0	0
S13	0	0	0.9184	0	0	0
S14	0	1.6108	0	0	0	0
S15	0	0	0	0	0	0
S16	0.3988	0.1215	0.2026	0.7289	0.7486	0.7815
S17	0.2047	0.1176	0.5358	0.9919	1.0204	1.0665
S18	0.0748	0.3842	0.5234	0.8062	0.8293	0.8667
S19	0.0486	0.3973	0.4455	0.5574	0.5732	0.5988
S20	0.3532	0.0149	0.2353	1.0324	1.0611	1.1083
S21	0.0547	0.3287	0.2423	1.2428	1.2780	1.3354
S22	0.1667	0	0.4888	0.2285	0.2357	0.2467
S23	0.0076	0.3669	0	0.8709	0.8941	0.9331
S24	0.1065	0.2701	0.6743	0.2888	0.2971	0.3104
S25	0.1327	0.1718	0.2765	0.1455	0.1496	0.1563

Table III.3: Estimated sparse test structure for science test in NELS:88

	MR	MK	ES	SR	CK	RK
MR	1.0000	0.9815	0.8999	0.7150	0.8668	-0.8366
MK	0.9815	1.0000	0.9617	0.8214	0.9391	-0.9032
ES	0.8999	0.9617	1.0000	0.9212	0.9903	-0.9577
SR	0.7150	0.8214	0.9212	1.0000	0.9345	-0.8885
CK	0.8668	0.9391	0.9903	0.9345	1.0000	-0.9535
RK	-0.8366	-0.9032	-0.9577	-0.8885	-0.9535	1.0000

Table III.4: Estimated Correlation between latent factors

III.6 Discussions

In this chapter, a Gaussian variational regularization method has been proposed for the estimation of the sparse item-trait relationship in M2PL and M3PL models. It has the advantage over the model selection methods based on information criteria as it simultaneously performs model estimation and selection of the sparse test structure. Moreover, it has the computational advantage over the exiting methods based on direct numerical approximation (e.g. Sun et al., 2016) by avoiding the calculation of the intractable likelihood with variational approximation and deriving closed-form EM updates. Simulation studies demonstrated that the proposed methods performs well in correctly estimating the sparse item-trait structure in both M2PL and M3PL models with the adaptive lasso penalization.

A future step of this chapter would be to consider adding additional individual-level covariates to the model such as gender, race, and age, etc. These additional characteristics of individuals would help obtain in-depth understanding of the individuals' response pattern compared to the MIRT analysis only considering latent traits. The interesting application of this future study would be the recommender system. In the online recommendation system, individual-level information such as gender or age has significant impact on the individuals' preferences. Hence, incorporating these additional covariates in analyzing their pattern and understanding their latent preferences would greatly help the prediction for the personalized recommendations.

Appendix of Chapter III

In this section, we present the detailed derivations of the regularized variational estimation procedures. Appendix A and B illustrate the GVEM algorithm as presented in Algorithm 3 for M2PL and M3PL models, respectively.

III.A Derivation in M2PL

Item discrimination parameters a_{jk} are updated by the following steps. Let $Q_j(\boldsymbol{\alpha}_j, b_j)$ be the variational lower bound only concerning jth test item. Then,

$$\begin{aligned} Q_j(\boldsymbol{\alpha}_j, b_j) &= \sum_{i=1}^{N}\Bigg(\log\frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + (\frac{1}{2} - Y_{ij})b_j + (Y_{ij} - \frac{1}{2})\boldsymbol{\alpha}_j^\top \mu_i^{(t)} - \frac{1}{2}\xi_{i,j} \\ &\quad -\eta(\xi_{i,j})\{b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top \mu_i^{(t)} + \boldsymbol{\alpha}_j^\top[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]\boldsymbol{\alpha}_j - \xi_{i,j}^2\}\Bigg) \\ &\quad +\frac{N}{2}\log|\Sigma_{\boldsymbol{\theta}}^{-1}| - \sum_{i=1}^{N}\frac{1}{2}Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]). \end{aligned}$$

The first and second derivatives of the jth variational lower bound with respect to $\boldsymbol{\alpha}_{jk}$ are

$$\begin{aligned} \frac{\partial Q_j(\boldsymbol{\alpha}_j, b_j)}{\partial \boldsymbol{\alpha}_{jk}} &= \sum_{i=1}^{N}\Bigg((Y_{ij} - \frac{1}{2})\mu_{i,k}^{(t)} + 2b_j\eta(\xi_{i,j})\mu_{i,k}^{(t)} - \eta(\xi_{i,j})(2\boldsymbol{\alpha}_{jk}[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]_{k,k} \\ &\quad +2\sum_{l\neq k}\boldsymbol{\alpha}_{jl}[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]_{l,k})\Bigg), \\ \frac{\partial^2 Q_j(\boldsymbol{\alpha}_j, b_j)}{\partial \boldsymbol{\alpha}_{jk}^2} &= -\sum_{i=1}^{N}\Bigg(2\eta(\xi_{i,j})[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^\top]_{k,k}\Bigg), \end{aligned}$$

respectively. Define a function S to be a soft threshold operator such that $\mathrm{S}(\delta, \lambda) = sign(\delta)(|\delta| - \lambda)_+$ Then $\hat{\boldsymbol{\alpha}}_{jk}$ is updated according to the following rule;

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_{jk} &= -\frac{S(-\partial^2 Q_j(\boldsymbol{\alpha}_j, b_j) * \boldsymbol{\alpha}^*_{jk} + \partial Q_j(\boldsymbol{\alpha}_j, b_j), \lambda)}{\partial^2 Q_j(\boldsymbol{\alpha}_j, b_j)} \\
&= \frac{S\Bigg(\sum_{i=1}^N \Bigg[(Y_{ij} - \frac{1}{2})\mu_{i,k} + 2b_j \eta(\xi_{i,j})\mu_{i,k} - 2\eta(\xi_{i,j}) \sum_{l \neq k} \boldsymbol{\alpha}_{jl} [\Sigma_i + (\mu_i)(\mu_i)^\top]_{l,k} \Bigg], \lambda \Bigg)}{\sum_{i=1}^N \Bigg(2\eta(\xi_{i,j}) [\Sigma_i + (\mu_i)(\mu_i)^\top]_{k,k} \Bigg)}.
\end{aligned} \tag{III.A.1}$$

III.B Derivation in M3PL

The likelihood with prior-penalty on $\mathbf{B}$ and $\mathbf{C}$ can be written as

$$\begin{aligned}
& L(\mathbf{A}, \mathbf{B}, \mathbf{C}, \boldsymbol{\theta}) \times \prod_{j=1}^J N(b_j \mid \mu_b, \sigma_b) \times \prod_{j=1}^J Beta(c_j \mid \alpha, \beta) \\
= \; & L(\mathbf{A}, \mathbf{B}, \mathbf{C}, \boldsymbol{\theta}) \times \prod_{j=1}^J (2\pi\sigma_b^2)^{-1/2} \exp\Bigg[-\frac{1}{2\sigma_b^2}(b_j - \mu_b)^2 \Bigg] \times \prod_{j=1}^J \frac{c_j^{\alpha-1}(1-c_j)^{(\beta-1)}}{B(\alpha, \beta)}
\end{aligned}$$

Then, the variational lower bound to the expected log-likelihood is

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

To get the updating rules, we maximize the log-likelihood with respect to parameters b_j, c_j for $j = 1, \ldots, J$. First for b_j, we get the following by setting the derivative with respect to b_j equal to zero.

$$\frac{\partial Q_j(b_j)}{\partial b_j} = \sum_{i=1}^{N}(1 - Y_{ij} + s_{ij}Y_{ij})\left[(\frac{1}{2} - Y_{ij}) - \eta(\xi_{i,j})\{2b_j - 2\boldsymbol{\alpha}_j^\top \mu_i\}\right] - \frac{1}{\sigma_b^2}(b_j - \mu_b) = 0$$

We update $b_j^{(t+1)}$ according to the following equation (III.7).

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

Similarly for c_j, the derivative of the partial log-likelihood is

$$\frac{\partial Q_j(c_j)}{\partial c_j} = -\frac{1}{1 - c_j}\sum_{i=1}^{N}(1 - Y_{ij} + s_{ij}Y_{ij}) + \frac{1}{c_j}\sum_{j=1}^{N} Y_{ij}(1 - s_{ij}) + \frac{\alpha - 1}{c_j} - \frac{\beta - 1}{1 - c_j}$$

By rearranging the equation by c_j, we get a closed-form update of $c_j^{(t+1)}$ as follows.

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

For M3PL, we penalize $\boldsymbol{\alpha}_j$'s with adaptive lasso penalty as well.

$$\begin{aligned}
\frac{\partial Q_j(\boldsymbol{\alpha}_j, b_j)}{\partial \boldsymbol{\alpha}_{jk}} &= \sum_{i=1}^{N}(1 - Y_{ij} + s_{ij}Y_{ij})\Bigg((Y_{ij} - \frac{1}{2})\mu_{i,k} + 2b_j\eta(\xi_{i,j})\mu_{i,k} - \eta(\xi_{i,j})\Big\{2\boldsymbol{\alpha}_{jk}[\Sigma_i + \mu_i\mu_i^\top]_{k,k} \\
&\quad +2\sum_{l \neq k}\boldsymbol{\alpha}_{jl}[\Sigma_i + \mu_i\mu_i^\top]_{l,k}\Big\}\Bigg), \\
\frac{\partial^2 Q_j(\boldsymbol{\alpha}_j, b_j)}{\partial \boldsymbol{\alpha}_{jk}^2} &= -\sum_{i=1}^{N}(1 - Y_{ij} + s_{ij}Y_{ij}) \times 2\eta(\xi_{i,j})[\Sigma_i + \mu_i\mu_i)^\top]_{k,k},
\end{aligned}$$

respectively. Then $\hat{\boldsymbol{\alpha}}_{jk}$ is updated according to the following rule;

$$\begin{aligned}
&\boldsymbol{\alpha}_{jk}^{(t+1)} \\
= \quad & -\frac{S(-\partial^2 Q_j(\boldsymbol{\alpha}_j, b_j) \times \boldsymbol{\alpha}_{jk}^{(t)} + \partial Q_j(\boldsymbol{\alpha}_j, b_j), \lambda)}{\partial^2 Q_j(\boldsymbol{\alpha}_j, b_j)} \\
= \quad & \frac{S\Bigg(\sum_{i=1}^N (1 - Y_{ij} + s_{ij}Y_{ij}) \Big\{ (Y_{ij} - \frac{1}{2})\mu_{i,k} + 2b_j \eta(\xi_{i,j})\mu_{i,k} - 2\eta(\xi_{i,j}) \sum_{l \neq k} \boldsymbol{\alpha}_{jl} [\Sigma_i + (\mu_i)(\mu_i)^\top]_{l,k} \Big\}, \lambda \Bigg)}{\sum_{i=1}^N (1 - Y_{ij} + s_{ij}Y_{ij}) \Bigg(2\eta(\xi_{i,j}) [\Sigma_i + (\mu_i)(\mu_i)^\top]_{k,k} \Bigg)}.
\end{aligned} \tag{III.B.3}$$

where λ is the sparsity parameter of choice.

