Regularization Approaches to Detect Differential Item Functioning: Multiple Covariates, Polytomous Response, and Multidimensional Traits

Ruoyi Zhu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Chun Wang, Chair

Elizabeth A. Sanders

Oscar L. Olvera Astivia

Program Authorized to Offer Degree:
College of Education

University of Washington

Abstract

Regularization Approaches to Detect Differential Item Functioning: Multiple Covariates, Polytomous Response, and Multidimensional Traits

Ruoyi Zhu

Chair of the Supervisory Committee:
Chun Wang
College of Education

In the field of educational measurement, ensuring that assessment instruments are free from bias is crucial. Differential Item Functioning (DIF) indicates a potential bias in test items, implying that different covariate groups, despite having the same latent trait level, have differing probabilities of responding correctly to an item. Such biases could lead to unfair advantages or disadvantages for certain groups. While several methods exist for DIF detection, there is a growing need for more efficient and robust techniques, especially in situations with multiple covariates and multidimensional latent traits. To address this gap, I propose utilizing regularization methods to detect DIF associated with multiple covariates and multidimensional polytomous response data. The proposed algorithm concurrently estimates DIF effects across numerous covariates, encompassing both continuous and categorical variables. An additional advantage of our approach is the elimination of the necessity for anchor items, thus simplifying the DIF detection process.

Furthermore, I introduce two distinct methods to model the impact of covariates on the covariance matrix of the multidimensional latent trait. The first method employs Cholesky decomposition of the covariance matrix, while the second method utilizes covariance regression, which can effectively handle high-dimensional latent traits and a large number of covariates. By incorporating these procedures into our regularization-based DIF detection framework, the proposed method can accurately and effectively recover the impact of these

covariates.

To assess the performance of the proposed methods, I conduct three simulation studies. The first study focuses on examining uniform DIF associated with three covariates in a two-dimensional graded response model (GRM). The second study explores non-uniform DIF associated with the same set of covariates within the framework of the two-dimensional GRM. Lastly, the third simulation study investigates the efficacy of the group Lasso algorithm under the condition of uniform DIF associated with three covariates using the two-dimensional two-parameter logistic (2D2PL) model. Following the simulation studies, I conduct a real data analysis using the Patient-Reported Outcomes Measurement Information System (PROMIS) dataset.

In conclusion, this dissertation presents an emerging approach to addressing the challenges of DIF detection within the complex settings of multiple covariates, polytomous responses, and multidimensional latent traits. By incorporating advanced regularization methods and innovative impact modeling techniques, I aim to enhance the efficacy and robustness of identifying potential biases in educational assessments. The encouraging results from the simulation studies and real data analysis suggest that the proposed methods could significantly contribute to improved fairness and precision in educational measurement.

ACKNOWLEDGMENTS

The work is supported by IES R305D200015, NSF, and the University of Washington BIRCH center M-PARC award.

My heartfelt thanks go to my mom and dad for their endless love and unwavering support. You both have been the bedrock of my security and well-being, providing a nurturing and loving home. Your sacrifices have been the cornerstone of my achievements, and they have not gone unnoticed.

Special thanks to my boyfriend, Carl, whose support has been my anchor during a significant portion of my PhD journey. I am also profoundly grateful to the friends I met in Minnesota and Seattle. Your hospitality and friendship created comforting havens that truly felt like home whenever I was in need. Thank you for being my family away from home and for all the laughter and support that helped sustain me through my journey.

A heartfelt thanks to my peers in the UW Measurement and Statistics program for making my graduate school years enriching and enjoyable.

Lastly, I am profoundly thankful to my advisor, Chun, whose patience and support have profoundly impacted my life and academic journey. Thank you, Chun, for providing indispensable guidance that has shaped me into a qualified scientist. My educational journey would have been significantly less fruitful without your mentorship and advisement. I am also thankful to my committee members—Liz, Elena, Oscar, and Julia—and my faculty co-mentors, Paul and Min, for your invaluable expertise and wise counsel throughout this research and my broader academic and professional pursuits.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Number Page

GLOSSARY

BIC: Bayesian Information Criterion

CTT: Classical Test Theory

DIF: Differential Item Functioning

GRM: Graded Response Model

IRT: Item Response Theory

LASSO: Least Absolute Selection and Shrinkage Operator

2D2PL: Two-dimensional two-parameter logistic model

Chapter 1

INTRODUCTION

Achieving fairness in educational assessments requires the development of accurate tools that identify and resolve biases, such as DIF. DIF occurs when individuals from different groups, despite having similar abilities, have differing probabilities of answering a question correctly, leading to unfairness and misrepresentation of true skills. Earlier methods for detecting DIF have provided valuable insights; however, today's assessments tend to involve various demographic variables, intricate response formats, high-dimensional latent traits, and typically lack sufficient known non-DIF items, making them more complex. More advanced methods are needed. This study introduces a suite of new regularization approaches designed to address the challenges of modern DIF detection.

1.1 Literature Review

1.1.1 Traditional and Emerging Methods in DIF Detection

The development of DIF detection methods has a long history, with the earliest methods tracing back to the 1950s. The earlest DIF detection results are based on Classical Test Theory (CTT), and one of the most famous CTT based DIF detection methods is the Mantel-Haenszel test (Holland & Thayer, 1986; Mantel, 1963; Mantel & Haenszel, 1959). The Mantel-Haenszel test method checks if test items perform equally across different groups when overall test scores are controlled. Specifically, examinees are grouped based on their total scores. For each item within these groups, a 2x2 table is created to calculate the success odds for one group versus another. These odds are then combined into a weighted average to reveal any significant differences in item responses, regardless of overall ability.

The Mantel-Haenszel test is simple, effective for binary data, and works well with large samples. However, it is limited to items with two possible answers and comparisons between only two groups. To address more complex situations, the Generalized Mantel-Haenszel

(GMH) test was developed for polytomous items (Zwick, Donoghue, & Grima, 1993). Additionally, researchers expanded the Mantel-Haenszel and GMH tests to evaluate DIF across multiple groups (Fidalgo, 2011).

The primary strength of the MH test lies in its simplicity and ease of computation, making it accessible to many practitioners. However, as mentioned in Swaminathan and Rogers (1990), CTT based methods assume each item equally contributes to measuring the overall trait. This assumption can miss complexities like interactions between group variables and latent traits, limiting detection to uniform DIF. And its reliance on observed scores rather than latent traits limit its applicability in more nuanced scenarios. The MH test has seen widespread use in large-scale standardized testing programs, to ensure that test items do not unfairly advantage or disadvantage any group. Its straightforward application makes it a staple in the initial stages of DIF detection.

As psychometric research advanced, the limitations of CTT-based methods like the Mantel-Haenszel test became more apparent, particularly in their inability to model latent traits and detect non-uniform DIF. This led to the adoption of Item Response Theory (IRT), which provides a more sophisticated framework for analyzing item characteristics. Advancements in IRT have led to more precise and adaptable DIF detection methods. IRT models facilitate detailed item-level analyses across ability continua. (Bechger, Maris, Verstralen, & Béguin, 2003). These models also make it possible to detect non-uniform DIF, which involves an interaction between group membership and the latent trait level, by including discrimination parameters.

Early IRT-based methods predominantly used the goodness-of-fit indices of IRT models, such as chi-square statistic-based methods. Cohen, Kim, and Baker (1993) introduced a chi-square statistic-based method to detect DIF within the GRM framework, an IRT approach tailored for polytomous response items. This method compares differences in item parameters between groups, calculating a chi-square statistic to assess the significance of these differences. However, this oversimplified procedure does not yield good DIF detection power, as shown in later studies.

In a different vein, as the field of DIF detection continues to evolve, researchers have sought methods that can provide comprehensive insights while balancing model assumptions

and flexibility. One such method is the Differential Functioning of Items and Tests (DFIT) framework. Building on the foundational work, Roju, Van der Linden, and Fleer (1995) developed internal measures of DIF and Differential Test Functioning (DTF) that are applicable to both unidimensional and multidimensional testing scenarios. Their framework, which includes Compensatory DIF (CDIF) and Noncompensatory DIF (NCDIF) indices, allows for a nuanced analysis of DIF by not assuming that other items in the test are unbiased. This approach was validated through Monte Carlo simulations, demonstrating its effectiveness in identifying known DIF and providing a robust comparison against traditional methods like Lord's Chi-square test. Oshima, Raju, and Flowers (1997) extended these methodologies to tests designed to measure multiple dimensions simultaneously. Their study not only confirmed the capability of this extended framework to accurately detect DIF and DTF in a multidimensional context but also highlighted the importance of scaling and linking item parameters properly across different groups to ensure the fairness and validity of tests. The drawback of the DFIT study is that the linking has to be done with iterative linking based on matching test response functions, which is inefficient and likely to be confounded with DIF itself (C. Wang, Zhu, & Xu, 2023).

The likelihood ratio test (LRT) provides a more robust method as well as more flexibility. The IRT-LR-DIF method was developed to improve the detection of DIF by combining the robust modeling capabilities of IRT with the statistical rigor of likelihood ratio testing. IRT provides a flexible and powerful framework. The method can be applied within the context of any of the common IRT models, making it versatile and adaptable to various testing scenarios. After the item parameters are estimated, the IRT-LR-DIF method employs a likelihood ratio test to compare nested models — one model that assumes the item parameters are the same across groups (no DIF) and another model that allows different parameters for different groups (presence of DIF). The test statistic from the likelihood ratio test evaluates whether the inclusion of different item parameters significantly improves the model fit, indicating the presence of DIF. The likelihood ratio test provides a robust statistical basis for testing the presence of DIF, offering clear metrics and criteria for decision-making in test analysis.

After the general idea of the LRT-based method was proposed, many further studies

advanced the LRT-based method. The original IRT-LR-DIF method has been further developed and refined to handle more complex data structures and to improve accuracy by using sophisticated statistical techniques to ensure that the likelihood ratio test is appropriately calibrated and that the type I error rates (the probability of incorrectly declaring DIF when there is none) are controlled. Also, different algorithms have been studied. For example, the IRT-LR-DIF can start by assuming all items are DIF-free and then drop that constraint for each item iteratively, or it can start with assuming all items have DIF and different item parameters for different covariate groups and then add constraints to each item iteratively. The latter algorithm tends to have higher power than the former but requires known anchor items at the beginning to make the model identifiable(Woods, 2009). As it is intensively studied, the IRT-LR-DIF method has been integrated into many software packages, such as the mirt package in R (Chalmers, 2012), which facilitates its application and improves accessibility for researchers and practitioners.

The LRT-based method is flexible enough to detect DIF associated with polytomous response data and multi-dimensional latent traits, generating results with good detection power. It can also handle situations where no known anchor items are available, and the well-developed software packages make this method easy to implement. However, the biggest limitation of this method is that it still performs pairwise comparisons, making it challenging to apply when DIF is associated with multiple covariates or impossible when the number of covariates is large or there are continuous covariates.

A complementary area within IRT-based DIF detection focuses on logistic regression. Unlike previously reviewed IRT-based methods, the logistic regression models the probability of a correct response as a function of group membership and ability level, including interaction terms. This method involves estimating logistic regression equations to assess the impact of group membership on item responses while accounting for the underlying ability. By including group membership as variables, this method make it possible to detect different kind of covariates and multiple covariates. Swaminathan and Rogers (1990) introduce a logistic regression procedure that utilizes a 2PL IRT model to account for both the effects of group variables and the interaction between group membership and ability levels. Expanding on this framework, Zumbo (1999) enhances the method by adapting the logistic

regression model to handle ordered response categories, thus broadening the applicability of the DIF detection method to a wider range of testing scenarios and item types.

Further development in this area is presented in the work on the lordif R package by Choi, Gibbons, and Crane (2011), which integrates IRT and ordinal logistic regression in an iterative hybrid framework. The lordif R package incorporates an iterative purification algorithm, which refines the matching criterion by using group-specific item parameters for items identified with DIF when there is no anchors provided. This process allows for more accurate trait estimation and the re-evaluation of items for DIF across multiple iterations, enhancing the precision and reliability of the DIF detection process. The package provides a comprehensive platform for both detecting DIF and assessing the impact of identified DIF on the overall test scores.

Building on the flexibility of logistic regression methods, researchers developed the Moderated Nonlinear Factor Analysis (MNLFA). The MNLFA framework, introduced by Bauer and Hussong (2009), offers a broader and more flexible framework for measurement invariance that is particularly useful in integrative data analysis, which combines data from multiple studies. MNLFA extends the capabilities of traditional item response theory by utilizing a 2PL IRT model to accommodate variations across different study samples and items. Similar to logistic regression methods, this model integrates covariates that can influence item functioning, enabling the testing of the effects of multiple continuous and categorical covariates simultaneously on the measurement model. Furthermore, the MNLFA model allows the factor mean and variance to differ across levels of the exogenous moderators, offering a more robust framework for understanding the nuances of complex covariate effects. MNLFA offers a powerful approach for capturing complex DIF patterns and provides a solid foundation for regularization DIF detection methods, which will be reviewed in the next section.

Similar to MNLFA, which extends the traditional factor analysis model to handle multiple groups and interactions, another branch of study in DIF involves the assessment of measurement invariance within Structural Equation Modeling (SEM) using the multiple-indicator multiple-cause (MIMIC) model. The MIMIC model was first introduced by Jöreskog and Goldberger (1975). This model addresses the need for a method capable of si-

multaneously handling measurement errors in observed variables and the effects of external variables on latent constructs. The structure of a MIMIC model comprises a measurement part that links latent variables to their indicators and a structural part that specifies how exogenous variables directly affect both the latent variables and observed indicators. The model's approach to modifying direct paths to items from background variables linked to group membership enables a nuanced analysis of how these variables influence item properties and factor means (Finch, 2005). However, as discussed in Woods (2009), one potential limitation of MIMIC methods is their inefficacy in effectively testing for non-uniform DIF. Like MNLFA, the MIMIC framework itself is not a DIF detection method but provides a solid foundation for many emerging advanced DIF detection methods.

One recent such study is by W. Wang, Liu, and Liu (2022), who propose a novel framework for quantifying DIF through item-specific residuals derived from a regression model tailored to true item parameters. This approach eliminates the need for predefined anchor items, setting a versatile foundation to define DIF effect sizes. The methodology extends to include both analytical and numerical techniques to establish the null distribution of the test statistic, thereby facilitating robust statistical inference even amidst a mixture of DIF and non-DIF items. This method represents a pivotal improvement in ensuring the accuracy and reliability of DIF detection across varying testing scenarios.

The IRT and SEM based methods discussed so far rely on specific model assumptions that may not hold in all contexts. are computationally demanding due to the complexity of item parameter estimation, and they require large sample sizes and careful model specification to avoid misidentification of DIF. These complexity can be a barrier for some practitioners.

While the IRT and SEM based methods discussed so far have provided valuable insights and tools for DIF detection, they often rely on specific model assumptions that may not hold in all contexts. In addition, most of those methods are computationally demanding due to the complexity of item parameter estimation. These complexity can be a barrier for some practitioners. Nonparametric methods offer an alternative by making fewer assumptions about the underlying data distribution.

Developed by Shealy and Stout (1993), SIBTEST is a robust non-parametric method

that compares subgroup performances on test items matched by a reference variable related to the overall test score. This means it does not require the estimation of item response functions, making it applicable even when traditional IRT assumptions do not hold. The method provides a mechanism to both test for significance and estimate the amount of DIF present in test items. It does this through a comparison process that involves matching test takers based on their ability levels and then statistically examining the differences in their responses to specific items. A key aspect of SIBTEST is its use of a regression correction technique to control for type I error inflation, which can occur due to differences in ability distributions between groups being compared. This correction is crucial for maintaining the accuracy of DIF detection. SIBTEST is known for its robustness in various testing conditions and its flexibility to handle different kinds of data. The method has been shown to perform well in comparison with other popular DIF detection procedures and is capable of being adapted to complex test structures, including those involving polytomous items. While effective, Gierl, Gotzmann, and Boughton (2004) studied on the effectiveness and limitations of SIBTEST under conditions with a large percentage of DIF items and varying sample sizes, particularly highlighting issues with unbalanced DIF and the influence of sample size on detection rates. SIBTEST's performance is also sensitive to the choice of matching criterion and the proportion of DIF items within the test, and the matching process can make the method computational inefficient, especially when the dimension of latent trait is high.

Another notable emerging non-parametric method is the one proposed by Yuan, Liu, and Han (2021). They introduced a novel method called RCD-DIF (Relative Change of Difficulty Difference) for detecting DIF at both the test and item levels. This method involves a visual examination procedure based on two types of quantile-quantile (QQ) plots: the D-QQ plot and a graphical test through RCD-QQ plots. These plots are designed to visually identify DIF by comparing observed differences against simulated DIF-free data. A formal graphical test procedure is also provided for item-level DIF detection.

The RCD-DIF method identifies the model based on the principle that the relative change in difficulty (RCD) of an item does not depend on the mean ability of individual groups, thereby cleverly bypassing the model indetermination problem. Additionally, the

RCD-DIF method offers several advantages over traditional methods by providing a more integrated approach to identifying DIF items, controlling Type I error at both the item and test levels, and handling cases where the majority of items exhibit DIF, which might challenge traditional methods due to masking and swamping effects. However, the RCD-DIF method still requires pairwise analysis when there are multiple groups. Similar to LRT-based methods, the computational efficiency of this method may decrease when the number of groups is large, and continuous covariates cannot be well handled.

This literature review highlights the evolution of DIF detection methods from early CTT-based approaches like the Mantel-Haenszel test to advanced IRT-based methods like DFIT, LRT-LR-DIF, and logistic regression models. It also covers nonparametric methods like SIBTEST and RCD-DIF. It also introduced two important frameworks for DIF detection, the MNLFA and the MIMIC model, and introduced advanced recent studies base on those works like the robust regression method. Each method has its strengths and limitations, and a summary is provided in Table 1.1. Note that a "-" in the table means not applicable because the CTT model does not calculate the latent trait. "Yes" in the table indicates that the method can handle or can be generalized to handle the challenge, regardless of whether researchers have already expanded the method to address the challenge.

These advancements illustrate the ongoing development of DIF detection methods, transitioning from traditional approaches to more complex statistical models. This progression reflects the increasing complexity of testing scenarios and the various applications of these methods.

Method	**Challenges**				
	Non-uniform DIF	Multiple Covariates	Multidimensional Trait	Polytomous Response	No Anchor Items
Generalized Mantel–Haenszel Methods	No	No	-	Yes	Yes
Chi-square Statistic Methods	Yes	No	Yes	Yes	No
DFIT	Yes	No	Yes	Yes	Yes
IRT-LR-DIF	Yes	No	Yes	Yes	Yes
Logistic Regression Methods	Yes	Yes	Yes	Yes	Yes
Robust Regression Method	No	Yes	Yes	No	Yes
SIBTEST	Yes	No	No	Yes	Yes
RCD-DIF	Yes	No	No	No	Yes

Table 1.1: Comparison of Methods

1.1.2 Regularization Methods for Enhanced DIF Detection

Other recent research has explored regularization methods as a promising solution to detect DIF in various model settings. These approaches leverage modern machine learning techniques to enhance the accuracy and efficiency of DIF detection, particularly in complex model configurations where traditional methods may struggle. Regularization techniques, such as Lasso, introduce penalties that simplify model estimation and help in identifying the most relevant variables for DIF detection. This section delves into how these innovative methods are being applied to improve DIF analysis, offering a more nuanced understanding of measurement biases across diverse testing environments.

With the advancement of machine learning, psychometricians have begun to integrate regularization methods into measurement models and DIF detection since the 2010s. Several studies have adopted L_1 regularization across different models and data types. Magis, Tuerlinckx, and De Boeck (2015) proposes the LR lasso DIF method for detecting DIF among dichotomously scored items using logistic regression with a lasso penalty. The LR lasso DIF method applies a logistic regression model to all items simultaneously, incorporating item-specific intercepts, an effect of the sum score, and item-group interaction effects with a lasso penalty applied to all DIF parameters. The method was tested against traditional approaches like the Mantel-Haenszel method and another logistic regression method through a simulation study focusing on small sample performance, where it demonstrated superior performance in settings with item impact while performing comparably in larger samples.

Schauberger and Mair (2020) explores an innovative regularization method using the lasso principle to detect uniform DIF in item response models, particularly for polytomous items. This approach is suitable for a variety of models, including the generalized partial credit model, allowing for the simultaneous consideration of multiple covariates of different scale levels. The method uses a penalized likelihood approach that identifies DIF effects and provides trait estimates that correct for these effects across different covariates.

Additionally, Bauer, Belzak, and Cole (2020) presents a comprehensive study on the detection of DIF using regularization approaches, specifically focusing on intercept DIF

versus loading DIF parameters. The study explores the performance of regularization DIF (Reg-DIF) in comparison with the IRT-LR-DIF method across various simulation settings. Key findings include that Reg-DIF maintains better control of false positives compared to IRT-LR-DIF, especially when both retention and significance of DIF parameters are required. However, this control comes at the cost of reduced power, particularly when DIF is large in magnitude but not pervasive across items. The study suggests that a larger sample size is needed to adequately detect small-magnitude DIF. The use of the Bayesian Information Criterion (BIC) for tuning Reg-DIF is recommended over the Akaike Information Criterion (AIC) in most scenarios, except in smaller samples where small DIF effects are present, where AIC may be preferable. Overall, the study provides valuable insights into the strengths and limitations of using Reg-DIF and IRT-LR-DIF for DIF detection, highlighting the importance of sample size and DIF magnitude in the effectiveness of these methods.

Wallin, Chen, and Moustaki (2024) proposes a statistical model that incorporates latent classes to handle unknown groups and introduces item-specific DIF parameters to identify DIF effects. They utilize an L1-regularized estimator to simultaneously determine latent classes and detect DIF items, employing a computationally efficient Expectation-Maximization (EM) algorithm to manage the non-smooth optimization challenges of the regularization process. All these methods address the limitations of traditional DIF detection methods, which often rely on an item-by-item analysis with assumptions of DIF-free anchor items. Additionally, they overcome the limitations of traditional DIF detection tools that typically handle only one covariate at a time, providing a more flexible and comprehensive tool for researchers.

Furthermore, beyond the L_1 penalty lasso, other penalty types have been explored. Tutz and Schauberger (2015) introduces a generalized model for detecting DIF and a group lasso penalty (Yuan & Lin, 2006) for detecting DIF that incorporates covariates characterizing the test taker, such as gender, race, or age, while C. Wang et al. (2023) compares the effectiveness of lasso and adaptive lasso methods (Zou, 2006) against the classical likelihood ratio test, highlighting the advantages of regularization methods. Belzak and Bauer (2024) discuss advanced statistical techniques, including lasso and minimax concave penalty

(MCP; Breheny & Huang, 2011; Friedman et al., 2010; Zhang, 2010) to adjust for multiple covariates simultaneously, enhancing the detection of bias in item responses. Together, these works provide the possibility of choosing different regularization methods other than lasso for different testing purposes.

The studies mentioned above illustrate the efficacy of regularization methods in accurately detecting DIF items while managing multiple parameters simultaneously. However, some studies stand out not only for these capabilities but also for providing robust statistical inferences. Unlike approaches reliant on pre-specified anchor items, regularization methods often yield underestimated standard errors and incorrect p-values (Chen, Bauer, Belzak, & Brandt, 2022). For instance, Chen et al. (2022) scrutinizes the use of lasso regularization in DIF assessment, noting its effectiveness in simultaneous DIF detection across all items yet its tendency to produce underestimated standard errors and erroneous p-values. In contrast, Bayesian regularization with spike and slab priors offers a more theoretically sound approach, employing an inclusion probability criterion for selecting and inferring DIF parameters over empirical criteria like credible intervals.

In summary, regularization methods provide a powerful toolkit for DIF detection, particularly in modern testing scenarios that involve complex model structures and various data forms. Increasingly, research is emerging that leverages the advantages of regularization methods to address a wide range of complicated DIF detection scenarios. Researchers are making significant efforts to develop different models and parameterizations that are suitable for regularization techniques. Some have also dedicated effort to developing and adapting their estimation algorithms to improve the accuracy and efficiency of parameter estimation for regularization-based DIF methods. Meanwhile, others are interested in drawing valid inferences from the results of regularization methods. Overall, these research efforts together ensure that regularization in DIF detection not only addresses the limitations of traditional approaches but also leads to a more robust, efficient, and nuanced analysis in measurement invariance.

1.2 Structure of the Study

The existing literature indicates a growing interest in regularized methods for detecting DIF, yet a noticeable gap persists in the field regarding adaptable approaches suited to diverse testing scenarios. Recognizing this need for innovation, this disseration introduces a comprehensive framework aimed at enhancing DIF detection across multiple covariates, encompassing both continuous and categorical variables, and accommodating multi-dimensional polytomous response data. Moreover, this study emphasizes the importance of investigating not only the detection of DIF but also its impact. DIF detection involves identifying differences in item parameters among different examinee groups, while understanding its impact allows us to ascertain the true differences between individuals from these groups. While most studies focus on DIF detection, this research endeavors to bridge this gap by presenting two novel methods for estimating the impact of covariates on multivariate latent traits. These methods greatly expand established models such as the MNLFA model (Bauer & Hussong, 2009), employing techniques like Cholesky decomposition to ensure valid covariance matrices, and the covariance regression model (Hoff & Niu, 2012), which addresses complexities associated with high-dimensional latent trait spaces and a large number of covariates. By enhancing the model's capacity to accurately capture covariate effects, our methods offer promising avenues for advancing DIF research.

The dissertation is structured as follows: First, I introduce the framework of a multidimensional Graded Response Model with DIF effects. Then, I outline two innovative parameterization approaches for modeling the impact of covariates on latent traits. Next, I describe the EM (Expectation-Maximization) algorithm used for efficient estimation of item, DIF, and impact parameters. Subsequently, I rigorously test the proposed methods through simulation studies, focusing on both uniform and non-uniform DIF detection within a two-dimensional GRM framework. Additionally, I validate the feasibility of the proposed approach by analyzing real data from the Patient-Reported Outcomes Measurement Information System (PROMIS) dataset.

Chapter 2

MODELS

In this chapter, I introduce the Graded Response Model with DIF effect, focusing specifically on the effects of covariates on item parameters and the latent trait distribution. The chapter begins by presenting the multi-dimensional GRM with DIF. Following that, two innovative parameterization approaches are employed to model the impact of covariates on the latent trait distribution. These approaches are designed to enhance the accuracy and efficiency of the impact recovery algorithms, ultimately contributing to a better understanding of the extent to which latent traits genuinely differ among different covariate groups, independent of the existing DIF effects.

2.1 Covariates Effect on Item Parameters

Let N represent the total number of examinees, J denote the test length, K indicate the number of trait dimensions, P denote the number of covariates, and G denote the number of response categories. The response matrix, denoted as $\boldsymbol{U}$, is an N-by-J matrix given by:

$$\boldsymbol{U} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1j} & \dots & u_{1J} \\ u_{21} & u_{22} & \dots & u_{2j} & \dots & u_{2J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{i1} & u_{i2} & \dots & u_{ij} & \dots & u_{iJ} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & \dots & u_{Nj} & \dots & u_{NJ} \end{pmatrix} (i = 1, ..., N; j = 1, ..., J).$$

A dummy variable, denoted as Y_{ijg}, is used to indicate whether examinee i receives a score g for item j:

$$Y_{ijg} = \begin{cases} 1, \text{ if } u_{ij} = g \\ 0, \text{ if } u_{ij} \neq g \end{cases} \quad (g = 1, ..., G) \tag{2.1}$$

For a polytomously scored item j, the probability that person i with a latent trait vector $\boldsymbol{\theta}_i$ scores g or higher on item j is given by

$$P^*_{jg}(\boldsymbol{\theta}_i) = \frac{1}{1 + e^{-(\boldsymbol{a}_j^T \boldsymbol{\theta}_i + d_{jg} + \boldsymbol{X}_i \boldsymbol{\gamma}_j \boldsymbol{\theta}_i + \boldsymbol{X}_i \boldsymbol{\beta}_{jg})}} \quad (g = 1, ..., G-1). \tag{2.2}$$

When $g = 0$,

$$P^*_{j0}(\boldsymbol{\theta}_i) = 1 \tag{2.3}$$

When $g = G$,

$$P^*_{jG}(\boldsymbol{\theta}_i) = 0. \tag{2.4}$$

The probability that person i with a latent trait vector $\boldsymbol{\theta}_i$ scores exactly g on item j can be obtained by taking the difference between $P^*_{j(g-1)}(\boldsymbol{\theta}_i)$ and $P^*_{jg}(\boldsymbol{\theta}_i)$:

$$P_{jg}(\boldsymbol{\theta}_i) = P^*_{j(g-1)}(\boldsymbol{\theta}_i) - P^*_{jg}(\boldsymbol{\theta}_i) \quad (g = 1, ..., G). \tag{2.5}$$

In Equation 2.2, $\boldsymbol{a}_j$ is a K-by-1 vector of item discriminations, d_{jg} is a scalar of intercept for item J and response category g. $\boldsymbol{\theta}_i$ is a K-by-1 vector of latent trait for person i. In addition, $\boldsymbol{X}_i$ is a 1-by-P vector including all the grouping information related to DIF. The matrix $\boldsymbol{\gamma}_j$ is a P-by-K matrix of coefficients that indicates the DIF effect on item discrimination. The vector $\boldsymbol{\beta}_{jg}$ has dimensions P-by-1 and contains coefficients that signify the DIF effect on item difficulty. When $\boldsymbol{\gamma}_j = \mathbf{0}$, it implies that item j exhibits non-uniform DIF. If $\boldsymbol{\gamma}_j = \mathbf{0}$, but $\boldsymbol{\beta}_{jg} \neq \mathbf{0}$, it suggests that item j has uniform DIF. If both $\boldsymbol{\beta}_{jg}$ and $\boldsymbol{\gamma}_j$ are equal to $\mathbf{0}$, it indicates that item j has no DIF.

2.2 Covariates Effect on Latent Trait Distribution

The concept of "impact" refers to the genuine differences in latent traits observed among various covariate groups, regardless of the presence of DIF effects. To address this, two novel parameterization approaches are utilized to effectively model how covariates influence the distribution of latent traits. In both approaches, I assume that the latent trait, $\boldsymbol{\theta}_i$, follows a conditional multivariate normal distribution, specifically $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$. In Approach I, the influence of covariates on both the mean vector and covariance matrix is denoted as $\boldsymbol{\kappa}$ and $\boldsymbol{\omega}$, respectively. In Approach II, the impact of covariates on the mean vector and covariance matrix is symbolized by $\boldsymbol{\kappa}$ and $\boldsymbol{\phi}$, respectively.