ChapterIV

Extensions of Gaussian Variational EM

IV.1 Introduction

In this chapter, we discuss some of the interesting extensions of the proposed Gaussian Variational estimation approach. The first extension we study is the parameter estimation for the Multidimensional 4-Parameter Logistic (M4PL) model, an extension of the M2PL and M3PL models discussed previously. M4PL model can be simplified to the unidimensional 4-Parameter Logistic (4PL) model where we only model one latent ability at a time. The 4PL model has not been widely applied for a long time probably due to its challenges in parameter estimation and a lack of evidence supporting the need for such a complex model (Loken & Rulison, 2010). However, recent literature have expressed renewed interest in the 4PL model. For example, Liao, Ho, Yen, and Cheng (2012) and Rulison and Loken (2009) demonstrated that the 4PL can improve the accuracy of the assessments in computerized adaptive testing by incorporating the individuals' careless mistakes in the early stages. In addition, Reise and Waller (2003) and Waller and Reise (2010) argued that the 4PL model may be more appropriate for measuring psychopathology traits than the simpler 2PL or 3PL models since it is very common in psychopathology measurement for a highly able subject (i.e. subject with enough latent abilities) to be reluctant to self-report his or her attitudes. These findings support the need to include an upper asymptote parameter in modeling the probability of responses and develop estimation methods for the 4PL models. Furthermore, it would be necessary to study the parameter estimation for more general multidimensional 4PL models which can handle more complex educational and psychological data by modeling multiple latent traits at the same time.

Several parameter estimation methods for the 4PL model have been proposed in the previous literature. For instance, Loken and Rulison (2010) adopted a Bayesian approach with the Markov chain Monte Carlo (MCMC) sampler for the parameter estimation. Feuerstahler and Waller (2014) employed the marginal maximum likelihood method, which require less computation time than the Bayesian approach with the MCMC sampler. However, marginal maximum likelihood approach may not be stable and may produce deviant values in many cases (Baker & Kim, 2004). To solve this problem, Waller and Feuerstahler (2017) recently applied Bayes Model estimation for the 4PL model. That is, an augmented optimization objective was used that includes the likelihood and some additional prior beliefs on the item parameters to prevent deviant parameter estimates. However, these methods either involve MCMC sampling or computation of complete likelihood and thus would be computationally time-consuming for high dimensional data from the large scale assessment tests. In addition, they would be especially challenging for the parameter estimation in M4PL models that involve intractable multidimensional integrals in the calculation of likelihood due to its multidimensional latent varable structure. To tackle these challenges, we propose to apply Gaussian Variational EM approach to develop a computationally efficient and accurate EM algorithms for the parameter estimation in M4PL.

As our second extension, we demonstrate how the GVEM method can be applied to Differential Item Functioning (DIF) analysis for MIRT. In short, DIF occurs when groups (e.g. defined by gender or race) have different probabilities of responses for a given test item even when individuals share the same level of latent abilities. An item is labeled as having DIF when people with the same level of latent abilities but from different groups have an unequal probability of responses. An item is labeled as non-DIF when people with the same latent abilities have equal probability of getting a test item correct, regardless of group membership. If DIF occurs in a test, that means subgroups of the examinee population differ on the dimensions that are other than the goal dimensions (i.e. a set of latent abilities or traits intended to be measured by the test). Hence, the reported results in such a test may include test and item bias (Reckase, 2009). MIRT analyses can help identify any group differences that contribute to test bias and a clear representation of DIF using MIRT models has been previously discussed (Ackerman, 1992).

Naturally it is desirable for the assessment test to not have differential item functioning in any of its items. Hence, identifying the items that have DIF is crucial in evaluating and improving the test design. On the other perspective, studying the DIF would help practitioners uncover deeper understanding of individuals by studying whether certain subgroups have different responses or characteristics. Hence, it is of interest to study DIF for many practitioners in the field of education, psychology, epidemiology and medicine (e.g. Breslau, Javaras, Blacker, Murphy, & Normand, 2008; Crane et al., 2007; Kwakkenbos et al., 2014; Lewis, Yang, Jacobs, & Fitchett, 2012; Uebelacker, Strong, Weinstock, & Miller, 2009). We formulate the DIF analysis to assess group differences in a test as a regularization problem in MIRT and develop the estimation methods using the proposed GVEM approach. This could be considered as the extension of the regularized variational estimation studied in Chapter III. Our discussion of DIF analysis in this chapter is limited to the M2PL model; however it can be naturally extended to more complex MIRT models. We leave this for the future study.

The rest of this chapter is organized as follows. In Section IV.2, we introduce the M4PL model formulation in detail and present the mixture modeling approach for the M4PL. Section IV.2.1 presents the derivation of the GVEM algorithm for parameter estimation in

M4PL under the mixture modeling framework. Section IV.2.2 shows simulation studies conducted to evaluate the performance of the proposed method under various model conditions. Section IV.3 then discusses the Differential Item Functioning for MIRT as our second extension focusing on the binary group case such as defined by gender. Section IV.3.1 present the proposed estimation methods via GVEM approach. Simulation results in IV.3.2 shows that our proposed GVEM approach performs well for the DIF analysis in M2PL model in various model conditions.

IV.2 Multidimensional 4PL model

The unidimensional 4-Parameter Logistic model (4PL) was first proposed in Barton and Lord (1981), who introduced an upper asymptote parameter, d, to incorporate the scenario where a high-ability individual misses an easy test item. Essentially, $1 - d$ can be considered as the probability of making a mistake. The limitations of Barton and Lord's model are that all test items share a common upper asymptote parameter value and also the model was estimated with fixed d values. Recent studies (e.g. Linacre, 2004; Rouse, Finger, & Butcher, 1999; Rupp, 2003; Tavares, Andrade, & Pereira, 2004; Waller & Reise, 2010) have demonstrated that the upper asymptote d varies across test items in most cases. Hence, the 4PL formulation that allows the upper asymptote parameter to be item-specific (i.e. d_j for $j = 1, \ldots, J$) is considered more appropriate. The probability of correctly answering test items in 4PL model is formulated as follow;

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i) = c_j + (d_j - c_j)\frac{\exp(\alpha_j\theta_i - b_j)}{1 + \exp(\alpha_j\theta_i - b_j)} \tag{IV.1}$$

where Y_{ij} denotes the observed dichotomous response of examinee i to test item j for $i = 1, \ldots, N$ and $j = 1, \ldots, J$). Again, $Y_{ij} = 1$ denotes a correct response and $Y_{ij} = 0$ otherwise. As previously discussed, α_j denotes the item discrimination parameter and b_j denotes item difficulty parameter for the jth test time. The c_j denotes the guessing parameter and d_j

denotes the upper asympotote parameter. The parameter d_j is the maximum probability of response for the jth item. Hence, $1 - d_j$ can be considered as the slipping probability of a student who's able to correctly answer but missing the item by mistake. N and J are used to denote the number of the examinees and the test length.

The unidimensional 4PL model in (IV.1) can be extended to multidimensional 4PL (M4PL) model by substituting a single latent ability variable, θ_i, to a K-dim vector of multiple latent traits, $\boldsymbol{\theta}_i$, as follows;

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

where K is the dimension of latent traits. Due to the K dimensional latent structure in M4PL, the complete log-likelihood involves calculation of K-dim integrals, which is intractable as in M2PL and M3PL. The Bayesian approach of using MCMC sampler or marginal maximum likelihood approach could be considered as it has been discussed as the estimation methods for the unidimensional 4PL model. However, this would be computationally time-consuming when the dimension K is high and data is of high dimensions. Also, the marginal maximum likelihood estimates could be biased as shown in the 4PL model (Baker & Kim, 2004). Hence, we would like to apply our Gaussian Variational approach to approximate the intractable likelihood and to facilitate the parameter estimation in EM procedures.

Mixture modeling approaches have been recently developed for MIRT by introducing additional latent variables. For instance, Béguin and Glas (2001), von Davier (2009), and Martín, Del Pino, and De Boeck (2006) interpreted the 3PL model from the perspective of a two-response strategy (with and without guessing) and formulated it as a mixture model. Culpepper (2016, 2017) further developed a mixture modeling approach to reformulate the four-parameter normal ogive model (4PNOM) and multidimensional 4PNOM. In addition, Meng, Xu, Zhang, and Tao (2019) adopted the mixture modeling approach for unidimensional 4PL model to develop EM based estimation approach. Following these mixture framework, we present an alternative expression for the M4PL using a mixture model. Specifically, we introduce an additional latent variable, Z_{ij} to characterize the two response processes. Define Z_{ij} as a binary latent variable denoting the ith individual's capability on jth test item. That is, $Z_{ij} = 1$ if ith subject is capable of correctly answering jth test item based on his or her latent ability $\boldsymbol{\theta}_i$ and item specific parameters (i.e. $\boldsymbol{\alpha}_j$ and b_j). Naturally, $Z_{ij} = 0$ if he or she does not have enough latent skills required by jth test item to answer it correctly. The mixture modeling of M4PL is as follows;

$$\begin{aligned} Y_{ij} \mid Z_{ij} &\sim Bernoulli(d_j^{Z_{ij}} c_j^{1-Z_{ij}}) \\ Z_{ij} \mid \boldsymbol{\theta}_i &\sim Bernoulli(\frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1+\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}) \end{aligned} \tag{IV.3}$$

It can be easily shown that the marginal probability of responses Y_{ij} under the mixture model in (IV.3) is equivalent to that of M4PL given in (IV.2).

The mixture model framework for 4PL models offers new insight into the connection between 4PL MIRT models and the Cognitive Diagnosis Models (CDMs) (Meng et al., 2019). The Z_{ij} can be interpreted as the attribute profile in the CDM literature, often denoted as $\boldsymbol{\alpha}$. That is, $Z_{ij} = 1$ indicates that the ith examinee is capable of answering jth item and $Z_{ij} = 0$ otherwise. Then the distribution of responses Y_{ij} is the same as the deterministic input, noisy AND gate (DINA) model specification, where c_j corresponds to the guessing parameter and $1 - d_j$ corresponds to the slipping parameter. Meng et al. (2019) argues that 4PL can also be viewed as a generalization of the higher-order DINA model (De La Torre & Douglas, 2004) with only one latent attribute and more generally, multi-attribute higher-order DINA model may be considered as a sub-model of the M4PL. This emphasizes the importance of developing estimation methods for M4PL and study the connection between M4PL and CDM framework.

IV.2.1 GVEM for M4PL

In this section, we present the Gaussian variational EM algorithm for M4PL model with its mixture modeling specification. By Bayes Theorem, we have $P(Y_{ij}, Z_{ij}, \boldsymbol{\theta}_i \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) = P(Y_{ij} \mid Z_{ij}, \boldsymbol{\theta}_i, \mathbf{C}, \mathbf{D})P(Z_{ij} \mid \boldsymbol{\theta}_i, \mathbf{A}, \mathbf{B})P(\boldsymbol{\theta}_i)$. Then, the complete data log-likelihood for ith subject only can be written as

$$\begin{aligned}
& \log P(\mathbf{Y}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}_i \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \\
= & \sum_{j=1}^{J} \log P(Y_{ij}, Z_{ij}, \boldsymbol{\theta}_i \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \\
= & \sum_{j=1}^{J} \log P(Y_{ij} \mid Z_{ij}, \boldsymbol{\theta}_i, \mathbf{C}, \mathbf{D}) + \sum_{j=1}^{J} \log P(Z_{ij} \mid \boldsymbol{\theta}_i, \mathbf{A}, \mathbf{B}) + \log P(\boldsymbol{\theta}_i) \\
= & \sum_{j=1}^{J} \log(d_j^{Z_{ij}} c_j^{1-Z_{ij}})^{Y_{ij}}(1 - d_j^{Z_{ij}} c_j^{1-Z_{ij}})^{1-Y_{ij}} + \log P(\boldsymbol{\theta}_i) \\
& + \sum_{j=1}^{J} \log \Big(\frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} \Big)^{Z_{ij}} \Big(\frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} \Big)^{1-Z_{ij}} \\
= & \sum_{j=1}^{J} \Bigg[Y_{ij}\{Z_{ij} \log d_j + (1 - Z_{ij}) \log c_j\} + (1 - Y_{ij}) \log(1 - d_j^{Z_{ij}} c_j^{1-Z_{ij}}) \Bigg] + \log P(\boldsymbol{\theta}_i) \\
& + \sum_{j=1}^{J} \Bigg[Z_{ij}(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + \log \frac{1}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j)} \Bigg] \qquad \text{(IV.4)}
\end{aligned}$$

By using local variational method on log-sigmoid function in Eqn (IV.4) as discussed in previous chapters, we have

$$\begin{aligned}
& \log P(\mathbf{Y}_i, \boldsymbol{Z}_i, \boldsymbol{\theta}_i \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \\
\geq & \sum_{j=1}^{J} \Bigg[Y_{ij}\{Z_{ij} \log d_j + (1 - Z_{ij}) \log c_j\} + (1 - Y_{ij}) \log(1 - d_j^{Z_{ij}} c_j^{1-Z_{ij}}) \Bigg] + \log P(\boldsymbol{\theta}_i) \\
& + \sum_{j=1}^{J} Z_{ij}(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j) + \sum_{j=1}^{J} \log \frac{e^{\xi_{i,j}}}{1 + e^{\xi_{ij}}} + \sum_{j=1}^{J} \frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \xi_{ij}) \\
& - \sum_{j=1}^{J} \eta(\xi_{i,j})\{(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \xi_{i,j}^2\} \\
:= & \; l(\mathbf{Y}_i, \boldsymbol{\theta}_i, \mathbf{Z}_i, \boldsymbol{\xi}_i \mid \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) \qquad \text{(IV.5)}
\end{aligned}$$

Now, we would like to find the optimal variational densities for the latent variables $\boldsymbol{\theta}_i$ and $\boldsymbol{Z}_i$. Let us denote these variational densities as $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)$ and $q_z(Z_{ij})$ respectively where $\boldsymbol{Z}_i = \{Z_{i1}, Z_{i2}, \ldots, Z_{iJ}\}$. Firstly, in order to find the optimal $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)$ we take expectation of the lower bound to the log-likelihood (i.e. Eqn (IV.5)) with respect to Z_{ij}'s. Then we can show that $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)$ naturally follows the following form;

$$\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \propto \sum_{j=1}^{J}(E_z[Z_{ij}] - \frac{1}{2})\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \sum_{j=1}^{J} \eta(\xi_{i,j})(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i)^2 - \frac{1}{2}\boldsymbol{\theta}_i^\top \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\theta}_i$$

Then, the optimal choice of the variational density for $\boldsymbol{\theta}_i$ is $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \mu_i, \Sigma_i)$ with

$$\begin{aligned} \mu_i &= \Sigma_i \times \sum_{j=1}^{J}\{2\eta(\xi_{i,j})b_j + E_z[Z_{ij}] - \frac{1}{2}\boldsymbol{\alpha}_j^\top\} && \text{(IV.6)} \\ \Sigma_i &= \Sigma_{\boldsymbol{\theta}}^{-1} + 2\sum_{j=1}^{J} \eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top \boldsymbol{\alpha}_j. && \text{(IV.7)} \end{aligned}$$

for $i = 1, \ldots, N$.

Similarly, we take the expectation of the lower bound, Eqn (IV.5), with respect to $\boldsymbol{\theta}_i$'s in order to find the optimal choice of the variational distributions, $q_z(Z_{ij})$. We can easily show that the posterior distribution of Z_{ij} has the following form;

$$\begin{aligned} \log q_z(Z_{ij}) &\propto Y_{ij} Z_{ij} \log(d_j) + Y_{ij}(1 - Z_{ij}) \log(c_j) + (1 - Y_{ij}) \log(1 - d_j^{Z_{ij}} c_j^{1-Z_{ij}}) \\ &\quad + Z_{ij}(\boldsymbol{\alpha}_j^\top E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i] - b_j) \\ &= Z_{ij}\{Y_{ij} \log(d_j) + (1 - Y_{ij}) \log(1 - d_j) + \boldsymbol{\alpha}_j^\top E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i] - b_j\} \\ &\quad + (1 - Z_{ij})\{Y_{ij} \log(c_j) + (1 - Y_{ij}) \log(1 - c_j)\} \end{aligned}$$

This implies that the optimal choice of the variational density for Z_{ij} is $q_z(Z_{ij}) \sim Bernoulli(s_{ij})$ where

$$s_{ij} = \frac{1}{1 + \exp\left[Y_{ij} \log(\frac{c_j}{d_j}) + (1 - Y_{ij}) \log(\frac{1-c_j}{1-d_j}) - \boldsymbol{\alpha}_j^\top E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i] + b_j\right]} \qquad \text{(IV.8)}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, J$.

E step With the optimally chosen variational densities $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)$ and $q_z(Z_{ij})$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$, we can derive a closed form expression of the variational lower bound to the marginal log-likelihood, denoted by $E[\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \boldsymbol{\xi}]$, as follows in the E step;

$$\begin{aligned}
& E[\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \boldsymbol{\xi}] \\
:= \; & \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \left[\sum_{j=1}^{J} \sum_{Z_{ij}} l(\mathbf{Y}_i, \boldsymbol{\theta}_i, \boldsymbol{Z}_i, \boldsymbol{\xi}_i | \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}) q_z(Z_{ij}) \right] q_\theta(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \\
= \; & \sum_{i=1}^{N} \sum_{j=1}^{J} \left[Y_{ij}\{s_{ij} \log d_j + (1 - s_{ij}) \log c_j\} + (1 - Y_{ij})\{s_{ij} \log(1 - d_j) + (1 - s_{ij}) \log(1 - c_j)\} \right] \\
& - \sum_{i=1}^{N} \frac{1}{2} Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i + \mu_i \mu_i^\top]) + \frac{N}{2} \log |\Sigma_{\boldsymbol{\theta}}^{-1}| + \sum_{i=1}^{N} \sum_{j=1}^{J} \left[s_{ij}(\boldsymbol{\alpha}_j^\top \mu_i - b_j) + \log \frac{e^{\xi_{i,j}}}{1 + e^{\xi_{ij}}} \right. \\
& \left. + \frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top \mu_i - \xi_{ij}) - \eta(\xi_{i,j})\{b_j^2 - 2b_j \boldsymbol{\alpha}_j^\top \mu_i + \boldsymbol{\alpha}_j^\top [\Sigma_i + \mu_i \mu_i^\top] \boldsymbol{\alpha}_j - \xi_{i,j}^2\} \right] \qquad \text{(IV.9)}
\end{aligned}$$

M step To get the updating rules for model parameters, we simply take derivatives of variational lower bound in Eqn. (IV.9) and set them equal to zero. By repeating this procedure for each model parameter, we get the updating rules as in Eqn. (IV.10), (IV.11), (IV.12), and (IV.13) for $\boldsymbol{\alpha}_j$, b_j, c_j and d_j repectively. $\Sigma_{\boldsymbol{\theta}}$ and $\xi_{i,j}$ have the same updating rule as in M3PL, which can be referred to Chapter II.4.1.