In the first parameterization (Approach I), I expand the moderated nonlinear factor analysis (MNLFA) model introduced by Bauer and Hussong (2009). In Bauer and Hussong (2009), only a one-dimensional latent trait was considered, so they estimated one mean impact parameter and one variance impact parameter for each covariate. In our multi-dimensional case, I estimate a vector for the mean impact and a matrix for the variance-covariance impact. The covariate effect could be positive or negative. To ensure that the covariate-moderated covariance matrix is valid, a Cholesky decomposition is applied to the covariance matrix, breaking it down into the product of a lower triangular matrix and its conjugate transpose. The covariate effect is then incorporated into the lower triangular matrix. Expressed in equation form, the relationship between the covariates and the mean vector is as follows:

$$\boldsymbol{\mu}_i = \boldsymbol{\kappa}^T \boldsymbol{X}_{\boldsymbol{i}}^T, \tag{2.6}$$

$$\boldsymbol{L_i} = \boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i\omega})), \tag{2.7}$$

$$\boldsymbol{\Sigma_i} = \boldsymbol{L_i L_i^T}, \tag{2.8}$$

where $\boldsymbol{\Sigma_0} = \boldsymbol{L_0 L_0^T}$ is the covariance matrix for $\boldsymbol{X_i}$ at reference level. $\boldsymbol{\kappa}$ is a P-by-K matrix. $\boldsymbol{\omega}$ is a $P \times \frac{K(K+1)}{2}$ matrix. $\boldsymbol{X_i\omega}$ is a 1-by-$\frac{K(K+1)}{2}$ matrix and I reshape $\exp(\boldsymbol{X_i\omega})$ into a lower

triangular matrix. $\circ$ means entrywise product. While Bauer (2017) models the covariates effects on each element of $\boldsymbol{\Sigma_i}$ separately, which may result in non-positive-definite cases, I perform a Cholesky decomposition of $\boldsymbol{\Sigma_i}$ in Equation 2.8.

It is important to note that for the reference group, $\boldsymbol{X}_i = 0$, which automatically fixed the center of the scale. To identify the model, it is necessary to further set all diagonal elements of $\boldsymbol{\Sigma_0}$ to 1. The correlations between θs can be freely estimated.

In the second parameterization (Approach II), I maintain the same $\boldsymbol{\kappa}$ as in the first approach. However, I adopt a different parameterization for $\boldsymbol{\omega}$. Given that both the latent trait dimension K and the number of covariates P can be large, $\boldsymbol{\omega}$ becomes a high-dimensional matrix, which can pose challenges for optimization. Instead, I employ the following formulation:

$$\boldsymbol{\Sigma_i} = \boldsymbol{\Sigma_0} + \boldsymbol{\phi}^T \boldsymbol{X}_i^{*T} \boldsymbol{X}_i^* \boldsymbol{\phi}. \tag{2.9}$$

Here, $\boldsymbol{\phi}$ represents a $(P+1)$-by-K matrix of regression coefficients that captures the effect of $\boldsymbol{X_i}$ on $\boldsymbol{\Sigma_i}$. To ensure flexibility and prevent the assumption that the variance is smallest for the reference group, $\boldsymbol{X}_i^* = (1, \boldsymbol{X}_i)$ includes an intercept term. This approach is particularly suitable when dealing with a large latent trait dimension K.

Chapter 3

ESTIMATION

In this chapter, the equations and algorithms utilized to estimate the item and trait distribution parameters introduced in Chapter 2 are presented. As the latent variable $\boldsymbol{\theta}$ is not directly observable, the EM algorithm is employed for model estimation.

Let $\boldsymbol{\Delta}$ represent the set of model parameters, including item parameters ($\boldsymbol{a}$, $\boldsymbol{d}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$) and latent trait distribution parameters ($\boldsymbol{\kappa}$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\omega}$). The marginal likelihood, given the response indicators $\boldsymbol{Y}$ and grouping information $\boldsymbol{X}$, is defined as:

$$L(\boldsymbol{\Delta}) = \prod_{i=1}^{N} \int \prod_{j=1}^{J} L(\boldsymbol{a}_j, \boldsymbol{d}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\theta}_i) f(\boldsymbol{\kappa}, \boldsymbol{\omega} \mid \boldsymbol{X}_i, \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_0) \partial \boldsymbol{\theta}_i, \tag{3.1}$$

where

$$L(\boldsymbol{a}_j, \boldsymbol{d}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\theta}_i) = \prod_{g}^{G} P_{jg}(\boldsymbol{\theta}_i)^{Y_{ijg}}$$

represents the likelihood of item parameters, and $P_{jg}(\boldsymbol{\theta}_i)$ is defined in equation 2.5, and

$$f(\boldsymbol{\kappa}, \boldsymbol{\omega} \mid \boldsymbol{X}_i, \boldsymbol{\theta}_i, \boldsymbol{\Sigma}_0) = (2\pi)^{-\frac{K}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} e^{-0.5(\boldsymbol{\theta}_i - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)}$$

represents the density function of $\boldsymbol{\theta}_i$.

To estimate the DIF parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, a l_1 regularized estimator is obtained by maximizing the following objective function:

$$\log L(\boldsymbol{\Delta}) - \eta(||\boldsymbol{\beta}||_1 + ||\boldsymbol{\gamma}||_1), \tag{3.2}$$

where

$$||\boldsymbol{\beta}||_1 = \sum_{j}^{J} \sum_{p}^{P} \sum_{g}^{G} |\beta_{jpg}|, ||\boldsymbol{\gamma}||_1 = \sum_{j}^{J} \sum_{p}^{P} \sum_{k}^{K} |\gamma_{jpk}| \mathbf{1}_{a_{jk} \neq 0},$$

and η is the regularization tuning parameter. The parameter estimates are obtained using EM algorithms. First, I will explore the general idea of the EM algorithm, followed by a detailed explanation of two different EM algorithms in section 3.3.

3.1 Expectation Step

In the E-step of the EM algorithm, I compute the conditional expectation of the complete data log-likelihood with respect to $\boldsymbol{\theta}$. At the $(t+1)$th iteration of the EM cycle, the expression for $Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})$ is given by:

$$\begin{aligned}
&Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) = E_{h(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{u},\boldsymbol{\Delta}^{(t)})}(\log(L(\boldsymbol{\Delta}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta})))\\
&= \sum_{i}^{N}\Bigg[\int l(\boldsymbol{a},\boldsymbol{d},\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{\theta}_i)h(\boldsymbol{\theta}_i|\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{\Delta}^{(t)})\partial\boldsymbol{\theta}_i\\
&+ \int \log f(\boldsymbol{\kappa},\boldsymbol{\omega}\mid\boldsymbol{\theta}_i,\boldsymbol{X}_i,\boldsymbol{\Sigma}_0)h(\boldsymbol{\theta}\mid\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{\Delta}^{(t)})\partial\boldsymbol{\theta}_i\Bigg]\\
&= \sum_{i}^{N}\sum_{j}^{J}\Bigg[\int l(\boldsymbol{a}_j,\boldsymbol{d}_j,\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j|\boldsymbol{X}_i,\boldsymbol{Y}_{ij},\boldsymbol{\theta}_i)h(\boldsymbol{\theta}_i|\boldsymbol{X}_i,\boldsymbol{Y}_{ij},\boldsymbol{\Delta}_j^{(t)})\partial\boldsymbol{\theta}_i\Bigg]\\
&+ \sum_{i}^{N}\Bigg[\int \log f(\boldsymbol{\kappa},\boldsymbol{\omega}\mid\boldsymbol{\theta}_i,X_i,\boldsymbol{\Sigma}_0)h(\boldsymbol{\theta}_i\mid\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{\Delta}^{(t)})\partial\boldsymbol{\theta}_i\Bigg],
\end{aligned} \tag{3.3}$$

where

$$h(\boldsymbol{\theta}_i\mid\boldsymbol{X}_i,\boldsymbol{Y}_{ij},\boldsymbol{\Delta}_j^{(t)}) = \frac{L(\boldsymbol{a}_j^{(t)},\boldsymbol{d}_j^{(t)},\boldsymbol{\beta}_j^{(t)},\boldsymbol{\gamma}_j^{(t)}|\boldsymbol{X}_i,\boldsymbol{Y}_{ij},\boldsymbol{\theta}_i)f(\boldsymbol{\kappa}^{(t)},\boldsymbol{\omega}^{(t)}\mid\boldsymbol{\theta}_i,X_i,\boldsymbol{\Sigma}_0)}{\int L(\boldsymbol{a}_j^{(t)},\boldsymbol{d}_j^{(t)},\boldsymbol{\beta}_j^{(t)},\boldsymbol{\gamma}_j^{(t)}|\boldsymbol{X}_i,\boldsymbol{Y}_{ij},\boldsymbol{\theta}_i)f(\boldsymbol{\kappa}^{(t)},\boldsymbol{\omega}^{(t)}\mid\boldsymbol{\theta}_i,X_i,\boldsymbol{\Sigma}_0)\partial\boldsymbol{\theta}_i} \tag{3.4}$$

denotes the posterior density of $\boldsymbol{\theta}_i$ given the current estimates of $\boldsymbol{a}_j,\boldsymbol{d}_j,\boldsymbol{\beta}_j,\boldsymbol{\gamma}_j,\boldsymbol{\kappa}$, and $\boldsymbol{\omega}$ at iteration t.

To approximate the integral, I employ the Gauss-Hermite quadrature method. I define M_0 as the number of equidistant points sampled from the interval $[h_\theta, l_\theta]$ along each coordinate dimension, resulting in $M = (M_0)^K$ total quadrature samples. Each sample $\boldsymbol{q}_m$ $(m = 1,...,M)$ corresponds to a K-dimensional vector. Equation 3.3 can then be approximated as:

$$
\begin{aligned}
Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) &= \sum_{i}^{N}\sum_{j}^{J}\sum_{m}^{M} l(\boldsymbol{a}_j, \boldsymbol{d}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{q}_m) h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)}) \\
&+ \sum_{i}^{N}\sum_{m}^{M} \log f(\boldsymbol{\kappa}, \boldsymbol{\omega} \mid \boldsymbol{q}_m, X_i, \boldsymbol{\Sigma}_0) h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}),
\end{aligned} \tag{3.5}
$$

where

$$
h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)}) = \frac{L(\boldsymbol{a}_j, \boldsymbol{d}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{q}_m) f(\boldsymbol{\kappa}, \boldsymbol{\omega} \mid \boldsymbol{q}_m, X_i, \boldsymbol{\Sigma}_0)}{\sum_m^M L(\boldsymbol{a}_j, \boldsymbol{d}_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{q}_m) f(\boldsymbol{\kappa}, \boldsymbol{\omega} \mid \boldsymbol{q}_m, X_i, \boldsymbol{\Sigma}_0)}. \tag{3.6}
$$

The objective function in 3.2 becomes:

$$
Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \eta(||\boldsymbol{\beta}||_1 + ||\boldsymbol{\gamma}||_1). \tag{3.7}
$$

3.2 Maximization step

In the Maximization step of the EM algorithm, I update the parameter estimates $\hat{\boldsymbol{\Delta}}$ based on equation 3.5 obtained from the Expectation step.

3.2.1 Trait Distribution Estimation

In the M-step, the first task involves estimating the group effect on the mean vector and covariance matrix, as well as the correlation between latent traits (represented by the off-diagonal elements in $\boldsymbol{\Sigma}_0$). Three parameters, namely $\boldsymbol{\kappa}$, $\boldsymbol{\omega}$, and $\boldsymbol{\Sigma}_0$, need to be estimated.

To determine the optimal value of $\boldsymbol{\kappa}$, the partial derivative of Equation 3.5 with respect to $\boldsymbol{\kappa}$ is taken and set equal to zero. This results in the following equation:

$$
\begin{aligned}
\frac{\partial Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial \boldsymbol{\kappa}} &= \sum_{i}^{N}\sum_{m}^{M} h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) \frac{\partial -\frac{1}{2}(\boldsymbol{q}_m - \boldsymbol{\kappa}^T\boldsymbol{X}_i^T)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{q}_m - \boldsymbol{\kappa}^T\boldsymbol{X}_i^T)}{\partial \boldsymbol{\kappa}} \\
&= \sum_{i}^{N}\sum_{m}^{M} h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) \boldsymbol{X}_i^T(\boldsymbol{q}_m - \boldsymbol{\kappa}^T\boldsymbol{X}_i^T)^T\boldsymbol{\Sigma}_i^{-1} = 0.
\end{aligned} \tag{3.8}
$$

I have

$$\sum_i^N \sum_m^M h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) \boldsymbol{X}_i^T \boldsymbol{q}_m^T \boldsymbol{\Sigma}_i^{-1} = \sum_i^N \sum_m^M h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) \boldsymbol{X}_i^T \boldsymbol{X}_i \boldsymbol{\kappa} \boldsymbol{\Sigma}_i^{-1}. \tag{3.9}$$

This suggests that there is no closed-form solution for $\boldsymbol{\kappa}$. Then I try to find the estimate of the correlation matrix $\boldsymbol{\Sigma_0}$. Remember in the first approach, I use the parameterization $\boldsymbol{\Sigma_i} = \boldsymbol{L_i} \boldsymbol{L_i}^T$ where $\boldsymbol{L_i} = \boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i \omega}))$. So when estimating $\boldsymbol{\Sigma_0}$, I calculate the partial derivative with respect to $\boldsymbol{L_0}$ instead of $\boldsymbol{\Sigma_0}$ as following:

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\Delta} | \boldsymbol{\Delta}^{(t)})}{\partial \boldsymbol{L_0}} &= \sum_i^N \sum_m^M h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) \frac{\partial -\frac{1}{2} \log |\boldsymbol{\Sigma_i}| - \frac{1}{2} (\boldsymbol{q}_m - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{q}_m - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)}{\partial \boldsymbol{L_0}} \\ &= \sum_i^N \sum_m^M h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) \frac{\partial -\frac{1}{2} \log |\boldsymbol{\Sigma_i}| - \frac{1}{2} (\boldsymbol{q}_m - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{q}_m - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)}{\partial \boldsymbol{\Sigma_i}} \frac{\partial \boldsymbol{\Sigma_i}}{\partial \boldsymbol{L_0}}, \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} \frac{\partial \boldsymbol{\Sigma_i}}{\partial \boldsymbol{L_0}} &= \frac{\partial (\boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i \omega})))(\boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i \omega})))'}{\partial \boldsymbol{L_0}} \\ &= \text{reshape}(\exp(\boldsymbol{X_i \omega}))(\boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i \omega})))' \\ &+ \text{reshape}(\exp(\boldsymbol{X_i \omega}))'(\boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i \omega}))). \end{aligned} \tag{3.11}$$

Set the partial derivative to 0:

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\Delta} | \boldsymbol{\Delta}^{(t)})}{\partial \boldsymbol{L_0}} = \sum_i^N \sum_m^M & h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) (-\frac{1}{2} \boldsymbol{\Sigma}_i^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{q}_m - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)(\boldsymbol{q}_m - \boldsymbol{\kappa}^T \boldsymbol{X}_i^T)^T \boldsymbol{\Sigma}_i^{-1}) \\ & [\text{reshape}(\exp(\boldsymbol{X_i \omega}))(\boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i \omega})))^T \\ & + \text{reshape}(\exp(\boldsymbol{X_i \omega}))^T (\boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i \omega})))] \\ & = 0 \end{aligned} \tag{3.12}$$

It is observed that the parameter $\boldsymbol{L_0}$ does not possess a closed-form solution.