$$\boldsymbol{\alpha}_j = \Big[\sum_{i=1}^{N} 2\eta(\xi_{i,j})[\Sigma_i + \mu_i \mu_i^\top] \Big]^{-1} \times \sum_{i=1}^{N} \Big[s_{ij} - \frac{1}{2} + 2b_j \eta(\xi_{ij}) \Big] \mu_i^\top \qquad \text{(IV.10)}$$

$$b_j = \frac{\sum_{i=1}^{N} [2\eta(\xi_{i,j}) \boldsymbol{\alpha}_j^\top \mu_i - s_{ij} + \frac{1}{2}]}{\sum_{i=1}^{N} 2\eta(\xi_{i,j})} \qquad \text{(IV.11)}$$

$$c_j = \frac{\sum_{i=1}^{N} Y_{ij}(1 - s_{ij}) + \alpha - 1}{\sum_{i=1}^{N}(Y_{ij} - s_{ij} Y_{ij}) + \sum_{i=1}^{N}(1 - Y_{ij})(1 - s_{ij})} = \frac{\sum_{i=1}^{N} Y_{ij}(1 - s_{ij})}{\sum_{i=1}^{N}(1 - s_{ij})} \qquad \text{(IV.12)}$$

$$d_j = \frac{\sum_{i=1}^{N} Y_{ij} s_{ij}}{\sum_{i=1} s_{ij}} \tag{IV.13}$$

From the above updating rules, we observe that iterative EM updates for the parameter estimation are all in closed-form, which contributes to the efficient computation for the complex M4PL model.

IV.2.2 Simulation studies : M4PL

A series of simulation studies were conducted to evaluate the performance of the proposed GVEM algorithm for M4PL models. The number of test dimensions was fixed at 3 and test length was fixed at 45. In all cases, item discrimination parameters $\boldsymbol{\alpha}_j$ were simulated from $Unif(1, 2)$ distribution, and difficulty parameter b_j was simulated from the standard normal distribution. The probability of guessing and slipping were fixed at 0.05 for all test items. That is, there is 5% chance of guessing the test items correctly and 5% chance of making a mistake even with enough latent abilities.

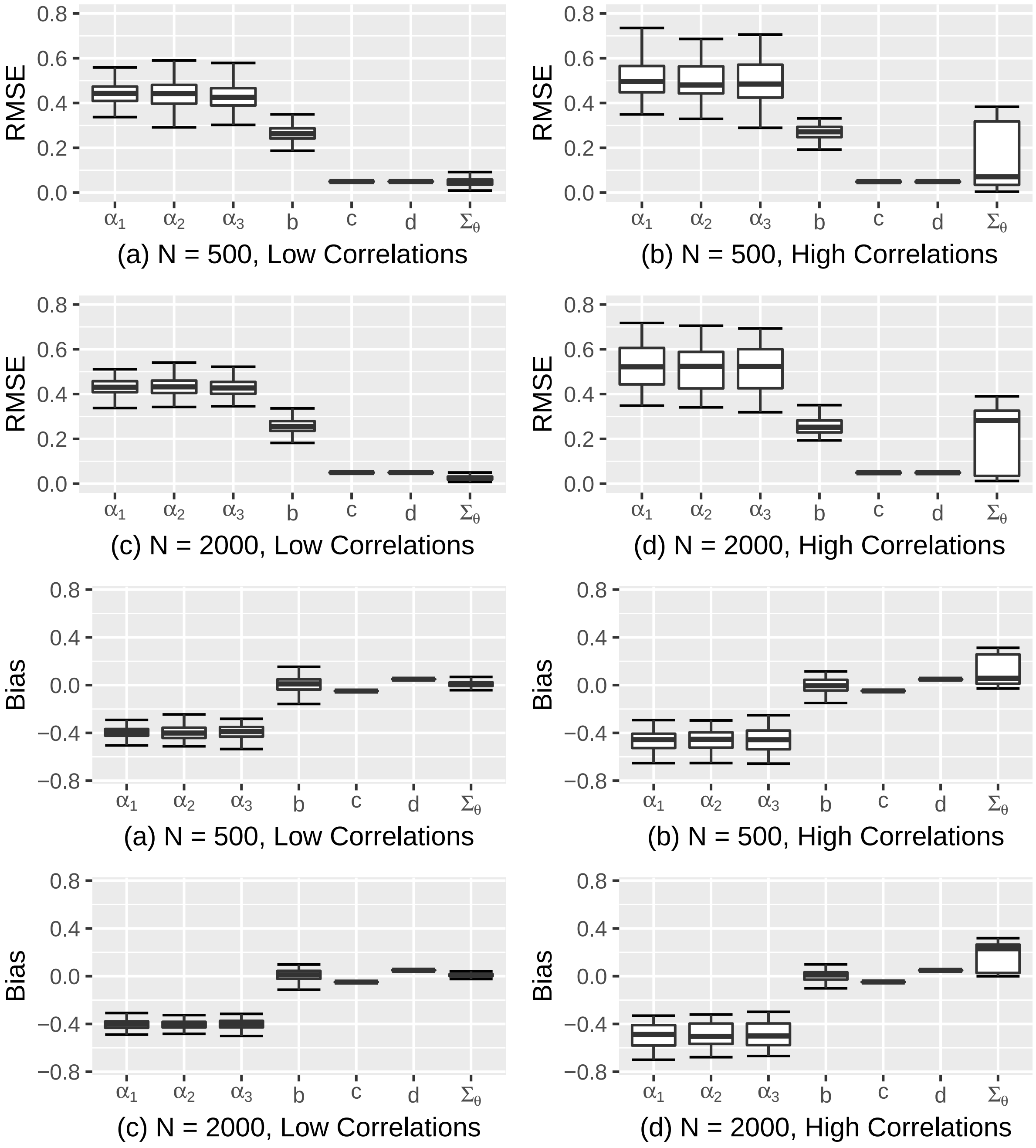
The manipulated conditions include: (i) correlations among the latent traits, and (iii) sample size. The latent traits $\boldsymbol{\theta}_i$ were generated from multivariate normal distribution, $N(0, \Sigma_{\boldsymbol{\theta}})$, where $\Sigma_{\boldsymbol{\theta}}$ is a covariance matrix whose diagonal elements were 1 and the off-diagonals were drawn from Uniform distribution. For the high correlation condition, the correlations were drawn from $Unif(0.5, 0.7)$ and for the low correlation condition, they were drawn from $Unif(0.1, 0.3)$. Sample size was set at either 500 or 2000 to consider both small sample and large sample scenarios.

In the confirmatory analysis, some of the item loading parameters are constrained to 0 based on the pre-specified item factor loading structure. In the simulation, there were 15 items loaded onto each factor with loading values set by $Unif(1, 2)$. In the exploratory analysis, we do not assume any constraint on the item discrimination parameter $\boldsymbol{A}$. Hence, the exploratory item factor analysis for M4PL is computationally more challenging scenario than confirmatory factor analysis. In our preliminary analysis in M4PL, we encountered several convergence issues with the exploratory factor analysis. This could be due to the fact that we need additional constraints in M4PL to ensure stable model estimation since the number of parameters to estimate is much higher and the structure is more complex in M4PL. For this section, we mainly present the results in confirmatory factor analysis with between item multidimensional structure. The more challenging simulation conditions will be further discussed in Section IV.4 as a topic for the future studies.

Figure IV.1 shows the distributions of RMSE and Bias of the model parameter under M4PL with between item multidimensional structure. Overall, higher factor correlation increases the RMSE and Biases as expected. In terms of varying sample sizes, N, the RMSE and Bias decreases with increasing sample size in low correlation conditions as observed in Figure IV.1 (a) and (c). Under the high correlation condition, however, we observe less decrease in both RMSE and Bias, which implies that parameter estimation in M4PL is more challenging especially with highly correlated latent factors. In all simulation conditions, we observe that the estimation error of c_j and d_j are relatively lower than the error of $\boldsymbol{\alpha}_j$. This suggests that the estimation of lower and upper asymptote values of the M4PL model is quite accurate and less challenging. However, inclusion of these additional asymptote parameters make the estimation of item discrimination $\boldsymbol{\alpha}_j$ more difficult. Computationally, the estimation is done in a few seconds for the simulation conditions discussed in this section. Overall, the performance of our proposed GVEM approach in challenging M4PL model is pretty impressive both in terms of the accuracy and computational time.

IV.3 Differential Item Functioning (DIF) Analysis

Measures of differential item functioning (DIF) are used to help ensure the fairness of tests in the applications of educational and psychometrics test. They can also be used to assess the effect of interventions and help building better intervention strategies for biomedical research. In this section, we study the DIF analysis in MIRT and develop the Gaussian Variational



(a) N = 500, Low Correlations

(b) N = 500, High Correlations

(c) N = 2000, Low Correlations

(d) N = 2000, High Correlations

(a) N = 500, Low Correlations

(b) N = 500, High Correlations

(c) N = 2000, Low Correlations

(d) N = 2000, High Correlations

Figure IV.1: Parameter recovery of the between-item M4PL models from confirmatory factor analysis

estimation procedure via regularization. For our discussion of DIF, we focus on the scenario with a binary predictor in the M2PL model to assess its effect on the probability of responses in assessment tests. The estimation procedure presented in this chapter is applicable to any binary predictor of interest. For simplicity, let us consider this binary predictor as gender, i.e. male or female, for the following discussion. We could understand the concept of DIF with respect to gender as *gender-biasedness* of each test items in the assessment test. Since gender effect may be present for each test items respectively, we consider gender coefficients specific to each item, β_{1j}. The model is as follows;

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i) = \frac{\exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j + \beta_{1j} G_i)}{1 + \exp(\boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - b_j + \beta_{1j} G_i)} \tag{IV.14}$$

where G_i denotes the gender status of the ith individual and non-zero β_{1j} implies the existence of gender-bias in the jth test item. Other model parameters are defined in the same fashion as in previously discussed MIRT models.