The same issue is encountered with the other parameter related to $\boldsymbol{\Sigma}_i$, the impact on covariance matrix, $\boldsymbol{\omega}$. Consequently, the approximation of the values of $\boldsymbol{\kappa}$, $\boldsymbol{\omega}$, and $\boldsymbol{\Sigma}_0$ is carried out using the gradient descent method, employing the optimization function "optim" in R.

Specifically, the value of $\boldsymbol{\kappa}$ is updated by utilizing the *optim* function with the objective function described in Equation 3.5. Subsequently, $\boldsymbol{L_0}$ undergoes optimization with the same log-likelihood function through the *optim* function, and the reference level covariance matrix $\hat{\boldsymbol{\Sigma}}_0$ is rescaled to ensure unit variances along the diagonal. This involves computing the square root of the diagonal elements of $\hat{\boldsymbol{\Sigma}}_0$ to derive standard deviations for each variable, followed by dividing each off-diagonal element by the corresponding standard deviation product. Finally, the *optim* function is employed to update $\boldsymbol{\omega}$ using the same objective function previously utilized for updating $\boldsymbol{\kappa}$ and $\boldsymbol{L_0}$. Due to the summations involved in the objective function, the algorithm may experience sluggishness. Various algorithms were tested in an attempt to expedite the estimation process. However, it was determined that the most efficient approach still necessitates iterating over the sample size N. For a comprehensive understanding of the computing algorithms employed, please refer to Section 3.5.

Due to lack of computing efficiency in the first approach, a different parameterization approach is considered, which results in more efficient computing. In the second parameterization approach, the Cholesky decomposition of $\boldsymbol{\Sigma}$ is avoided. Instead, $\boldsymbol{\Sigma_i} = \boldsymbol{\Sigma_0} + \boldsymbol{\phi}^T \boldsymbol{X}_i^{*T} \boldsymbol{X}_i^* \boldsymbol{\phi}$, where $\boldsymbol{X}_i^* = (1, \boldsymbol{X}_i)$, and $\boldsymbol{\mu}_i = \boldsymbol{\kappa}^{*T} \boldsymbol{X_i^{*T}}$ is similar to the first approach, but a column of K 0s is added to $\boldsymbol{\kappa^T}$ to form $\boldsymbol{\kappa}^{*T} = (\mathbf{0}_K, \boldsymbol{\kappa^T})$. Similar to $\boldsymbol{\phi}$, $\boldsymbol{\kappa^*}$ is a $(P+1)$-by-K matrix. Closed-form solutions of $\boldsymbol{\kappa}$, $\boldsymbol{\phi}$, and $\boldsymbol{\Sigma}_0$ in the M-step are obtained using this parameterization, as stated by Hoff and Niu (2012).

According to Hoff and Niu (2012), the following expressions are derived for maximizing the expected log-likelihood:

$$v_i = (1 + \boldsymbol{X}_i^* \boldsymbol{\phi} \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\phi}^T \boldsymbol{X}_i^{*T})^{-1}, \tag{3.13}$$

$$m_i = \sum_{m}^{M} h(\boldsymbol{q}_m \mid \boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\Delta}^{(t)}) v_i (\boldsymbol{q}_m - \boldsymbol{\kappa}^{*T} \boldsymbol{X}_{\boldsymbol{i}}^T)^T \boldsymbol{\Sigma}_0 \boldsymbol{\phi}^T \boldsymbol{X}_i^{*T}, \tag{3.14}$$

and

$$s_i = v_i^{1/2}. \tag{3.15}$$

To maximize the expected log-likelihood, a $2N$-by-$2(P+1)$ matrix $\tilde{\boldsymbol{X}}$ is constructed, with the ith row as $(\boldsymbol{X_i}, m_i \boldsymbol{X_i})$ and the $(N+i)$th row as $(\boldsymbol{0}_P^T, s_i \boldsymbol{X_i})$. Additionally, $\tilde{\boldsymbol{Y}}$ is a $2N$-by-K matrix, with the ith row as $\boldsymbol{q}_{\boldsymbol{m}}^T$ and the $(N+i)$th row as $\boldsymbol{0}_K^T$.

In each M-step, the estimates of ($\boldsymbol{\kappa}^*$, $\boldsymbol{\phi}$, $\boldsymbol{\Sigma}_0$), represented as ($\hat{\boldsymbol{\kappa}}^*$, $\hat{\boldsymbol{\phi}}$, and $\hat{\boldsymbol{\Sigma}}_0$), are updated as follows:

$$(\hat{\boldsymbol{\kappa}}^*, \hat{\boldsymbol{\phi}}) = \hat{\boldsymbol{C}} = \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{X}} (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1}, \tag{3.16}$$

$$\hat{\boldsymbol{\Sigma}_0} = \frac{1}{N} (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \hat{\boldsymbol{C}}^T)^T (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \hat{\boldsymbol{C}}^T). \tag{3.17}$$

These are the closed-form solutions for the impact parameters, which are convenient and efficient in computing.

3.2.2 Item Parameters Estimation

The cyclical coordinate descent algorithm for the lasso problem for Generalized Linear Models (Friedman, Hastie, & Tibshirani, 2010) is used to estimate the item and DIF parameters. The item parameters $\boldsymbol{a}_j$ and $\boldsymbol{d}_j$ do not have an L_1 penalty, and the part responsible for updating the unpenalized parameters in the M-step is the same as the Newton-Raphson method.

At iteration $t+1$, the discrimination parameter $\boldsymbol{a}_j$ and the boundary parameter $\boldsymbol{d}_j$ can be updated using the following formulas:

$$\boldsymbol{a}_{jk}^{(t+1)} = \boldsymbol{a}_{jk}^{(t)} - \frac{\partial_{\boldsymbol{a}_{jk}} Q(\boldsymbol{\Delta} | \boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{a}_{jk}} Q(\boldsymbol{\Delta} | \boldsymbol{\Delta}^{(t)})} \tag{3.18}$$

and

$$\boldsymbol{d}_{jg}^{(t+1)} = \boldsymbol{d}_{jg}^{(t)} - \frac{\partial_{\boldsymbol{d}_{jg}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{d}_{jg}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}. \tag{3.19}$$

Specifically, I define $\nu_{ijg} = P^*_{ijg}(\boldsymbol{\theta}_i) - P^*_{ijg}(\boldsymbol{\theta}_i)^2$, and the partial derivatives are given by:

$$\frac{\partial \log Q}{\partial \boldsymbol{a}_{jk}} = \sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{g=1}^{G} \frac{q_{mk} Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})(\nu_{ij(g-1)} - \nu_{ijg})}{P_{jg}(\boldsymbol{\theta}_i)}, \tag{3.20}$$

$$\frac{\partial^2 \log Q}{\partial \boldsymbol{a}^2_{jk}} = \sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{g=1}^{G} -\frac{q^2_{mk} Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})(\nu_{ij(g-1)} - \nu_{ijg})^2}{P_{jg}(\boldsymbol{\theta}_i)^2}, \tag{3.21}$$

$$\frac{\partial \log Q}{\partial \boldsymbol{d}_{jg}} = \sum_{i=1}^{N}\sum_{m=1}^{M} \nu_{ijg}\left(\frac{Y_{ij(g+1)} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{j(g+1)}(\boldsymbol{\theta}_i)} - \frac{Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{jg}(\boldsymbol{\theta}_i)}\right) \quad (g = 1, ..., G-1), \tag{3.22}$$

$$\frac{\partial^2 \log Q}{\partial \boldsymbol{d}^2_{jg}} = \sum_{i=1}^{N}\sum_{m=1}^{M} -\nu^2_{ijg}\left(\frac{Y_{ij(g+1)} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{j(g+1)}(\boldsymbol{\theta}_i)^2} + \frac{Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{jg}(\boldsymbol{\theta}_i)^2}\right) \quad (g = 1, ..., G-1). \tag{3.23}$$

The estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ involves maximizing equation 3.7 with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. The DIF parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are subject to L_1 penalties. For estimating the lasso-penalized parameters, I employ the soft threshold approach, which is generally defined as follows: Given an estimate $s \in \mathbb{R}$, which is shrunk towards zero by computing

$$\hat{\theta}_\tau = \arg\min_{\theta \in \mathbb{R}} \{0.5\theta^2 - s\theta + \tau|\theta|\}, \tag{3.24}$$

where τ is the tuning parameter. The soft-threshold of s and τ is represented by

$$\text{soft}(s, \tau) \equiv \text{sign}(s)(|s| - \tau)_+, \tag{3.25}$$

which yields the global minimizer of the convex objective function $0.5\theta^2 - s\theta + \tau|\theta|$.

As for our DIF parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, denote $\boldsymbol{\Theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta})$, and $\boldsymbol{\Theta}^{(t)}$ is the estimate of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$

at iteration t, denoted as $(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\beta}^{(t)})$. I calculate a quadratic approximation of equation 3.7 at $\boldsymbol{\Theta}^{(t)}$. The lasso estimator of the objective function in equation 3.2 can then be updated as follows:

$$\begin{aligned}
\hat{\boldsymbol{\Theta}} &= \text{argmax}\{Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \eta(||\boldsymbol{\Theta}||_1)\} \\
&= \text{argmin}\{-Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) + \eta(||\boldsymbol{\Theta}||_1)\} \\
&= \text{argmin}\{-Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - (\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(t)})^T \partial_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - (\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(t)})^T \frac{\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{2} (\boldsymbol{\Theta} - \boldsymbol{\Theta}^{(t)}) + \eta||\boldsymbol{\Theta}||_1\} \\
&= -\frac{\text{soft}(\partial_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \boldsymbol{\Theta}^{(t)} \partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}), \eta)}{\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}
\end{aligned} \tag{3.26}$$

The update for $\boldsymbol{\gamma}_j$ and $\boldsymbol{\beta}_j$ is given by:

$$\boldsymbol{\gamma}_{jk}^{(t+1)} = -\frac{\text{soft}(\partial_{\boldsymbol{\gamma}_{jk}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \boldsymbol{\gamma}_{jk}^{(t)} \partial^2_{\boldsymbol{\gamma}_{jk}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}))}{\partial^2_{\boldsymbol{\gamma}_{jk}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}, \tag{3.27}$$

and

$$\boldsymbol{\beta}_{jg}^{(t+1)} = -\frac{\text{soft}(\partial_{\boldsymbol{\beta}_{jg}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \boldsymbol{\beta}_{jg}^{(t)} \partial^2_{\boldsymbol{\beta}_{jg}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}))}{\partial^2_{\boldsymbol{\beta}_{jg}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}, \tag{3.28}$$

where

$$\frac{\partial \log Q}{\partial \boldsymbol{\gamma}_{jk}} = \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{g=1}^{G} \frac{\boldsymbol{X}_i q_{mk} Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})(\nu_{ij(g-1)} - \nu_{ijg})}{P_{jg}(\boldsymbol{\theta}_i)}, \tag{3.29}$$

$$\frac{\partial^2 \log Q}{\partial \boldsymbol{\gamma}_{jk}^2} = \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{g=1}^{G} -\frac{\boldsymbol{X}_i^T \boldsymbol{X}_i q_{mk}^2 Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})(\nu_{ij(g-1)} - \nu_{ijg})^2}{P_{jg}(\boldsymbol{\theta}_i)^2}, \tag{3.30}$$

$$\frac{\partial \log Q}{\partial \boldsymbol{\beta}_{jg}} = \sum_{i=1}^{N} \sum_{m=1}^{M} \nu_{ijg} \left(\frac{\boldsymbol{X}_i Y_{ij(g+1)} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{j(g+1)}(\boldsymbol{\theta}_i)} - \frac{Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{jg}(\boldsymbol{\theta}_i)}\right) \quad (g = 1, ..., G-1), \tag{3.31}$$

$$\frac{\partial^2 \log Q}{\partial \boldsymbol{\beta}_{jg}^2} = \sum_{i=1}^{N}\sum_{m=1}^{M} -\boldsymbol{X}_i^T \boldsymbol{X}_i \nu_{ijg}^2 \left(\frac{Y_{ij(g+1)} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{j(g+1)}(\boldsymbol{\theta}_i)^2} + \frac{Y_{ijg} h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t)})}{P_{jg}(\boldsymbol{\theta}_i)^2}\right) \quad (g = 1, ..., G-1). \tag{3.32}$$

In equation 3.39, it is observed that $\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}^{(t)}|\boldsymbol{\Delta}^{(t)}) < 0$, so $-\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}^{(t)}|\boldsymbol{\Delta}^{(t)}) > 0$. Consequently, I have:

$$\begin{aligned} &- \frac{\text{soft}(\partial_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \boldsymbol{\Theta}^{(t)} \partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}), \eta)}{\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})} \\ &= \text{soft}(\boldsymbol{\Theta}^{(t)} - \frac{\partial_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}, \eta) \end{aligned} \tag{3.33}$$

In the computation, instead of updating $\boldsymbol{a}_j$, $\boldsymbol{d}_j$, $\boldsymbol{\gamma}_j$, and $\boldsymbol{\beta}_j$ using equations 3.18, 3.19, 3.27, and 3.28, a score vector containing the first derivatives of the objective function is calculated as:

$$\boldsymbol{s} = (\frac{\partial \log Q}{\partial \boldsymbol{a}_{jk}}, \frac{\partial \log Q}{\partial \boldsymbol{d}_{jg}}, \frac{\partial \log Q}{\partial \boldsymbol{\gamma}_{jk}}, \frac{\partial \log Q}{\partial \boldsymbol{\beta}_{jg}}), \tag{3.34}$$

and a Hessian matrix is constructed as follows:

$$\boldsymbol{H} = \begin{pmatrix} \frac{\partial^2 \log Q}{\partial \boldsymbol{a}_{jk}^2} & \frac{\partial^2 \log Q}{\partial \boldsymbol{d}_{jg} \partial \boldsymbol{a}_{jk}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\gamma}_{jk} \partial \boldsymbol{a}_{jk}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\beta}_{jg} \partial \boldsymbol{a}_{jk}} \\ \frac{\partial^2 \log Q}{\partial \boldsymbol{a}_{jk} \partial \boldsymbol{d}_{jg}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{d}_{jg}^2} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\gamma}_{jk} \partial \boldsymbol{d}_{jg}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\beta}_{jg} \partial \boldsymbol{d}_{jg}} \\ \frac{\partial^2 \log Q}{\partial \boldsymbol{a}_{jk} \partial \boldsymbol{\gamma}_{jk}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{d}_{jg} \boldsymbol{\gamma}_{jk}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\gamma}_{jk}^2} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\beta}_{jg} \boldsymbol{\gamma}_{jk}} \\ \frac{\partial^2 \log Q}{\partial \boldsymbol{a}_{jk} \partial \boldsymbol{\beta}_{jg}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{d}_{jg} \partial \boldsymbol{\beta}_{jg}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\gamma}_{jk} \partial \boldsymbol{\beta}_{jg}} & \frac{\partial^2 \log Q}{\partial \boldsymbol{\beta}_{jg}^2} \end{pmatrix}. \tag{3.35}$$

Subsequently, I solve $\frac{\partial_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}$ by solving the equation:

$$\boldsymbol{H} \frac{\partial_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})} = -\boldsymbol{s}. \tag{3.36}$$

Finally, $\boldsymbol{a}_j$ and $\boldsymbol{d}_j$ are updated by inserting the corresponding elements in $\frac{\partial \boldsymbol{\Delta} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2 \boldsymbol{\Delta} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}$. As for $\boldsymbol{\gamma}_j$ and $\boldsymbol{\beta}_j$, I calculate $\text{soft}(\boldsymbol{\Theta}^{(t)} - \frac{\partial_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}, \eta)$ using the corresponding element values of $\frac{\partial_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Theta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}$ in $\frac{\partial_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}$.