Our goal is identify the test items that have DIF. That is, we would like to study if certain gender group has higher chance of correctly answering the test items and if so, which test items they are. Hence, we formulate the problem of estimating gender effect on test items in MIRT as a variable selection problem, in which we penalize the coefficients for the observed predictors G_i so that we can detect the true non-zero coefficients among the β_{1j}'s for $j = 1, \dots, J$. We solve the following optimization problem with L_1 penalty,

$$(\hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda, \hat{\boldsymbol{\beta}}_{1\lambda}) = argmax_{\mathbf{A},\mathbf{B},\boldsymbol{\beta}_1} l(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1; \mathbf{Y}) - \lambda \sum_{j=1}^{J} |\beta_{1j}| \tag{IV.15}$$

where λ is the penalization parameter and $\boldsymbol{\beta}_1 = [\beta_{1j}]_{j=1,\dots,J} = \{\beta_{11}, \dots, \beta_{1J}\}$. As discussed in regularized variational estimation in Chapter III, we use variational lower bound in place of the complete log-likelihood. In essence, we empirically estimates the biasedness of the test items via the sparsity of the coefficient vector $\boldsymbol{\beta}_1$ under the GVEM framework. The non-zero β_{1j}'s implies that the associated jth test items are biased toward certain gender group. Optimal penalization parameter λ is chosen by the information criteria similarly as discussed in Chapter III.

IV.3.1 Algorithm Details

In this section, we illustrate how the regularized estimation using Gaussian Variaitonal EM approach can be applied to the DIF analysis in MIRT. The only difference between M2PL model and model equation (IV.14) is the existence of the observed binary predictor, G_i for $i = 1, \dots, N$. Hence the derivation of the EM procedures is similar to our previous discussion of the regularized GVEM in M2PL models.

We first find the optimal choice of variational distributions for $\boldsymbol{\theta}_i$, $i = 1, \dots, N$. By following the variational inference theory as in previous chapters, it can be shown that the variational distributions $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)$, $i = 1, \dots, N$ that minimizes the KL divergence with the posterior distributions $P(\boldsymbol{\theta}_i | \mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1)$ takes the following form;

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

which implies that the optimal variational distribution is $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i | \mu_i, \Sigma_i)$ where the mean parameter is

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

and the covariance matrix is

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

E-step Following the Gaussian variational approach developed previously, we evaluate the closed-form lower bound of the expected log likelihood with respect to the variational distributions q_i's under the model (IV.14) as follows;

$$\begin{aligned}
&E(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi}) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{J} \Bigg(\log \frac{\exp(\xi_{i,j})}{1+\exp(\xi_{i,j})} + Y_{ij}(\boldsymbol{\alpha}_j^\top \mu_i - b_j + \beta_{1j} G_i) + \frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top \mu_i - \beta_{1j} G_i - \xi_{i,j}) \\
&-\eta(\xi_{i,j})\{b_j^2 - 2b_j \boldsymbol{\alpha}_j^\top \mu_i + \boldsymbol{\alpha}_j^\top [\Sigma_i + (\mu_i)(\mu_i)^\top] \boldsymbol{\alpha}_j + \beta_{1j}^2 G_i - 2b_j \beta_{1j} G_i \\
&+2\beta_{1j} G_i \boldsymbol{\alpha}_j^\top \mu_i - \xi_{i,j}^2\} \Bigg) + \frac{N}{2} \log |\Sigma_{\boldsymbol{\theta}}^{-1}| - \sum_{i=1}^{N} \frac{1}{2} Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i + \mu_i \mu_i^\top]).
\end{aligned}$$

M-Step In this step, we maximize the $E(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi})$ to update the model parameters. This is simply achieved by setting the derivative of $E(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi})$ with respect to $(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi})$ to be zero, respectively. See appendix for more details. Then, $\boldsymbol{\alpha}_j$ is updated according to

$$\boldsymbol{\alpha}_j = \frac{1}{2} \Big[\sum_{i=1}^{N} \eta(\xi_{i,j})(\Sigma_i + \mu_i \mu_i^\top) \Big]^{-1} \sum_{i=1}^{N} \Big(Y_{ij} - \frac{1}{2} + 2b_j \eta(\xi_{i,j}) - 2\beta_{1j} G_i \eta(\xi_{i,j}) \Big) \mu_i^\top. \tag{IV.18}$$

The updating rule for b_j can be derived similarly, which is

$$b_j = \frac{1}{2} \Big[\sum_{i=1}^{N} \eta(\xi_{i,j}) \Big]^{-1} \sum_{i=1}^{N} \Big(\frac{1}{2} - Y_{ij} + 2\eta(\xi_{i,j}) \{ \boldsymbol{\alpha}_j^\top \mu_i + \beta_{1j} G_i + \beta_{2j}^\top \mu_i G_i \} \Big). \tag{IV.19}$$

Setting the derivative of the variational lowerbound with respect to $\xi_{i,j}$ equal to zero, we get the following updating rule for $\xi_{i,j}$;

$$\begin{aligned}
\xi_{i,j}^2 &= E[(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \beta_{1j} G_i)^2] \qquad \text{(IV.20)} \\
&= b_j^2 + \boldsymbol{\alpha}_j^\top \Big[\Sigma_i + \mu_i \mu_i^\top \Big] \boldsymbol{\alpha}_j + \beta_{1j}^2 G_i - 2b_j \boldsymbol{\alpha}_j^\top \mu_i - 2b_j \beta_{1j} G_i + 2\beta_{1j} G_i \boldsymbol{\alpha}_j^\top \mu_i.
\end{aligned}$$

The updating rule for $\Sigma_{\boldsymbol{\theta}}$ is

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

which is same with the updating rule for M2PL model presented in Chapter II.

To update the coefficient vector $\boldsymbol{\beta}_1$ with L_1 penalty, we use the coordinate descent algorithm by Friedman et al. (2010) as in Section III.3. The coordinate descent algorithm update each of the β_{1j} iteratively according to the following updating rule;

$$\beta_{1j} = \frac{S\Bigg(\sum_{i=1}^{N} Y_{ij}G_i - \frac{1}{2}G_i - \eta(\xi_{i,j})\{2G_i\boldsymbol{\alpha}_j^\top \mu_i - 2b_jG_i\}, \lambda\Bigg)}{\sum_{i=1}^{N} 2\eta(\xi_{i,j})G_i} \tag{IV.22}$$

where λ is the sparsity parameter of choice and the function S is a soft threshold operator such that $\mathrm{S}(\delta, \lambda) = sign(\delta)(|\delta| - \lambda)_+$. Refer to the appendix for detailed derivation of the updating rules presented above.

IV.3.2 Simulation Studies : DIF

In this section, we present the simulation results conducted to assess the performance of the proposed regularized GVEM approach for the DIF analysis in MIRT. We focus on M2PL models as discussed in previous sections. The number of test dimensions was fixed at 3 and test length was fixed at 45. The manipulated conditions include: (i) multidimensional structure, i.e. between-item multidimensionality and within-item multidimensionality; (ii) correlations among the latent traits, and (iii) sample size. Similarly as in the previous studies, for the between-item multidimensional structure, there were 15 items loaded onto each factor; whereas for the within-item multidimensional structure, about one third of the items were loaded onto one, two, and three factors respectively. In all cases, item discrimination parameters were simulated from $Unif(1, 2)$ distribution, and difficulty parameter b_j was simulated from the standard normal distribution. The latent traits $\boldsymbol{\theta}_i$ were generated from multivariate normal distribution, $N(0, \Sigma_{\boldsymbol{\theta}})$, where $\Sigma_{\boldsymbol{\theta}}$ is a covariance matrix whose diagonal elements were 1 and the off-diagonals were drawn from Uniform distribution. Three different correlation conditions were studied. That is, the correlations were drawn from $Unif(0.5, 0.7)$ for the high correlation condition, $Unif(0.1, 0.3)$ for the low correlation condition and lastly they were set to zero for the condition with no factor correlations. Sample size was set at either 500 or 2000 to study to effect of varying sample sizes.

50 replications were conducted for each condition. Evaluation criteria include the correct estimation rates(%) of the nonzero structure of the coefficients for the Gender predictor. Again, the goal here is to detect any gender-biased test items via regularization. Additionally we present the false positive and false negative rates(%) in each condition to further study the performance of the proposed method in correctly estimating the gender biasedness of test items. In the context of our DIF problem, false positive rate is the probability of falsely identifying test items as gender biased. Similarly, false negative rate is the probability of falsely identifying test items as not biased for any gender group. Naturally, we would like to observe low false positive and low false negative rates.

Figure IV.2 and Figure IV.3 show the correct estimation rates(%) under the between-item and within-item multidimensional structure, respectively. Overall, the correct estimation rates are pretty high with most above 90% rates although we do observe several cases with lower rates in more challenging scenario, which is within item model with large factor correlations and small sample size (i.e. $N = 500$). The performance gets better on average as sample size increases in all correlation and multidimensional structure conditions. However, this pattern is not very strong probably since the correct estimation rate are already pretty high in all conditions.