3.3 Two EMM algorithms

In this section, I discuss two complete algorithms used for the regularization method: Lasso EMM and Group Lasso EMM.

3.3.1 Lasso EMM

The penalty terms in the regularization methods shrink the estimated coefficients towards zero, which helps in reducing model complexity. However, the lasso regularization can introduce bias in the parameter estimates, particularly when the true underlying model has non-zero coefficients. To address this bias, it is common practice to perform an additional EM (Expectation-Maximization) step without any penalty after the initial EM step with regularization. This re-estimation step aims to correct the biased parameter estimates obtained in the first EM cycle. However, our previous research (C. Wang et al., 2023) has shown that this approach doesn't always work, especially under certain simulation conditions. The reason lies in the fact that in the first EM cycle, the biased updates from the M-step are utilized in the second E-step, leading to potentially inaccurate expectations. Consequently, this biased expectation can further influence the subsequent M-step, exacerbating the introduction of more bias into the estimation process.

To overcome these limitations, we proposed a novel algorithm in C. Wang et al. (2023) called the lasso Expectation-Maximization-Maximization (EMM) algorithm. Unlike the traditional approach, the lasso EMM algorithm incorporates a re-estimation step after each EM cycle. This re-estimation step serves to refine the parameter estimates, mitigating the impact of bias introduced in each EM iteration and enhancing the overall accuracy of the estimation process.

The lasso EMM algorithm is summarized in Algorithm 1, where $\boldsymbol{a}_0$, $\boldsymbol{d}_0$, $\boldsymbol{\gamma}_0$,$\boldsymbol{\beta}_0$,$\boldsymbol{\kappa}_0$,$\boldsymbol{\Sigma}_0$,$\boldsymbol{\omega}_0$ represent the initial values of the model parameters, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the data of covariates and response, η is the tuning parameter, and ε_1, and ε_2 are the convergence tolerance.

The parameters are updated iteratively in the algorithm until convergence is achieved, as determined by the tolerance thresholds ε_1 and ε_2. The non-zero DIF parameters are denoted as $\boldsymbol{\gamma}_{jk\neq 0}^{(l)}$ and $\boldsymbol{\beta}_{jg\neq 0}^{(l)}$, indicating that only the non-zero DIF parameters are updated. The

second approach for impact parameter estimation is presented in this section. The intricate details of the algorithm for the first impact estimation approach are provided in section 3.5.

Our research (C. Wang et al., 2023) has shown that the EMM algorithm is more efficient than performing the EM algorithm twice, which was confirmed in our simulations. This efficiency stems from the fact that substantial parameter changes occur in the early cycles of EM, but these changes taper off in the later cycles, making the EMM results very similar to those of the EMM+EM result. Therefore, the second EM re-estimation step can be omitted, saving time and computational resources. I plan to continue using the EMM approach as I explore different types of penalties later on.

3.3.2 Group Lasso EMM

When identifying DIF, it's not uncommon to deal with a large number of covariates that may be correlated, or to have covariates with multiple categories. In such scenarios, the Group Lasso technique becomes a valuable tool for efficiently and simultaneously detecting omnibus DIF across all covariate groups. This method is particularly advantageous when dealing with high-dimensional data, as it encourages sparsity in the selection of DIF items.

This subsection introduces the group lasso penalty and utilizes the block co-ordinate gradient descent algorithm, as introduced by Meier, Van De Geer, and Bühlmann (2008). Notably, in this approach, I group the same DIF parameter for different covariate groups, rather than combining different DIF parameters, such as discrimination $\boldsymbol{\gamma}$ and boundary $\boldsymbol{\beta}$, together. The forthcoming formulas pertain to uniform DIF only, where I group $\boldsymbol{\beta}$s across different covariate groups.

The objective function incorporating the group lasso penalty is given by:

$$S_\eta(\boldsymbol{\beta}) = Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \eta_g \sum_{j}^{J} \sum_{g}^{G} ||\boldsymbol{\beta}_{jg}||_2, \tag{3.37}$$

where η_g represents the group lasso tuning parameter. For convenience, I will continue to use η to denote η_g in the later part of this section.

The E-step remains unchanged from before, and the M-step for impact parameters and item parameters follows the same procedure as before. To estimate the DIF parameters

with the L_2 penalty, I employ the block co-ordinate gradient descent algorithm.

To estimate the L_2 penalty at iteration $t+1$ of the EM cycle, similar to the lasso estimator, a second-order Taylor series expansion of the objective function is calculated at $\hat{\boldsymbol{\beta}}^{(t)}$:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \text{argmax}\{Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \eta \sum_{j}^{J}\sum_{g}^{G} ||\boldsymbol{\beta}_{jg}||_2\} \\ &= \text{argmin}\{-Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) + \eta \sum_{j}^{J}\sum_{g}^{G} ||\boldsymbol{\beta}_{jg}||_2\} \\ &= \text{argmin}\{-Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^T \partial_{\boldsymbol{\beta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)})^T \frac{\partial^2_{\boldsymbol{\beta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})}{2} (\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}) \\ &+ \eta \sum_{j}^{J}\sum_{g}^{G} ||\boldsymbol{\beta}_{jg}^{(t+1)}||_2\},\end{aligned} \tag{3.38}$$

where $\boldsymbol{\beta}_{jg}^{(t+1)} = \boldsymbol{\beta}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}$. Finally, I have

$$\hat{\boldsymbol{\beta}} = \text{argmin}\{-Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)}) - \boldsymbol{\epsilon}^{(t+1)T} \nabla Q - \frac{1}{2}\boldsymbol{\epsilon}^{(t+1)T} H^{(t)} \boldsymbol{\epsilon}^{(t+1)} + \eta \sum_{j}^{J}\sum_{g}^{G} ||\boldsymbol{\beta}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2\}. \tag{3.39}$$

where ∇Q is the score vector containing the first derivatives $\partial_{\boldsymbol{\beta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})$, and $H^{(t)}$ is the Hessian matrix containing $\partial^2_{\boldsymbol{\beta}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t)})$.

For item j, let u denote the subgradient of $||\boldsymbol{\beta}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2$. The equation can be defined as follows:

$$u = \begin{cases} \frac{\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}}{||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2}, & \text{if } \hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)} \neq \mathbf{0} \\ \in \{u : ||u||_2 \leq 1\}, & \text{if } \hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)} = \mathbf{0} \end{cases}. \tag{3.40}$$

The approximated version of the objective function in 3.39 is denoted as $Q_{\eta}^{(t)}(\boldsymbol{\epsilon}^{(t+1)})$. The subgradient equation $\partial_{\epsilon_{jg}} Q_{\eta}^{(t)}(\boldsymbol{\epsilon}^{(t+1)}) = -\nabla Q_{jg} - \boldsymbol{\epsilon}_{jg}^{(t+1)T} H_{jg}^{(t)} + \eta u = 0$ is satisfied with $\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)} = 0$ if:

$$||u||_2 = ||\frac{\nabla Q_{jg} + \boldsymbol{\epsilon}_{jg}^{(t+1)T} H_{jg}^{(t)}}{\eta}||_2 \leq 1 \tag{3.41}$$

$$||\nabla Q_{jg} + \boldsymbol{\epsilon}_{jg}^{(t+1)T} H_{jg}^{(t)}||_2 \leq \eta \tag{3.42}$$

$$||\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_j^{(t)}||_2 \leq \eta. \tag{3.43}$$

The minimizer of $Q_\eta^{(t)}(\boldsymbol{\epsilon}_{jg})$ is given by:

$$\hat{\boldsymbol{\epsilon}}_{jg}^{(t+1)} = -\hat{\boldsymbol{\beta}}_{jg}^{(t)}. \tag{3.44}$$

Alternatively, if $\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)} \neq 0$, then the subgradient equation becomes:

$$\partial_{\epsilon_{jg}} Q_\eta^{(t)}(\boldsymbol{\epsilon}_{jg}^{(t+1)}) = -\nabla Q_{jg} - \boldsymbol{\epsilon}_{jg}^{(t+1)T} H_{jg}^{(t)} + \eta \frac{\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}}{||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2} = 0 \tag{3.45}$$

$$\partial_{\epsilon_{jg}} Q_\eta^{(t)}(\boldsymbol{\epsilon}_{jg}^{(t+1)}) = -\nabla Q_{jg} - (\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}) H_{jg}^{(t)} + \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_{jg}^{(t)} + \eta \frac{\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}}{||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2} = 0 \tag{3.46}$$

$$\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_{jg}^{(t)} = -(\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}) H_{jg}^{(t)} + \eta \frac{\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}}{||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2} \tag{3.47}$$

$$\hat{\boldsymbol{\beta}}_{jg}^{(t+1)} = \hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)} = \frac{(\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_{jg}^{(t)}) ||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2}{\eta - H_{jg}^{(t)} ||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2} \tag{3.48}$$

Taking the norm of both sides of 3.47, I observe that:

$$||\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_{jg}^{(t)}||_2 = (\frac{\eta}{||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2} - H_{jg}^{(t)}) ||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2 \tag{3.49}$$

$$||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2 = \frac{\eta - ||\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_{jg}^{(t)}||_2}{H_{jg}^{(t)}} \tag{3.50}$$

By substituting 3.50 into 3.48, I obtain:

$$\boldsymbol{\epsilon}_{jg}^{(t+1)} = -(H_{jg}^{(t)})^{-1}\{\nabla Q_{jg} - \eta\frac{\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_{jg}^{(t)}}{||\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t)} H_{jg}^{(t)}||_2}\}. \tag{3.51}$$

This $\boldsymbol{\epsilon}_{jg}^{(t+1)}$ represents the direction of movement for β_{jg} in the $t+1$ iteration. For the step length, I perform a Backtracking-Armijo line search as follows:

If $\boldsymbol{\epsilon}_{jg}^{(t+1)} \neq 0$, I let $\alpha^{(t+1)}$ be the largest value in $\{\alpha_0\delta^l\}_{l\geq 0}$ such that:

$$S_\eta(\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \alpha^{(t+1)}\boldsymbol{\epsilon}_{jg}^{(t+1)}) - S_\eta(\hat{\boldsymbol{\beta}}_{jg}^{(t)}) \leq \alpha^{(t+1)}\sigma\tau^{(t+1)}, \tag{3.52}$$

where $\alpha_0 = 1$, $\delta = 0.5$, and $\sigma = 0.1$ are chosen according to the suggestions by Meier et al. (2008), and τ represents the improvement in the objective function $S_\eta(\cdot)$ when using a linear approximation for the log-likelihood, i.e.:

$$\tau_{jg}^{(t+1)} = -\boldsymbol{\epsilon}_{jg}^{(t+1)T}\nabla Q_{jg} + \eta||\hat{\boldsymbol{\beta}}_{jg}^{(t)} + \boldsymbol{\epsilon}_{jg}^{(t+1)}||_2 - \eta||\hat{\boldsymbol{\beta}}_{jg}^{(t)}||_2. \tag{3.53}$$

As a result, the $\hat{\boldsymbol{\beta}}_{jg}^{(t+1)}$ is updated as follows:

$$\hat{\boldsymbol{\beta}}_{jg}^{(t+1)} = \hat{\boldsymbol{\beta}}_{jg}^{(t)} + \alpha^{(t+1)}\boldsymbol{\epsilon}_{jg}^{(t+1)}. \tag{3.54}$$

The complete group lasso algorithm is presented in the table of Algorithm 2 below.

3.4 Tuning Parameter Selection

To determine the optimal value of the tuning parameter η, which yields accurate DIF detection results, I calculate the Bayesian Information Criterion (BIC) for both the lasso EMM and group lasso EMM algorithms. For each η value, I obtain a set of L_1 regularized Marginal Maximum Likelihood Estimators, denoted as $\boldsymbol{\Delta}_\eta$. The BIC is computed using the following formula:

$$BIC_{\hat{\boldsymbol{\Delta}}_\eta} = -2\max_{\hat{\boldsymbol{\Delta}}_\eta}\log M + (\|\hat{\boldsymbol{\beta}}\|_0 + \|\hat{\boldsymbol{\gamma}}\|_0)\log N, \tag{3.55}$$

where the log-marginal likelihood, denoted as $\log M$, is calculated as follows:

$$\log M = \sum_{i=1}^{N} \log \int (\prod_{j=1}^{J} l(\boldsymbol{a_j}, \boldsymbol{d_j}, \boldsymbol{\beta}_j, \boldsymbol{\gamma} | \boldsymbol{X}_i, \boldsymbol{u}_{ij}, \boldsymbol{\theta}_i)) f(\boldsymbol{\kappa}, \boldsymbol{\omega} \mid \boldsymbol{\theta}_i, , \boldsymbol{X}_i, \mu_0, \boldsymbol{\Sigma}_0) \partial \boldsymbol{\theta}_i$$
$$= \sum_{i=1}^{N} \log \sum_{g=1}^{G} \exp \sum_{j=1}^{J} [u_{ij} \log(P_j(\boldsymbol{q}_g)) + (1 - u_{ij}) \log(1 - P_j(\boldsymbol{q}_g))] f(\boldsymbol{\kappa}, \boldsymbol{\omega} \mid \boldsymbol{q}_g, , \boldsymbol{X}_i, \mu_0, \boldsymbol{\Sigma}_0) (Step)^K. \tag{3.56}$$

Here, $Step = \frac{h_\theta - l_\theta}{M_0}$. The first term in equation 3.55 controls the bias of the estimator, while the second term penalizes the model's complexity. Our goal is to select the tuning parameter η that results in the smallest BIC value.

3.5 Programming Details

In Section 3.2.1, it was noted that there are no closed-form solutions for the parameters $\boldsymbol{\kappa}$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\omega}$ in impact estimation Approach I. Consequently, I adopted an approximation method using the *optim* function in R. However, this approach may lead to computational inefficiency as it requires iterative approximation of the optimization problem. In each iteration, the complete-data log-likelihood function must be calculated, including the summation over the sample size N.

A similar computational inefficiency was encountered when performing the cyclical coordinate descent method to solve the DIF parameters in the M-step, as the complete-data likelihood function also needed to be summed over N in each iteration. To address this issue and improve efficiency, the decision was made to write that part of the code in c++.

This section presents the efforts made to explore potential ways to speed up the computing process of the complete likelihood function for the impact parameters estimation. Specifically, I examine the dimensions of matrices to explore potential alternatives for matrix operations, particularly in relation to the sample size, denoted as N.

To begin, I first find the total number of covariate group combinations denoted as T. For instance, if I have three covariates—Covariate 1 and Covariate 2 are binary, and Covariate 3 is continuous (integer) age ranging from 21 to 84—the total number of covariate group combinations would be 2 * 2 * 64 = 256.

I create a T-by-P matrix denoted as $\boldsymbol{\Upsilon}$ to store these combinations:

$$\boldsymbol{\Upsilon} = \begin{pmatrix} 0 & 0 & 21 \\ 0 & 0 & 22 \\ 0 & 0 & 23 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 84 \end{pmatrix}.$$

For $t = 1,2,...,T$, denote N_t as the number of individuals with covariate combination $\boldsymbol{\Upsilon}_t$. For example, 200 individuals are in the first covariate group $\boldsymbol{\Upsilon}_1 = (0, 0, 21)$ (200 individuals have covariate 1 value 0, Covariate 2 value 0, Covariate 3 value 21), then $N_1 = 200$.

The complete-data log-likelihood function is formulated as follows:

$$f(\boldsymbol{\kappa}) = \sum_{i}^{N} \sum_{g}^{G} h(\boldsymbol{\theta}_g \mid \boldsymbol{X}_i, \boldsymbol{u}_i, \boldsymbol{\Delta}^{(t)})(-\frac{1}{2}(\boldsymbol{\theta}_g - \boldsymbol{\kappa}\boldsymbol{X}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\theta}_g - \boldsymbol{\kappa}\boldsymbol{X}_i)) \tag{3.57}$$

Some terms are denoted as follows:

$$\boldsymbol{h} = h(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{u}, \boldsymbol{\Delta}^{(t)}) = \begin{pmatrix} h(\boldsymbol{\theta}_1 \mid \boldsymbol{X}_1, \boldsymbol{u}_1, \boldsymbol{\Delta}^{(t)}) & ... & h(\boldsymbol{\theta}_G \mid \boldsymbol{X}_1, \boldsymbol{u}_1, \boldsymbol{\Delta}^{(t)}) \\ h(\boldsymbol{\theta}_1 \mid \boldsymbol{X}_2, \boldsymbol{u}_2, \boldsymbol{\Delta}^{(t)}) & ... & h(\boldsymbol{\theta}_G \mid \boldsymbol{X}_2, \boldsymbol{u}_2, \boldsymbol{\Delta}^{(t)}) \\ \vdots & \vdots & \vdots \\ h(\boldsymbol{\theta}_1 \mid \boldsymbol{X}_N, \boldsymbol{u}_N, \boldsymbol{\Delta}^{(t)}) & ... & h(\boldsymbol{\theta}_G \mid \boldsymbol{X}_N, \boldsymbol{u}_N, \boldsymbol{\Delta}^{(t)}) \end{pmatrix}$$

is a $N - by - G$ matrix.

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K) = \begin{pmatrix} \boldsymbol{\theta}_{11} & ... & \boldsymbol{\theta}_{1K} \\ \boldsymbol{\theta}_{21} & ... & \boldsymbol{\theta}_{2K} \\ \vdots & \vdots & \vdots \\ \boldsymbol{\theta}_{G1} & ... & \boldsymbol{\theta}_{GK} \end{pmatrix}$$

is a $G - by - K$ matrix.

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{11} & ... & \boldsymbol{X}_{1P} \\ \boldsymbol{X}_{21} & ... & \boldsymbol{X}_{2P} \\ \vdots & \vdots & \vdots \\ \boldsymbol{X}_{N1} & ... & \boldsymbol{X}_{NP} \end{pmatrix}$$

is $N - by - P$ vector.

$$\boldsymbol{\kappa} = \begin{pmatrix} \boldsymbol{\kappa}_{11} & ... & \boldsymbol{\kappa}_{1P} \\ \boldsymbol{\kappa}_{21} & ... & \boldsymbol{\kappa}_{2P} \\ \vdots & \vdots & \vdots \\ \boldsymbol{\kappa}_{K1} & ... & \boldsymbol{\kappa}_{KP} \end{pmatrix}$$

is $K - by - P$ matrix.

In the *optim* function, the parameters to be optimized must be input as a vector. Therefore, I transform $\boldsymbol{\kappa}$ into a vector of length $K \times P$. This transformation is performed prior to using it as the starting value in the *optim* function, and then the vector form of $\boldsymbol{\kappa}$ is reshaped back to a $K - by - P$ matrix after obtaining the result vector from the *optim* function.

I denote the vectorized form of $\boldsymbol{\kappa}$ as $vec(\boldsymbol{\kappa})$, which is expressed as:

$$vec(\boldsymbol{\kappa}) = \begin{pmatrix} \boldsymbol{\kappa}_{11} & \boldsymbol{\kappa}_{12} & ... & \boldsymbol{\kappa}_{K(P-1)} & \boldsymbol{\kappa}_{KP} \end{pmatrix},$$

and is a vector of length $K \times P$.

Recall that

$$\boldsymbol{\Sigma_i} = \boldsymbol{L_i} \boldsymbol{L_i'}, \tag{3.58}$$

and

$$\boldsymbol{L_i} = \boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i}\boldsymbol{\omega}))], \tag{3.59}$$

where $\boldsymbol{\Sigma_0} = \boldsymbol{L_0}\boldsymbol{L_0'}$ is the K-by-K covariance matrix for $\boldsymbol{X_i}$ at reference level. $\boldsymbol{\omega}$ is a

$P \times \frac{K(K+1)}{2}$ matrix. $\boldsymbol{X_i\omega}$ is a 1-by-$\frac{K(K+1)}{2}$ matrix and I reshape $\exp(\boldsymbol{X_i\omega})$ into a lower triangular matrix. Equation 3.58 is a Cholesky decomposition of $\boldsymbol{\Sigma_i}$.

The need to iterate over N or T arises from the structure of $\boldsymbol{\Sigma_i}$ in Equation 3.60, which is calculated as shown in Equations 3.58 and 3.59. The process used for reshaping prevents the data from being combined into a single matrix over the sample size N or the number of group combinations T. Detailed algorithms for iterating over N and T to calculate the complete-data log-likelihood are systematically detailed in Algorithms 3 and 4.

When working with categorical covariates where the total number of group combinations T is smaller than the sample size N, the use of Algorithm 4 is advised. On the other hand, Algorithm 3 becomes more applicable in scenarios involving continuous covariates or when the total group combination T surpasses the sample size N.

The estimation methods for $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\omega}$ are similar to those used for $\boldsymbol{\kappa}$. Currently, there is no way to represent these in matrix form instead of looping over N or T. Depending on whether N is less than or greater than T, I need to modify Algorithms 3 and 4 for functions of $vec(\boldsymbol{\Sigma}_0)$ or $vec(\boldsymbol{\omega})$. This includes using the full version of $f(\boldsymbol{\omega})$, as presented in Equation 3.60, within the summation term.

$$f(\boldsymbol{\omega}) = \sum_{i}^{N} \sum_{g}^{G} h(\boldsymbol{\theta}_g \mid \boldsymbol{X}_i, \boldsymbol{u}_i, \boldsymbol{\Delta}^{(t)})(-\frac{1}{2} \log |\boldsymbol{\Sigma_i}| - \frac{1}{2}(\boldsymbol{\theta}_g - \boldsymbol{\kappa X}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\theta}_g - \boldsymbol{\kappa X}_i)). \quad (3.60)$$

Algorithm 1: Lasso EMM Algorithm

Input : $\boldsymbol{a}_0$,$\boldsymbol{d}_0$,$\boldsymbol{\gamma}_0$,$\boldsymbol{\beta}_0$,$\boldsymbol{\kappa}_0^*$,$\boldsymbol{\Sigma}_0$,$\boldsymbol{\phi}_0$,$\boldsymbol{X}$,$\boldsymbol{Y}$,η,ε_1,ε_2

Output: $\hat{\boldsymbol{a}}$,$\hat{\boldsymbol{d}}$,$\hat{\boldsymbol{\gamma}}$,$\hat{\boldsymbol{\beta}}$,$\hat{\boldsymbol{\kappa}^*}$,$\hat{\boldsymbol{\Sigma}}_0$,$\hat{\boldsymbol{\phi}}$

set $t = 1$, $\boldsymbol{a}^{(0)} = \boldsymbol{a}_0$, $\boldsymbol{d}^{(0)} = \boldsymbol{d}_0$, $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}_0$, $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}_0$, $\boldsymbol{\kappa}^{(0)} = \boldsymbol{\kappa}_0$, $\boldsymbol{\Sigma}_0^{(0)} = \boldsymbol{\Sigma}_0$,
$\boldsymbol{\phi}^{(0)} = \boldsymbol{\phi}_0$, $\delta_1^{(0)} = 1$;

while $\delta_1^{(t-1)} > \varepsilon_1$ **do**

- Calculate $h(\boldsymbol{q}_m|\boldsymbol{X}_i, \boldsymbol{Y}_{ij}, \boldsymbol{\Delta}^{(t-1)})$;
- Calculate $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$;
- Update $(\hat{\boldsymbol{\kappa}}^*, \hat{\boldsymbol{\phi}}) = \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{N}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{C}}^T)^T(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{C}}^T)$;
- **for** $j = 1, ...J$ **do**
  - set $l = 1$, $\delta_2^{(0)} = 1$;
  - **while** $\delta_2^{(l-1)} > \varepsilon_2$ **do**
    - Solve $\boldsymbol{H}\frac{\partial_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})} = -\boldsymbol{s}$, where $\frac{\partial_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})} = (\boldsymbol{\Delta}_a, \boldsymbol{\Delta}_d, \boldsymbol{\Delta}_\gamma, \boldsymbol{\Delta}_\beta)$;
    - $\boldsymbol{a}_j^{(l)} = \boldsymbol{a}_j^{(l-1)} - \boldsymbol{\Delta}_a$, $\boldsymbol{d}_j^{(l)} = \boldsymbol{d}_j^{(l-1)} - \boldsymbol{\Delta}_d$;
    - $\boldsymbol{\gamma}_{jk\neq 0}^{(l)} = \text{soft}(\boldsymbol{\gamma}_{jk\neq 0}^{(l-1)} - \boldsymbol{\Delta}_\gamma), -\frac{\eta}{\partial^2_{\boldsymbol{\gamma}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})})$;
    - $\boldsymbol{\beta}_{jg\neq 0}^{(l)} = \text{soft}(\boldsymbol{\beta}_{jg\neq 0}^{(l-1)} - \boldsymbol{\Delta}_\beta), -\frac{\eta}{\partial^2_{\boldsymbol{\beta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})})$;
    - $\delta_2^{(l)} = ||\boldsymbol{a}_j^{(l)} - \boldsymbol{a}_j^{(l-1)}|| + ||\boldsymbol{d}_j^{(l)} - \boldsymbol{d}_j^{(l-1)}|| + ||\boldsymbol{\gamma}_j^{(l)} - \boldsymbol{\gamma}_j^{(l-1)}|| + ||\boldsymbol{\beta}_j^{(l)} - \boldsymbol{\beta}_j^{(l-1)}||$;
    - $l = l + 1$;
- **for** $j = 1, ...J$ **do**
  - set $l = 1$, $\delta_2^{(0)} = 1$;
  - **while** $\delta_2^{(l-1)} > \varepsilon_2$ **do**
    - Solve $\boldsymbol{H}\frac{\partial_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})} = -\boldsymbol{s}$;
    - $(\boldsymbol{a}_j^{(l)}, \boldsymbol{d}_j^{(l)}, \boldsymbol{\gamma}_{jk\neq 0}^{(l)}, \boldsymbol{\beta}_{jg\neq 0}^{(l)}) = (\boldsymbol{a}_j^{(l-1)}, \boldsymbol{d}_j^{(l-1)}, \boldsymbol{\gamma}_{jk\neq 0}^{(l-1)}, \boldsymbol{\beta}_{jg\neq 0}^{(l-1)}) - \frac{\partial_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})}{\partial^2_{\boldsymbol{\Delta}} Q(\boldsymbol{\Delta}^{(l)}|\boldsymbol{\Delta}^{(t)})}$,
    - $\delta_2^{(l)} = ||\boldsymbol{a}_j^{(l)} - \boldsymbol{a}_j^{(l-1)}|| + ||\boldsymbol{d}_j^{(l)} - \boldsymbol{d}_j^{(l-1)}|| + ||\boldsymbol{\gamma}_j^{(l)} - \boldsymbol{\gamma}_j^{(l-1)}|| + ||\boldsymbol{\beta}_j^{(l)} - \boldsymbol{\beta}_j^{(l-1)}||$;
    - $l = l + 1$;
- $\delta_1^{(t)} = ||\boldsymbol{a}^{(t)} - \boldsymbol{a}^{(t-1)}|| + ||\boldsymbol{d}^{(t)} - \boldsymbol{d}^{(t-1)}|| + ||\boldsymbol{\gamma}^{(t)} - \boldsymbol{\gamma}^{(t-1)}|| + ||\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}|| + ||\boldsymbol{\kappa}^{(t)} - \boldsymbol{\kappa}^{(t-1)}|| + ||\boldsymbol{\phi}^{(t)} - \boldsymbol{\phi}^{(t-1)}|| + ||\boldsymbol{\Sigma}_0^{(t)} - \boldsymbol{\Sigma}_0^{(t-1)}||$;
- $t = t + 1$;

Algorithm 2: Uniform DIF Detection via group LASSO

Input : $\boldsymbol{a}_0$,$\boldsymbol{d}_0$,$\boldsymbol{\gamma}_0$,$\boldsymbol{\beta}_0$,$\boldsymbol{\kappa}_0^*$,$\boldsymbol{\Sigma}_0$,$\boldsymbol{\phi}_0$,$\boldsymbol{X}$,$\boldsymbol{Y}$,η,ε_1,ε_2

Output: $\hat{\boldsymbol{a}}$,$\hat{\boldsymbol{d}}$,$\hat{\boldsymbol{\gamma}}$,$\hat{\boldsymbol{\beta}}$,$\hat{\boldsymbol{\kappa}^*}$,$\hat{\boldsymbol{\Sigma}}_0$,$\hat{\boldsymbol{\phi}}$

set $t=1$, $\boldsymbol{a}^{(0)}=\boldsymbol{a}_0$, $\boldsymbol{d}^{(0)}=\boldsymbol{d}_0$, $\boldsymbol{\gamma}^{(0)}=\boldsymbol{\gamma}_0$, $\boldsymbol{\beta}^{(0)}=\boldsymbol{\beta}_0$, $\boldsymbol{\kappa}^{(0)}=\boldsymbol{\kappa}_0$, $\boldsymbol{\Sigma}_0^{(0)}=\boldsymbol{\Sigma}_0$, $\boldsymbol{\phi}^{(0)}=\boldsymbol{\phi}_0$, $\delta_1^{(0)}=1$

while $\delta_1^{(t-1)} > \varepsilon_1$ **do**

  Calculate $h(\boldsymbol{q}_m|\boldsymbol{X}_i,\boldsymbol{Y}_{ij},\boldsymbol{\Delta}^{(t-1)})$

  Calculate $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$

  Update $(\hat{\boldsymbol{\kappa}}^*,\hat{\boldsymbol{\phi}}) = \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{N}(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}\hat{\boldsymbol{C}}^T)^T(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{X}}\hat{\boldsymbol{C}}^T)$

  for $j=1,...J$ **do**

    set $t_2=1$, $\delta_2^{(0)}=1$

    while $\delta_2^{(t_2-1)} > \varepsilon_2$ **do**

      $a_{jk}^{(t_2)} = a_{jk}^{(t_2-1)} - \frac{\partial_{\boldsymbol{a}_{jk}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t_2-1)})}{\partial^2_{a_{jk}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t_2-1)})}$, $d_{jg}^{(t_2)} = d_{jg}^{(t_2-1)} - \frac{\partial_{\boldsymbol{d}_{jg}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t_2-1)})}{\partial^2_{d_{jg}} Q(\boldsymbol{\Delta}|\boldsymbol{\Delta}^{(t_2-1)})}$

      if $||\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)} H_{jg}^{(t_2-1)}||_2 \le \eta$ **then**

        $\hat{\boldsymbol{\beta}}_{jg}^{(t_2)} = \boldsymbol{0}$

      else

        $\boldsymbol{\epsilon}_{jg}^{(t_2)} = -(H_{jg}^{(t_2-1)})^{-1}\{\nabla Q_{jg} - \eta\frac{\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)} H_{jg}^{(t_2-1)}}{||\nabla Q_{jg} - \hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)} H_{jg}^{(t_2-1)}||_2}\}$

        $\tau_{jg}^{(t_2)} = -\boldsymbol{\epsilon}_{jg}^{(t_2)T}\nabla Q_{jg} + \eta||\hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)} + \boldsymbol{\epsilon}_{jg}^{(t_2)}||_2 - \eta||\hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)}||_2$

        $\alpha^{(t_2)}$ is the max value in $\{\alpha^{(0)}\delta^l\}_{l\ge 0}$ such that

        $S_\eta(\hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)} + \alpha^{(t_2)}\boldsymbol{\epsilon}_{jg}^{(t_2)}) - S_\eta(\hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)}) \le \alpha^{(t_2)}\sigma\tau^{(t_2)}$.