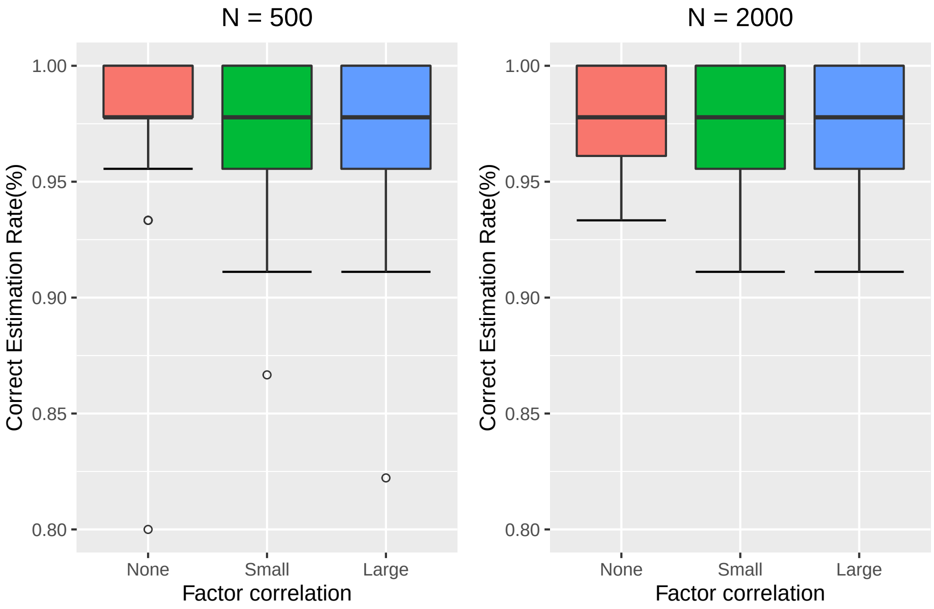


Figure IV.2: Correct Estimation Rates(%) under the between-item multidimensional structure

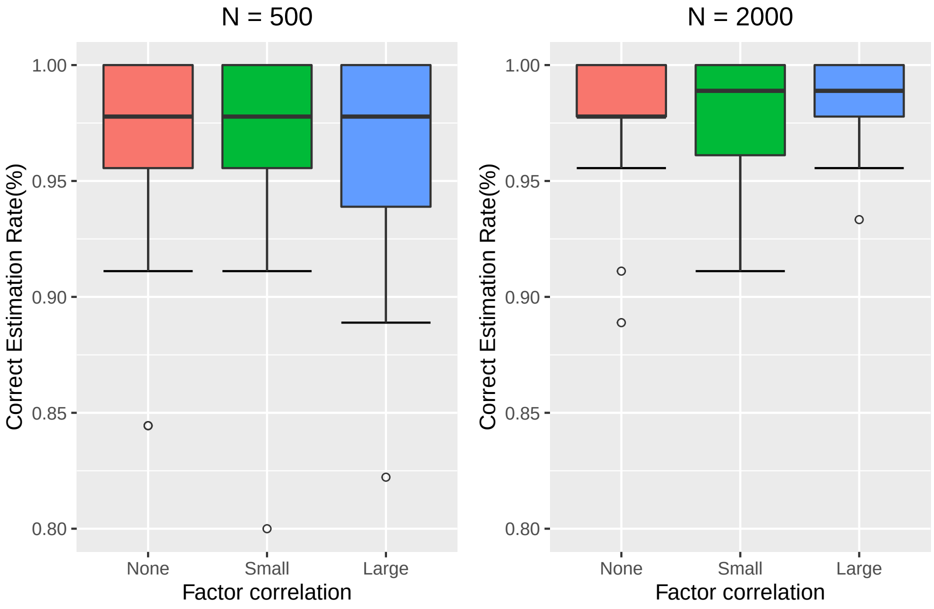


Figure IV.3: Correct Estimation Rates(%) for the within-item multidimensional structure

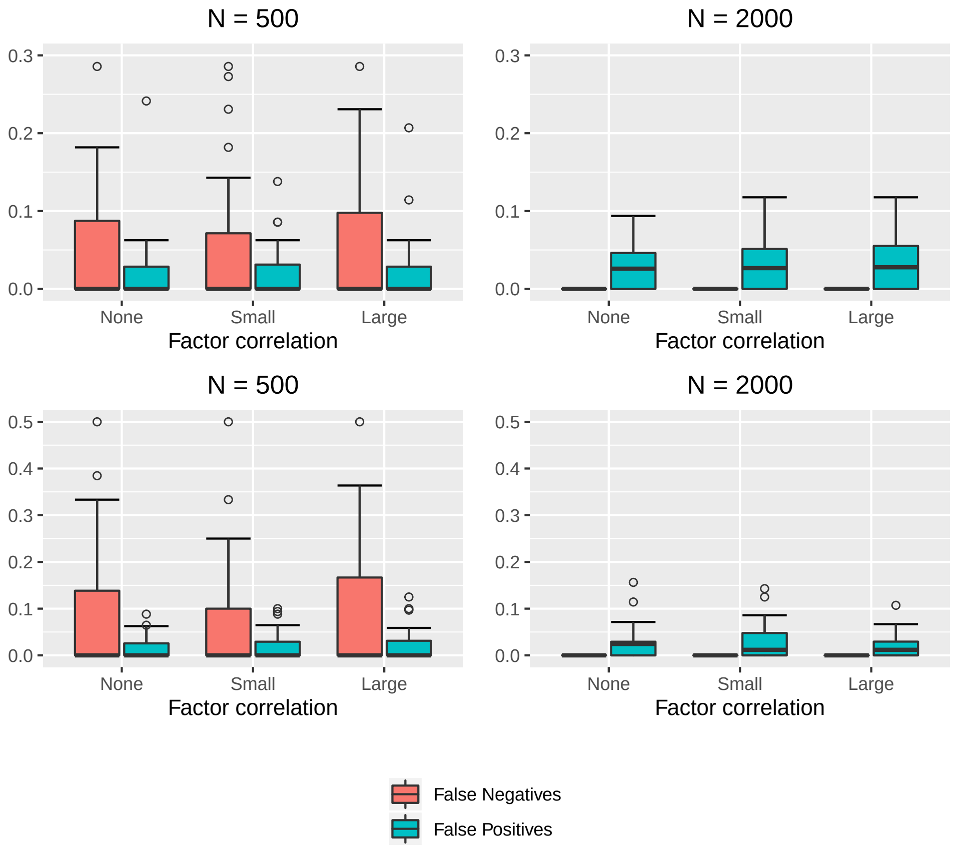


Figure IV.4: False Positive and False Negative rates in between-item model (first row) and within- item model (second row)

Figure IV.4 show the false positive and false negative rates in between-item model (first row) and within- item model (second row), respectively. We can clearly observe from false positive and false negative rates that the performance of the regularized estimation for DIF gets better with increasing sample sizes. Especially, false negative rages decreases to 0 in all conditions with large sample size (i.e. $N = 2000$).

IV.4 Discussions

In this chapter, we discussed two interesting extensions of our proposed Gaussian variational estimation approach; the parameter estimation in Multidimensional 4-Parameter Logistic (M4PL) model and Differential Item Functioning (DIF) analysis in MIRT. We developed

Gaussian Variational EM algorithm for M4PL as a computationally efficient and accurate estimation method for M4PL parameters. In addition, we applied the regularized variational EM approach for DIF analysis in order to assess the item bias in assessment tests with an example with gender. Through the discussions on the two extensions of GVEM, we illustrate how well the GVEM approach can be applied in various aspects of analyzing educational and psychological assessment data. A series of simulation studies demonstrate that the proposed method performs pretty well in the parameter estimation for M4PL and DIF analysis for M2PL model.

However, there are some challenges remained. For M4PL model, we observed that the parameter estimation gets less stable and encounters convergence issues with exploratory factor analysis and within item multidimensional structure. This could be due to the identifiability issue since we estimate too many parameters in the M4PL exploratory analysis and have less constraints than in confirmatory analysis. This suggests that penalization on model parameters may be helpful in reducing the number of parameters to estimate and thus making the estimation more stable. We leave this for the future research. Another interesting future research is the DIF analysis involving multiple covariates and even interactions between the latent traits and covariate. For example, one's probability of response would be dependent on a combination of his or her characteristics such as gender, race, income group, and ethnicity. Hence, it would be interesting to study the estimation methods for scenario with a combination of binary predictor (e.g. gender) and multiple categorical covariate (e.g. race or ethnicity). In addition, estimation problem with interaction between latent traits $\boldsymbol{\theta}_i$ and covariate (e.g. G_i) needs further study. Discussions on the model identifiability and necessary conditions to achieve consistent estimation would be necessary. It would be interesting to see if the Gaussian Variational approach can be developed as the estimating algorithm.

Appendix of Chapter IV

In this section, we provide the detailed derivation of the estimation procedures which are presented in Chapter IV. In Appendix A, we present the detailed derivation of the E and M steps for the proposed GVEM method for M4PL model. Appendix B presents the detailed derivation of the Gaussian Variational regularized estimation procedure for the DIF analysis in M2PL model.

IV.A Derivation of EM steps for M4PL

With the optimal choice of variational distributions $q_{\boldsymbol{\theta}}$ and q_z, we have the closed form expression of variational lower bound to the marginal log-likelihood.

$$E(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}) := \sum_{i=1}^{N}\int_{\boldsymbol{\theta}_i}\left[\sum_{j=1}^{J}\sum_{Z_{ij}} l(\mathbf{Y}_i,\boldsymbol{\theta}_i,\mathbf{Z}_i,\boldsymbol{\xi}_i \mid \mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D})q_z(Z_{ij})\right]q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$$

To evaluate this, we first need to choose optimal variational distributions that would give us a nice closed form solution.

Choice of $q_{\boldsymbol{\theta}}$ The expectation of the variational lower bound with respect to variational density of Z_{ij} is

$$\begin{aligned}
& E_Z(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}) \\
= & \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{Z_{ij}} l(\mathbf{Y}_i,\boldsymbol{\theta}_i,\mathbf{Z}_i,\boldsymbol{\xi}_i \mid \mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}) q_z(Z_{ij}) \\
= & \sum_{i=1}^{N}\Bigg[\sum_{j=1}^{J} Y_{ij}\{E_z[Z_{ij}]\log d_j + (1-E_z[Z_{ij}])\log c_j\} + \sum_{j=1}^{J}(1-Y_{ij})E_z[\log(1-d_j^{Z_{ij}}c_j^{1-Z_{ij}})] \\
& + \log P(\boldsymbol{\theta}_i) + \sum_{j=1}^{J} E_z[Z_{ij}](\boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i - b_j) + \sum_{j=1}^{J}\log\frac{e^{\xi_{i,j}}}{1+e^{\xi_{i,j}}} + \sum_{j=1}^{J}\frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i - \xi_{i,j}) \\
& - \sum_{j=1}^{J}\eta(\xi_{i,j})\{(b_j - \boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i)^2 - \xi_{i,j}^2\}\Bigg]
\end{aligned}$$

Then we can show that variational distributions $q_{\boldsymbol{\theta}}$ follows the following form;

$$\log q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \propto \sum_{j=1}^{J}(E_z[Z_{ij}] - \frac{1}{2})\boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i - \sum_{j=1}^{J}\eta(\xi_{i,j})(b_j - \boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i)^2 - \frac{1}{2}\boldsymbol{\theta}_i^\top\Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}_i$$

Thus, the optimal choice of is $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) \sim N(\boldsymbol{\theta}_i \mid \mu_i, \Sigma_i)$ with

$$\begin{aligned}
\mu_i &= \Sigma_i \times \sum_{j=1}^{J}\{2\eta(\xi_{i,j})b_j + E_z[Z_{ij}] - \frac{1}{2}\boldsymbol{\alpha}_j^\top\} && \text{(IV.A.1)} \\
\Sigma_i &= \Sigma_{\boldsymbol{\theta}}^{-1} + 2\sum_{j=1}^{J}\eta(\xi_{i,j})\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j && \text{(IV.A.2)}
\end{aligned}$$

where we denote $E_z[Z_{ij}] = s_{ij}$ in the following derivation.