        $\hat{\boldsymbol{\beta}}_{jg}^{(t_2)} = \hat{\boldsymbol{\beta}}_{jg}^{(t_2-1)} + \alpha^{(t_2)}\boldsymbol{\epsilon}_{jg}^{(t_2)}$

      end

      $\delta_2^{(t_2)} = ||\hat{\boldsymbol{A}}_j^{(t_2)} - \hat{\boldsymbol{A}}_j^{(t_2-1)}|| + ||\hat{\boldsymbol{D}}_j^{(t_2)} - \hat{\boldsymbol{D}}_j^{(t_2-1)}|| + ||\alpha^{(t_2)}\boldsymbol{\epsilon}_{jg}^{(t_2)}||$

      $t_2 = t_2+1$

    end

  end

  $\delta_1^{(t_1)} = ||\hat{\boldsymbol{A}}^{(t_1)} - \hat{\boldsymbol{A}}^{(t_1-1)}|| + ||\hat{\boldsymbol{D}}^{(t_1)} - \hat{\boldsymbol{D}}^{(t_1-1)}|| + ||\hat{\boldsymbol{\beta}}^{(t_1)} - \hat{\boldsymbol{\beta}}^{(t_1-1)}||$

  $t_1 = t_1+1$

end

Algorithm 3: Impact on mean estimation Loop over N

Define a function of $vec(\boldsymbol{\kappa})$ called 'object'. Start Function;
Define a scalar Sum = 0 ;
Reshape $vec(\boldsymbol{\kappa})$ into $\boldsymbol{\kappa}$;
for *i=1,...,N* **do**
  Calculate the product $\boldsymbol{\kappa X_i}$ (Note: the product $\boldsymbol{\kappa X_i}$ is 1-by-K);
  Calculate the product $\exp(\boldsymbol{X_i\omega})$ (Note: $\exp(\boldsymbol{X_i\omega})$ is 1-by-$K(K+1)/2$);
  Reshape $\exp(\boldsymbol{X_i\omega})$ into a K-by-K lower triangular matrix;
  $\boldsymbol{\Sigma}_i = \boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{X_i\omega}))$ (Note: $\boldsymbol{\Sigma}_i$ is K-by-K);
  $\text{Sum} = \text{Sum} + \sum(\boldsymbol{h}_i * -0.5 * diag[(\boldsymbol{X} - rep(\boldsymbol{\kappa X_i}, G)) * \boldsymbol{\Sigma}_i^{-1} * (\boldsymbol{X} - rep(\boldsymbol{\kappa X_i}, G))'];$
  (Note: In the equation above, $\boldsymbol{h}_i$ is the ith row of $\boldsymbol{h}$ which is also a vector of length G. To avoid sum over G, I use $rep()$ to replicate $\boldsymbol{\kappa X_i}$ by G times to match the dimension of $\boldsymbol{X}$ which is G-by-K.
  $(\boldsymbol{X} - rep(\boldsymbol{\kappa X_i}, G)) * \boldsymbol{\Sigma}_i^{-1} * (\boldsymbol{X} - rep(\boldsymbol{\kappa X_i}, G))'$ is a G-by-G matrix. the function $diag()$ take its diagonal values to form a vector of length G. $\sum$ takes the summation of G values in the vector to form a scalar);
Return −Sum.

Algorithm 4: Impact on mean estimation Loop over T

Define a function of $vec(\boldsymbol{\kappa})$ called 'object'. Start Function;
Define a scalar Sum = 0 ;
Reshape $vec(\boldsymbol{\kappa})$ into $\boldsymbol{\kappa}$;
for *t=1,...,T* **do**
  Calculate the product $\boldsymbol{\kappa\Upsilon_t}$ (Note: $\boldsymbol{\kappa\Upsilon_t}$ is 1-by-K);
  Calculate the product $\exp(\boldsymbol{\Upsilon_t\omega})$;
  Reshape $\exp(\boldsymbol{\Upsilon_t\omega})$ into a K-by-K lower triangular matrix;
  $\boldsymbol{\Sigma}_i = \boldsymbol{L_0} \circ \text{reshape}(\exp(\boldsymbol{\Upsilon_t\omega}))$ (Note: $\boldsymbol{\Sigma}_i$ is K-by-K);
  Define ind1= the row indicators of individuals with covariate combination $\boldsymbol{\Upsilon_t}$ (Note: ind1 is a vector of length N_t);
  $\text{Sum} = \text{Sum} + \sum(\text{colsum}(\boldsymbol{h}[\text{ind1},]) * -0.5 * diag[(\boldsymbol{X} - rep(\boldsymbol{\kappa\Upsilon_t}, G)) * \boldsymbol{\Sigma}_i^{-1} * (\boldsymbol{X} - rep(\boldsymbol{\kappa\Upsilon_t}, G))'];$
  (Note:$\boldsymbol{h}[\text{ind1},]$ is a matrix of N_t-by-G. $colsum(\boldsymbol{h}[\text{ind1},]$ is a vector of length G);
Return −Sum.

Chapter 4

SIMULATION

I conducted three simulation studies to systematically evaluate the regularization methods. The first study focused on the application of the lasso penalty method under the condition of uniform DIF, involving both categorical and continuous covariates, with the presence of impact. In the second study, the lasso penalty method was examined in scenarios characterized by non-uniform DIF, again with both categorical and continuous covariates, and the existence of impact. For the third study, a previously generated dataset in C. Wang et al. (2023) was utilized to assess the performance of the group lasso method.

4.1 Simulation Study I

In the first simulation study, I evaluated the lasso EMM algorithm's ability to detect Uniform DIF associated with two covariates in a 2DGRM setting. The test length was held constant at 12, consistent with Bauer et al. (2020). The item parameters for the graded response model are generated following the method outlined in Jiang et al. (2016). The item loadings demonstrate a simple structure, with the first six items being loaded on the first factor and the second six items loaded on the second factor. Item parameters are defined as follows: $\boldsymbol{a}$ is drawn from a uniform distribution with boundaries of 1.1 and 2.8, while b_1, b_2, and b_3 are drawn from uniform distributions with boundaries of -2 to -0.67, -0.67 to 0.67, and 0.67 to 2, respectively. For each item, boundary parameters d are computed as the product of a and each b, and they are arranged in descending order (Jiang et al., 2016). The generated item parameters are presented in Table 4.1.

The design for the covariates was borrowed from Belzak & Bauer (2020). In their study, three covariates were simulated to represent Study (binary), Gender (binary), and Age (9 categories). A correlation of .30 was observed between Gender and Study, -.51 between Age and Study, and -.15 between Gender and Age. In this study, I divided their design into

Table 4.1: Simulated True Item Parameters

Item	1	2	3	4	5	6
a_1	1.551	1.733	2.074	2.644	1.443	2.627
a_2	0	0	0	0	0	0
d_1	2.037	2.464	2.160	3.125	1.475	2.854
d_2	-0.074	-1.108	0.733	-1.716	0.168	0.260
d_3	-1.526	-3.268	-2.288	-3.063	-2.801	-2.048
Item	7	8	9	10	11	12
a_1	0	0	0	0	0	0
a_2	2.706	2.223	2.169	1.205	1.450	1.400
d_1	3.526	3.582	1.777	1.185	1.980	2.145
d_2	0.826	-1.374	1.105	-0.439	0.575	-0.836
d_3	-2.627	-2.187	-1.821	-1.186	-2.352	-1.360

two conditions with manipulated correlations between covariates. In the first condition, two binary covariates were independently generated using a Bernoulli distribution with a probability of 0.5, serving as a baseline for comparison. For the second design, a binary and a continuous covariate were simulated. The binary covariate was again generated from a Bernoulli distribution with a probability of 0.5, denoted as covariate 1. The generation of the continuous covariate followed the approach of Belzak & Bauer (2020): for individuals with a covariate 1 value of 0, it was generated from a Binomial distribution with 7 trials and a success probability of 0.5, then added by 10; for those with a covariate 1 value of 1, it was generated from a Binomial distribution with 6 trials and a success probability of 0.5, then added by 9. The values for the second covariate were then adjusted by subtracting 13 and dividing by 3, resulting in the scaled covariate 2 having a mean around 0 and a variance around 0.25, which matches the variance of covariate 1. This method resulted in an approximate correlation of -0.5 between the two covariates.

Within each design, I manipulated two levels of sample size (500 and 2000) and two levels of DIF item proportion (33% and 66%). Additionally, two levels of differential item functioning (DIF) magnitude were simulated in each condition. For items affected by DIF, the same magnitude of DIF effects was uniformly applied across the three boundary parameters of each item (Cohen et al., 1993). The first half of the items was assigned a small

magnitude DIF of 0.5, while the latter half received a large magnitude DIF of 1. In the condition with a 33% DIF prevalence, items 3, 4, 9, and 10 were the designated DIF items. In the scenario where 66% of the items exhibited DIF, items 3, 4, 5, 6, 9, 10, 11, and 12 were affected. The specifications of the DIF parameters are detailed in Table 4.2. $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ stand for covariate 1 and covariate 2, respectively.

Table 4.2: Simulated True DIF Parameters

	DIF %		$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\beta}_3$	
Item	33%	66%	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$
1								
2								
3	√	√	.5		.5		.5	
4	√	√	.5	.5	.5	.5	.5	.5
5		√		.5		.5		.5
6		√	.5	.5	.5	.5	.5	.5
7								
8								
9	√	√	1		1		1	
10	√	√	1	1	1	1	1	1
11		√		1		1		1
12		√	1	1	1	1	1	1

I generated two levels of impact using the previously described impact parameterization methods, and the impact parameters were estimated using each method accordingly. For the first method, I generated latent variables from bivariate normal distributions with a mean vector of $\mathbf{0}$ and variances of 1. In our earlier study, C. Wang et al. (2023), we manipulated the correlation between two trait dimensions to be either 0.25 or 0.85, and observed that the DIF detection results were not significantly affected by these variations. Consequently, I fixed the correlation between the two trait dimensions at 0.5 in this study. Aligning with the conditions described in Curran, Cole, Bauer, Hussong, and Gottfredson (2016), I chose impact parameters close to the magnitudes of medium mean and small variance impact conditions. Three impact parameters were simulated for this first estimation method as

follows:

$$\kappa = \begin{pmatrix} 0.37 & 0.37 \\ 0 & 0 \end{pmatrix},$$

$$\omega = \begin{pmatrix} 0.15 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.15 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Three impact parameters for the second impact estimation method are simulated as:

$$\kappa = \begin{pmatrix} 0 & 0 \\ 0.37 & 0.37 \\ 0.22 & 0.22 \end{pmatrix},$$

$$\phi = \begin{pmatrix} 0.15 & 0.15 \\ 0.15 & 0.15 \\ 0.15 & 0.15 \end{pmatrix}, \Sigma_0 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

To offer a more intuitive understanding of the impact size we generated, I am providing the mean vectors and covariance matrices for select covariate groups. Using the first impact estimation method, the impact parameter matrix $\kappa = \begin{pmatrix} 0.37 & 0.37 \\ 0 & 0 \end{pmatrix}$ yields specific mean vectors for different focal groups. For individuals with the first covariate at 1 and the second at 0, i.e., $\boldsymbol{X} = (1, 0)$, the mean vector is $\boldsymbol{\mu} = (0.37, 0.37)^T$. For those with the first covariate at 0 and the second at 1, i.e., $\boldsymbol{X} = (0, 1)$, the mean vector is $\boldsymbol{\mu} = (0, 0)^T$. Lastly, for individuals with both covariates at 1, i.e., $\boldsymbol{X} = (1, 1)$, the mean vector is $\boldsymbol{\mu} = (0.37, 0.37)^T$. The impact parameter matrix $\omega = \begin{pmatrix} 0.15 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.15 \end{pmatrix}$ yields specific covariance matrices for different focal groups. For individuals with the first covariate at 1 and the second at 0, i.e., $\boldsymbol{X} = (1, 0)$, the covariance matrix is $\boldsymbol{\Sigma} = \begin{pmatrix} 1.35 & 0.675 \\ 0.675 & 1.35 \end{pmatrix}$. For those with the first covariate at 0 and the second at 1, i.e., $\boldsymbol{X} = (0, 1)$, the covariance matrix is $\boldsymbol{\Sigma} = \begin{pmatrix} 1.35 & 0.675 \\ 0.675 & 1.35 \end{pmatrix}$. Lastly, for individuals with both covariates at 1, i.e., $\boldsymbol{X} = (1, 1)$, the

covariance matrix is $\boldsymbol{\Sigma} = \begin{pmatrix} 1.822 & 0.911 \\ 0.911 & 1.822 \end{pmatrix}$.

Similarly, In the approach employing the second impact estimation method, the impact parameter matrix is given as $\kappa = \begin{pmatrix} 0 & 0 \\ 0.37 & 0.37 \\ 0.22 & 0.22 \end{pmatrix}$, which yields particular mean vectors for different focal groups. Groups with $\boldsymbol{X} = (1, 0)$ are attributed a mean vector $\boldsymbol{\mu} = (0.37, 0.37)^T$. Groups with $\boldsymbol{X} = (0, 1)$ are attributed a mean vector $\boldsymbol{\mu} = (0.22, 0.22)^T$, while those with $\boldsymbol{X} = (1, 1)$ corresponds to $\boldsymbol{\mu} = (0.59, 0.59)^T$. The impact parameter matrix is set to be $\phi = \begin{pmatrix} 0.15 & 0.15 \\ 0.15 & 0.15 \\ 0.15 & 0.15 \end{pmatrix}$, resulting in distinct covariance matrices for various groups. The reference level group, $\boldsymbol{X} = (0, 0)$, is provided with a covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1.023 & 0.523 \\ 0.523 & 1.023 \end{pmatrix}$. Groups with $\boldsymbol{X} = (1, 0)$ receive $\boldsymbol{\Sigma} = \begin{pmatrix} 1.09 & 0.59 \\ 0.59 & 1.09 \end{pmatrix}$, and $\boldsymbol{X} = (0, 1)$ leads to $\boldsymbol{\Sigma} = \begin{pmatrix} 1.09 & 0.59 \\ 0.59 & 1.09 \end{pmatrix}$. Finally, for $\boldsymbol{X} = (1, 1)$, the covariance matrix is $\boldsymbol{\Sigma} = \begin{pmatrix} 1.202 & 0.703 \\ 0.703 & 1.202 \end{pmatrix}$.