Choice of q_z The expectation of variational lower bound with respect to $\boldsymbol{\theta}_i$ is

$$
\begin{aligned}
& E_{\boldsymbol{\theta}}(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}) \\
= & \sum_{i=1}^{N}\sum_{j=1}^{J}\int l(\mathbf{Y}_i,\boldsymbol{\theta}_i,\mathbf{Z}_i,\boldsymbol{\xi}_i)q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \\
= & \sum_{i=1}^{N}\Bigg[\sum_{j=1}^{J}Y_{ij}\{Z_{ij}\log d_j+(1-Z_{ij})\log c_j\}+\sum_{j=1}^{J}(1-Y_{ij})\log(1-d_j^{Z_{ij}}c_j^{1-Z_{ij}}) \\
& +E_{\boldsymbol{\theta}}[\log P(\boldsymbol{\theta}_i)]+\sum_{j=1}^{J}Z_{ij}(\boldsymbol{\alpha}_j^{\top}E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i]-b_j)+\sum_{j=1}^{J}\log\frac{e^{\xi_{i,j}}}{1+e^{\xi_{ij}}} \\
& +\sum_{j=1}^{J}\frac{1}{2}(b_j-\boldsymbol{\alpha}_j^{\top}E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i]-\xi_{ij})-\sum_{j=1}^{J}\eta(\xi_{i,j})\{E_{\boldsymbol{\theta}}[(b_j-\boldsymbol{\alpha}_j^{\top}\boldsymbol{\theta}_i)^2]-\xi_{i,j}^2\}\Bigg]
\end{aligned}
$$

Then, posterior distribution of Z_{ij} has the following form;

$$
\begin{aligned}
\log q_z(Z_{ij}) \propto\ & Y_{ij}Z_{ij}\log(d_j)+Y_{ij}(1-Z_{ij})\log(c_j)+(1-Y_{ij})\log(1-d_j^{Z_{ij}}c_j^{1-Z_{ij}}) \\
& +Z_{ij}(\boldsymbol{\alpha}_j^{\top}E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i]-b_j) \\
=\ & Y_{ij}Z_{ij}\log(d_j)+Y_{ij}(1-Z_{ij})\log(c_j)+(1-Y_{ij})\{Z_{ij}\log(1-d_j)+ \\
& (1-Z_{ij})\log(1-c_j)\}+Z_{ij}(\boldsymbol{\alpha}_j^{\top}E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i]-b_j) \\
=\ & Z_{ij}\{Y_{ij}\log(d_j)+(1-Y_{ij})\log(1-d_j)+\boldsymbol{\alpha}_j^{\top}E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i]-b_j\} \\
& +(1-Z_{ij})\{Y_{ij}\log(c_j)+(1-Y_{ij})\log(1-c_j)\}
\end{aligned}
$$

This implies that the optimal choice of the variational density is $q_z(Z_{ij})\sim Bernoulli(s_{ij})$ where

$$
s_{ij}=\frac{1}{1+\exp\left[Y_{ij}\log(\frac{c_j}{d_j})+(1-Y_{ij})\log(\frac{1-c_j}{1-d_j})-\boldsymbol{\alpha}_j^{\top}E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i]+b_j\right]} \qquad \text{(IV.A.3)}
$$

E step With optimally chosen q_z and $q_{\boldsymbol{\theta}}$, we can derive the closed form variational lower bound.

$$\begin{aligned}
& E[\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\boldsymbol{\xi}] \\
= & \sum_{i=1}^{N}\Bigg[\sum_{j=1}^{J} Y_{ij}\{E_z[Z_{ij}]\log d_j + (1-E_z[Z_{ij}])\log c_j\} + \sum_{j=1}^{J}(1-Y_{ij})\{E_z[Z_{ij}]\log(1-d_j) \\
& +(1-E_z[Z_{ij}])\log(1-c_j)\} + E_{\boldsymbol{\theta}}[\log P(\boldsymbol{\theta}_i)] + \sum_{j=1}^{J} E_z[Z_{ij}](\boldsymbol{\alpha}_j^\top E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i] - b_j) + \sum_{j=1}^{J}\log\frac{e^{\xi_{i,j}}}{1+e^{\xi_{ij}}} \\
& +\sum_{j=1}^{J}\frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i] - \xi_{ij}) - \sum_{j=1}^{J}\eta(\xi_{i,j})\{E_{\boldsymbol{\theta}}[(b_j - \boldsymbol{\alpha}_j^\top\boldsymbol{\theta}_i)^2] - \xi_{i,j}^2\}\Bigg] \\
= & \sum_{i=1}^{N}\sum_{j=1}^{J}\Bigg[Y_{ij}\{s_{ij}\log d_j + (1-s_{ij})\log c_j\} + (1-Y_{ij})\{s_{ij}\log(1-d_j) + (1-s_{ij})\log(1-c_j)\}\Bigg] \\
& -\sum_{i=1}^{N}\frac{1}{2}Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i + \mu_i\mu_i^\top]) + \frac{N}{2}\log|\Sigma_{\boldsymbol{\theta}}^{-1}| + \sum_{i=1}^{N}\sum_{j=1}^{J}\Bigg[s_{ij}(\boldsymbol{\alpha}_j^\top\mu_i - b_j) + \log\frac{e^{\xi_{i,j}}}{1+e^{\xi_{ij}}} \\
& +\frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top\mu_i - \xi_{ij}) - \eta(\xi_{i,j})\{b_j^2 - 2b_j\boldsymbol{\alpha}_j^\top\mu_i + \boldsymbol{\alpha}_j^\top[\Sigma_i + \mu_i\mu_i^\top]\boldsymbol{\alpha}_j - \xi_{i,j}^2\}\Bigg]
\end{aligned}$$

M step We update model parameters by taking derivatives of variational lower bound and setting them equal to zero. First, the updating rule for $\boldsymbol{\alpha}_j$ can be derived as follows. Here, Q_j denotes the variational lower bound for the jth item only. By setting the derivative of the variational lower bound with respect to $\boldsymbol{\alpha}_j$ as zero, we get

$$\frac{\partial Q_j(\boldsymbol{\alpha}_j)}{\partial\boldsymbol{\alpha}_j} = \sum_{i=1}^{N}\Bigg[(s_{ij} - \frac{1}{2})\mu_i^\top + 2b_j\eta(\xi_{i,j})\mu_i^\top - 2\eta(\xi_{i,j})[\Sigma_i + \mu_i\mu_i^\top]\boldsymbol{\alpha}_j\Bigg].$$

Then, $\boldsymbol{\alpha}_j$ is updated according to

$$\boldsymbol{\alpha}_j \;=\; \frac{1}{2}\Bigg[\sum_{i=1}^{N}\eta(\xi_{i,j})[\Sigma_i + \mu_i\mu_i^\top]\Bigg]^{-1}\sum_{i=1}^{N}\Bigg[\{s_{ij} - \frac{1}{2} + 2b_j\eta(\xi_{i,j})\}\mu_i^\top\Bigg].$$

Similarly for b_j, we set the following derivative equal to zero.

$$\frac{\partial Q_j(b_j)}{\partial b_j} = \sum_{i=1}^{N}[-s_{ij} + \frac{1}{2} - \eta(\xi_{i,j})\{2b_j - 2\boldsymbol{\alpha}_j^\top \mu_i\}] = 0$$

Thus, we update b_j by

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

For c_j, we have

$$\frac{\partial Q_j(c_j)}{\partial c_j} = \sum_{i=1}^{N}\Big[Y_{ij}(1 - s_{ij})\frac{1}{c_j} - (1 - Y_{ij})(1 - s_{ij})\frac{1}{1 - c_j}\Big] = 0$$

We update c_j according to

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

Lastly for d_j

$$\frac{\partial Q_j(d_j)}{\partial d_j} = \sum_{i=1}^{N}\Big[Y_{ij}s_{ij}\frac{1}{d_j} - (1 - Y_{ij})s_{ij}\frac{1}{1 - d_j}\Big] = 0$$

Thus, we update d_j according to

$$d_j = \frac{\sum_{i=1}^{N} Y_{ij}s_{ij}}{\sum_{i=1}^{N} s_{ij}}$$

$\Sigma_{\boldsymbol{\theta}}$ and $\xi_{i,j}$ have the same updating rule as in the EM steps for M3PL.

IV.B Derivation for DIF analysis

The derivation of the variational lower bound of the expected log likelihood and the optimal variational distributions $q_i(\boldsymbol{\theta}_i)$ are simply extensions of the derivations shown for M2PL models and thus are straightforward to derive. Hence, in this Appendix we start directly with the EM steps.

E-step Similarly as presented for M2PL model, we evaluate the closed-form lower bound of the expected log likelihood with respect to the variational distributions $q_{\boldsymbol{\theta}}$'s under the model (IV.14) as follows;

$$
\begin{aligned}
&E(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi})\\
&:= \sum_{i=1}^{N} \int_{\boldsymbol{\theta}_i} \Big[l(\mathbf{Y}_i, \boldsymbol{\theta}_i, \boldsymbol{\xi}_i \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1) \Big] q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i\\
&= \sum_{i=1}^{N}\sum_{j=1}^{J} \Bigg(\log \frac{\exp(\xi_{i,j})}{1+\exp(\xi_{i,j})} + Y_{ij}(\boldsymbol{\alpha}_j^\top E_i[\boldsymbol{\theta}_i] - b_j + \beta_{1j} G_i) + \frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top E_i[\boldsymbol{\theta}_i] - \beta_{1j} G_i - \xi_{i,j})\\
&-\eta(\xi_{i,j})\{E_i[(b_j - \boldsymbol{\alpha}_j^\top \boldsymbol{\theta}_i - \beta_{1j} G_i)^2] - \xi_{i,j}^2\} \Bigg) + \frac{N}{2}\log|\Sigma_{\boldsymbol{\theta}}^{-1}| - \sum_{i=1}^{N} \frac{1}{2} Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i + \mu_i \mu_i^\top])\\
&= \sum_{i=1}^{N}\sum_{j=1}^{J} \Bigg(\log \frac{\exp(\xi_{i,j})}{1+\exp(\xi_{i,j})} + Y_{ij}(\boldsymbol{\alpha}_j^\top \mu_i - b_j + \beta_{1j} G_i) + \frac{1}{2}(b_j - \boldsymbol{\alpha}_j^\top \mu_i - \beta_{1j} G_i - \xi_{i,j})\\
&-\eta(\xi_{i,j})\{b_j^2 - 2b_j \boldsymbol{\alpha}_j^\top \mu_i + \boldsymbol{\alpha}_j^\top [\Sigma_i + (\mu_i)(\mu_i)^\top]\boldsymbol{\alpha}_j + \beta_{1j}^2 G_i - 2b_j \beta_{1j} G_i\\
&+2\beta_{1j} G_i \boldsymbol{\alpha}_j^\top \mu_i - \xi_{i,j}^2\} \Bigg) + \frac{N}{2}\log|\Sigma_{\boldsymbol{\theta}}^{-1}| - \sum_{i=1}^{N} \frac{1}{2} Tr(\Sigma_{\boldsymbol{\theta}}^{-1}[\Sigma_i + \mu_i \mu_i^\top])
\end{aligned}
$$

where $E_{\boldsymbol{\theta}}[\boldsymbol{\theta}_i] = \mu_i$ and $Cov[\boldsymbol{\theta}_i] = \Sigma_i$, which are the model parameters for the variational distributions $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_i)$.

M-Step In this step, we maximize the $E(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi})$ to update the model parameters. This is simply achieved by setting the derivative of $E(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi})$ with respect to $(\mathbf{A}, \mathbf{B}, \boldsymbol{\beta}_1, \boldsymbol{\xi})$ to be zero, respectively.

First, consider the $\boldsymbol{\alpha}_j$. The derivative of the variational lower bound w.r.t. $\boldsymbol{\alpha}_j$ is as follows;

$$\begin{aligned}\frac{\partial E(\mathbf{A},\mathbf{B},\boldsymbol{\beta}_1,\boldsymbol{\xi})}{\partial \boldsymbol{\alpha}_j} &= \sum_{i=1}^{N}(Y_{ij}-\frac{1}{2})\mu_i^\top + 2b_j\eta(\xi_{i,j})\mu_i^\top - 2\eta(\xi_{i,j})[\Sigma_i+\mu_i\mu_i^\top]\boldsymbol{\alpha}_j \\ &\quad -2\beta_{1j}G_i\eta(\xi_{i,j})(\mu_i)^\top.\end{aligned}$$

Setting it equal to zero, we get the updating rule for $\boldsymbol{\alpha}_j$.

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

Now similarly, we take derivative with respect to b_j.

$$\frac{\partial E(\mathbf{A},\mathbf{B},\boldsymbol{\beta}_1,\boldsymbol{\xi})}{\partial b_j} = \sum_{i=1}^{N}(-Y_{ij}+\frac{1}{2}-\eta(\xi_{i,j})\{2b_j-2\boldsymbol{\alpha}_j^\top\mu_i-2\beta_{1j}G_i\}) = 0$$

This implies that b_j is updated according to

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

To update the coefficient vector $\boldsymbol{\beta}_1$ with L_1 regularization, we first evaluate first and second order derivatives of the variational lower bound with respect to β_{1j}, which are

$$\frac{\partial E(\mathbf{A},\mathbf{B},\boldsymbol{\beta}_1,\boldsymbol{\xi})}{\partial \beta_{1j}} = \sum_{i=1}^{N}(Y_{ij}G_i-\frac{1}{2}G_i-\eta(\xi_{i,j})\{2\beta_{1j}G_i-2b_jG_i+2G_i\boldsymbol{\alpha}_j^\top\mu_i\})$$

and

$$\frac{\partial^2 E(\mathbf{A},\mathbf{B},\boldsymbol{\beta}_1,\boldsymbol{\xi})}{\partial \beta_{1j}^2} = -\sum_{i=1}^{N}2\eta(\xi_{i,j})G_i.$$

By the coordinate descent algorithm by Friedman et al. (2010), β_{1j} is updated as follows;

$$\beta_{1j} = -\frac{S(-\frac{\partial^2 E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\beta}_1,\boldsymbol{\xi})}{\partial \beta_{1j}^2}\times\beta_{1j}+\frac{\partial E(\boldsymbol{A},\boldsymbol{B},\boldsymbol{\beta}_1,\boldsymbol{\xi})}{\partial\beta_{1j}},\lambda)}{\frac{\partial^2 E(\mathbf{A},\mathbf{B},\boldsymbol{\beta}_1,\boldsymbol{\xi})}{\partial\beta_{1j}^2}}$$

where λ is the sparsity parameter of choice and the function S is a soft threshold operator such that $\mathrm{S}(\delta, \lambda) = sign(\delta)(|\delta| - \lambda)_{+}$. Evaluating above, we can show that each of the β_{1j} is updated iteratively according to the following updating rule;

$$\beta_{1j} = \frac{S\Bigg(\sum_{i=1}^{N} Y_{ij} G_i - \frac{1}{2} G_i - \eta(\xi_{i,j}) \{ 2 G_i \boldsymbol{\alpha}_j^{\top} \mu_i - 2 b_j G_i \}, \lambda \Bigg)}{\sum_{i=1}^{N} 2\eta(\xi_{i,j}) G_i}$$

for $j = 1, \ldots, J$.

BIBLIOGRAPHY

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91.

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

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques.* CRC Press.

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), i–8.

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561.

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

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

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

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

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

Breheny, P., & Huang, J. (2011). Coordinate descent algorithm for nonconvex penalized regression, with application to biological feature selection. *The Annals of Applied Statistics*, *5*, 232-253.

Breslau, J., Javaras, K. N., Blacker, D., Murphy, J. M., & Normand, S. L. T. (2008). Differential item functioning between ethnic groups in the epidemiological assessment of depression. *The Journal of Nervous and Mental Disease*, *196*(4), 297.

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

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

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

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

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

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

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

Cho, A. E., Zhang, X., Wang, C., & Xu, G. (2020). Gaussian Variational Estimation for Multidimensional Item Response Theory.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., . . . Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research*, *16*(1), 69.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, *81*(4), 1142–1163.

Culpepper, S. A. (2017). The prevalence and implications of slipping on low-stakes, large-scale assessments. *Journal of Educational and Behavioral Statistics*, *42*(6), 706–725.

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

De La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* L. Erlbaum Associates.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348-1360.

Feuerstahler, L. M., & Waller, N. G. (2014). Estimation of the 4-parameter model with marginal maximum likelihood. *Multivariate Behavioral Research*, *49*(3), 285–285.

Gulliksen, H. (1950). *Theory of mental tests.* John Wiley & Sons In.

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

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

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference.

The Journal of Machine Learning Research, *14*(1), 1303–1347.

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

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

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

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

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

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

Kwakkenbos, L., Willems, L. M., Baron, M., Hudson, M., Cella, D., Van Den Ende, C. H., ... Canadian Scleroderma Research Group (2014). The comparability of English, French and Dutch scores on the Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-F): an assessment of differential item functioning in patients with systemic sclerosis. *PLOS One*, *9*(3).

Lewis, T. T., Yang, F. M., Jacobs, E. A., & Fitchett, G. (2012). Racial/ethnic differences in responses to the everyday discrimination scale: a differential item functioning analysis. *American Journal of Epidemiology*, *175*(5), 391–401.

Liao, W. W., Ho, R. G., Yen, Y. C., & Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, *40*(10), 1679–1694.

Linacre, J. (2004). Discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions*, *18*(1), 959–960.

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

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three parameter logistic model. *Educational and Psychological Measurement*, *28*, 989-1020.

Martín, E. S., Del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*(3), 183–203.

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

McKinley, R. L., & Reckase, M. D. (1982). *The Use of the General Rasch Model with Multidimensional Item Response Data.* American Coll Testing Program, Iowa City, IA.

Meng, X., Xu, G., Zhang, J., & Tao, J. (2019). Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *British Journal of Mathematical and Statistical Psychology.*

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

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

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

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

Parisi, G. (1988). *Statistical field theory.* Addison-Wesley.

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

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

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*(2), 164.

Revuelta, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, *38*(7), 549–562.

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

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

Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. *Journal of Personality Assessment*, *72*(2), 282–307.

Rulison, K. L., & Loken, E. (2009). I've fallen and i can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, *33*(2), 83–101.

Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing*, *3*(4), 365–384.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, *33*(1), 75–102.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 161–169.

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology*, *5*(4), 417.

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

Tavares, H. R., Andrade, D. F. D., & Pereira, C. A. D. B. (2004). Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology*, *27*(4), 679–685.

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

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

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

Uebelacker, L. A., Strong, D., Weinstock, L. M., & Miller, I. W. (2009). Use of item response theory to understand differential functioning of DSM-IV major depression symptoms by race, ethnicity and gender. *Psychological Medicine*, *39*(4), 591–601.

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

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

Waller, N. G., & Feuerstahler, L. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate Behavioral*

Research, *52*(3), 350–370.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. E. Embretson (Ed.). *Measuring psychological constructs: Advances in model based approaches*, 147–173. American Psychological Association.

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

Wang, T., Xu, P. R., & Zhu, L. X. (2012). Non-convex penalized estimation in high-dimensional models with single-index structure. *Journal of Multivariate Analysis*, *109*, 221–235.

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

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

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

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.