To measure the simulated DIF size, the weighted Average Area Between Curves (wABC) as described by Edelen, Stucky, and Chandra (2015), which quantifies the average area between the expected score curves from two groups, is calculated for each item and each covariate. The calculation of wABC in Edelen et al. (2015) is as follows:

The probability of response $u = g$ is given by the equation:

$$P_{jg}(\boldsymbol{\theta}) = P^*_{j(g-1)}(\boldsymbol{\theta}) - P^*_{jg}(\boldsymbol{\theta}) \quad (g = 1, ..., G), \tag{4.1}$$

where $P^*_{jg}(\boldsymbol{\theta}) = \frac{1}{1+e^{-(\boldsymbol{\alpha}_j^T \boldsymbol{\theta} + d_{jg} + \boldsymbol{X}_i \boldsymbol{\gamma}_j \boldsymbol{\theta} + \boldsymbol{X}_i \boldsymbol{\beta}_{jg})}}$ for $g = 1, ..., G-1$, $P^*_{j0}(\boldsymbol{\theta}) = 1$, and $P^*_{jG}(\boldsymbol{\theta}) = 0$.

When $\boldsymbol{X}_i = \boldsymbol{0}$, the expected score curve for the reference group is denoted by

$$E_j^{(R)}(\boldsymbol{\theta}) = \sum_{g=0}^{G-1} u \cdot P_{jg}(\boldsymbol{\theta}), \tag{4.2}$$

and when $\boldsymbol{X}_i \neq \mathbf{0}$, the expected score curve for the focal group is denoted by

$$E_j^{(F)}(\boldsymbol{\theta}) = \sum_{g=0}^{G-1} u \cdot P_{jg}(\boldsymbol{\theta}). \tag{4.3}$$

Then the wABC for the reference group is

$$\begin{aligned} \text{wABC}_{R_j} &= \int_{\theta} |E_j^{(R)}(\boldsymbol{\theta}) - E_j^{(F)}(\boldsymbol{\theta})| \cdot f(\boldsymbol{\theta}|\boldsymbol{\mu}^{(R)}, \boldsymbol{\Sigma}^{(R)}) d\boldsymbol{\theta} \\ &= \sum_{m=1}^{M} |E_j^{(R)}(\boldsymbol{q}_m) - E_j^{(F)}(\boldsymbol{q}_m)| \cdot f(\boldsymbol{q}_m|\boldsymbol{\mu}^{(R)}, \boldsymbol{\Sigma}^{(R)}), \end{aligned} \tag{4.4}$$

and the wABC for the focal group is

$$\begin{aligned} \text{wABC}_{F_j} &= \int_{\theta} |E_j^{(R)}(\boldsymbol{\theta}) - E_j^{(F)}(\boldsymbol{\theta})| \cdot f(\boldsymbol{\theta}|\boldsymbol{\mu}^{(F)}, \boldsymbol{\Sigma}^{(F)}) d\boldsymbol{\theta} \\ &= \sum_{m=1}^{M} |E_j^{(R)}(\boldsymbol{q}_m) - E_j^{(F)}(\boldsymbol{q}_m)| \cdot f(\boldsymbol{q}_m|\boldsymbol{\mu}^{(F)}, \boldsymbol{\Sigma}^{(F)}), \end{aligned} \tag{4.5}$$

where $\boldsymbol{q}_m$ is the Gaussian rectangular quadrature nodes to approximate the integral over $\boldsymbol{\theta}$.

Then the wABC for item j is given by

$$\text{wABC}_j = \text{wABC}_{F_j} * (N_F/N_{\text{Total}}) + \text{wABC}_{R_j} * (N_R/N_{\text{Total}}). \tag{4.6}$$

The wABC for two types of covariate designs are detailed in Tables 4.3 and 4.4. In Table 4.3, we present the wABC for the design with two binary covariates. The first row reports the DIF size between the reference group and the focal group, where individuals have $X = (1, 0)$—that is, a value of 1 for the first binary covariate and 0 for the second. Similarly, the second row details the DIF size for individuals in the focal group with $X = (0, 1)$, where the first covariate is 0 and the second is 1. The third row describes the DIF size for individuals with $X = (1, 1)$, indicating a value of 1 for both covariates.

The wABC for the design incorporating one binary and one continuous covariate is detailed in Table 4.4. For the continuous covariate (covariate 2), which is standardized, individuals around the 50th percentile have a zero value for covariate 2. This is because the distribution of covariate 2 is nearly symmetric. Individuals at lower quantiles have negative

values for covariate 2, while those at higher quantiles have positive values. In Table 4.4, the DIF sizes at six quantile levels — 0.1%, 10%, 25%, 75%, 90%, and 99.9% — are reported.

Table 4.3: Study I wABC for Two Binary Covariates

Condition	Item	3	4	5	6	9	10	11	12
No Impact	Covariate 1	0.173	0.134		0.155	0.350	0.507		0.431
	Covariate 2		0.134	0.198	0.155		0.507	0.375	0.431
	Covariate 1 + 2	0.173	0.272	0.198	0.305	0.350	0.944	0.375	0.829
Impact 1	Covariate 1	0.164	0.133		0.145	0.325	0.487		0.416
	Covariate 2		0.131	0.195	0.149		0.496	0.371	0.421
	Covariate 1 + 2	0.159	0.257	0.189	0.275	0.316	0.875	0.355	0.769

Table 4.4: Study I wABC for One Binary and One Continuous Covariate

Condition	Item	3	4	5	6	9	10	11	12
No Impact	Covariate 1	0.173	0.134		0.155	0.350	0.507		0.431
	Covariate 2 (0.1%)		0.171	0.265	0.208		0.629	0.525	0.530
	Covariate 2 (10%)		0.086	0.132	0.103		0.329	0.261	0.274
	Covariate 2 (25%)		0.043	0.066	0.051		0.168	0.130	0.139
	Covariate 2 (75%)		0.044	0.066	0.051		0.170	0.128	0.142
	Covariate 2 (90%)		0.088	0.131	0.102		0.339	0.252	0.285
	Covariate 2 (99.9%)		0.179	0.263	0.205		0.663	0.490	0.568
Impact 2	Covariate 1	0.164	0.133		0.145	0.325	0.487		0.416
	Covariate 2 (0.1%)		0.170	0.264	0.206		0.626	0.523	0.528
	Covariate 2 (10%)		0.085	0.132	0.103		0.326	0.261	0.272
	Covariate 2 (25%)		0.043	0.066	0.051		0.167	0.130	0.139
	Covariate 2 (75%)		0.044	0.066	0.051		0.169	0.127	0.142
	Covariate 2 (90%)		0.088	0.131	0.101		0.336	0.250	0.284
	Covariate 2 (99.9%)		0.179	0.262	0.204		0.661	0.490	0.567

I conducted the simulation study using self-written R code. Initially, for each condition, I generated sets of item parameters, DIF parameters, and impact parameters, which were consistently used across all replications. Random sets of latent traits and responses were generated for each simulatee in every replication. Each condition included 25 replications. Throughout each replication, I executed the algorithm with a series of tuning parameters. For conditions with a small sample size of 500, I adjusted the tuning parameters to range

from 10 to 35, increasing in increments of 5. For larger sample size conditions of 2000, I set the tuning range from 25 to 50, also increasing in increments of 5.

I calculated starting values using the mirt package in R for each replication. The starting values for the item parameters $\boldsymbol{a}$ and $\boldsymbol{d}$ were estimated in a 2PL model computed by the mirt function, employing all simulated individuals in the reference group. To obtain the starting value for the DIF caused by the first covariate $\boldsymbol{\beta}_1$, all simulatees with the covariate $X_1 = 1$ were used, regardless of their values for covariate 2. A mirt model was fitted on the subsample, and the parameter $\boldsymbol{d}$ estimates from this model, minus the $\boldsymbol{d}$ estimates from the reference group model, were used as the starting value for $\boldsymbol{\beta}_1$. Similarly, starting values for $\boldsymbol{\beta}_2$ were determined by employing all simulatees with $X_2 = 1$, irrespective of covariate 1 values. The difference in $\boldsymbol{d}$ estimates between this model and the reference model was taken as the starting value for $\boldsymbol{\beta}_2$. For no-impact conditions, starting values for the impact parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\omega}$/ $\boldsymbol{\phi}$ are fixed to be $\mathbf{0}$ and they are not estimated in the analysis. For conditions with impact, starting values were set to $\boldsymbol{\kappa}_0 = 0.5$ and $\boldsymbol{\omega}_0 = 0.2$ and $\boldsymbol{\phi}_0 = 0.2$.

A comparison study was conducted using the regDIF package, as documented in Belzak and Bauer (2024). In regDIF, only unidimensional models can be fitted; therefore, I fitted two separate unidimensional models for the two-dimensional data, one for each latent trait. By default, the algorithm in regDIF computes 100 tuning parameter values, starting with a value large enough so that all DIF parameter estimates are equal to zero. To conserve computing resources, the number of tuning parameters (num.tau) was reduced to 10, which is close to the number used in our algorithm. In regDIF, users are not required to input any starting values, and there is no mechanism to restrict the function to detecting only uniform DIF; instead, both intercept and slope DIF are evaluated together in all detections by regDIF. Therefore, in the uniform DIF setting, I allowed the regDIF function to detect DIF on both slope and intercept, but only counted DIF detected on the intercept when calculating the power and type I error of the method.

Type I error rates and power are vital metrics for evaluating the performance of the proposed method. These metrics are presented across various sample sizes and DIF proportions in Tables 4.5 and 4.6, as well as Figures 4.1 and 4.2. Type I error refers to the detection of a nonexistent DIF effect in the sample model, which does not exist in the population-

generating model. It is calculated by dividing the number of wrongly flagged DIF items by the total test length, then averaging over 25 replications. Conversely, power measures the method's ability to correctly identify a true DIF effect (intercept or slope) that is present in both the population-generating model and the sample's estimated model. It is calculated by dividing the number of correctly flagged DIF items by the total test length, then averaging over 25 replications. The objective of DIF detection is to obtain results that maintain low Type I error rates while maximizing power.

The first eight rows of the tables report outcomes where there is no impact: the first four detail two independent binary covariates, and the following four describe a scenario with correlated binary and continuous covariates. Rows 9-12 offer insights into two independent binary covariates with impact simulated via Approach I (Cholesky decomposition method), while rows 13-16 showcase results for correlated binary and continuous covariates with impact assessed through Approach II (covariance regression method). Each of these conditions varies the sample size (500 vs. 2000) and the DIF proportion (33% vs. 66%), creating four unique sets of results per condition, summing up to 16 different scenarios.

In each simulation condition, both tables and figures provide a report on the Type I error rates and power for each covariate, as well as the omnibus level. At the omnibus level, an item is identified as exhibiting DIF if it is detected on either covariate 1 or covariate 2. There are instances where an item is supposed to exhibit DIF on only one covariate but the algorithm flags the item as having DIF on both covariates. In this case, the detection is deemed correct at omnibus level (because the item indeed has DIF) but incorrect at a certain covariate level. For example, if covariate 2 is erroneously marked as having DIF on item 3, it is considered a Type I error at the covariate 2 level. However, this is justified at the omnibus level because item 3 does exhibit DIF, albeit only on covariate 1 and not on covariate 2. Therefore, the Type I error rate for each covariate can be higher than the omnibus Type I error rate under certain conditions. For instance, in the scenario with no impact, two independent binary covariates, a sample size of 2000, and 66% DIF, the Type I error rate for covariate 2 (0.04) exceeds the omnibus Type I error rate (0.03). Similarly, the power for a single covariate does not necessarily have to be lower than the omnibus power for the same reasons that affect Type I error rates.

Tables 4.5 and 4.6 show that the power of the lasso EMM method is consistently high, and the type I error is well-controlled. An increase in sample size typically results in higher power and type I error rates. Although the regDIF method exhibits strong power, it produces higher type I error rates, especially when the DIF proportion is high. Considering we selected only 10 tuning parameters within the default range for their method, enhancing the regDIF results by using more tuning parameters could be feasible. However, regularization methods often see a decrease in power as type I error is reduced. Thus, it remains challenging for the regDIF method to achieve higher power while maintaining control over type I error. Since controlling type I error is critical in DIF detection, any improvements to regDIF would prioritize this metric. Even if the type I error of regDIF could be controlled to the same level as our method, without any drop in power after model fine-tuning, the proposed lasso-EMM algorithm would still likely yield higher power than regDIF in these simulation settings, as demonstrated in Table 4.6.

It is also observed that in conditions where the two covariates are correlated, the power drops compared to those conditions where the two covariates are independent. There are two reasons for this drop. First, in linear regression, the issue that arises when independent variables are correlated is called multicollinearity. Multicollinearity can lead to unstable parameter estimates and inflated standard errors. Extending this concept to DIF detection, correlated covariates may also make the DIF effect harder to detect, thus potentially leading to false negatives. Second, from Table 4.4, we can see that the generated DIF size for 80% of individuals is lower than that in Table 4.3. Therefore, this may also result in lower power compared to the conditions with independent covariates.

Some outcomes from our model also show inflated Type I error rates, suggesting that our model selection criteria need refinement. The Generalized Information Criterion (GIC) could be a valuable tool for future adjustments. Overall, our algorithm outperforms regDIF in controlling Type I error while still providing strong power in DIF detection.

Tables 4.7 and 4.8 present the Mean Absolute Error (MAE) for Item Parameter Recovery and DIF Parameter Recovery, respectively, using our method. In Table 4.8, the means are computed across the DIF items for the three boundary parameters of a single item, denoted as $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13})$ for covariate 1 and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23})$ for covaraite 2. Given that

Table 4.5: Study I Type I error

Impact	Covariates	N	DIF%	Omnibus	Covariate 1	Covariate 2	regDIF
None	Two Independent Binary	500	33%	0.020	0.010	0.012	0.118
			66%	0.060	0.033	0.047	0.325
		2000	33%	0.075	0.040	0.048	0.208
			66%	0.030	0.013	0.040	0.483
	Correlated Binary+ Continuous	500	33%	0.065	0.060	0.008	0.075
			66%	0.042	0.056	0.014	0.183
		2000	33%	0.04	0.04	0.000	0.15
			66%	0.080	0.087	0.000	0.3
Impact	Two Independent Binary	500	33%	0.095	0.055	0.044	0.10
			66%	0.104	0.076	0.042	0.316
		2000	33%	0.150	0.085	0.065	0.208
			66%	0.060	0.033	0.020	0.417
	Correlated Binary+ Continuous	500	33%	0.030	0.030	0.004	0.113
			66%	0.010	0.007	0.007	0.091
		2000	33%	0.030	0.025	0.008	0.163
			66%	0.050	0.053	0.013	0.159

Table 4.6: Study I Power

Impact	Covariates	N	DIF%	Omnibus	Covariate 1	Covariate 2	regDIF
None	Two Independent Binary	500	33%	0.41	0.35	0.50	0.50
			66%	0.555	0.5	0.553	0.513
		2000	33%	0.94	0.90	0.88	0.925
			66%	0.79	0.707	0.74	0.658
	Correlated Binary+ Continuous	500	33%	0.33	0.31	0.34	0.40
			66%	0.552	0.542	0.451	0.533
		2000	33%	0.84	0.84	0.62	0.825
			66%	0.82	0.805	0.761	0.75
Impact	Two Independent Binary	500	33%	0.70	0.63	0.74	0.533
			66%	0.489	0.375	0.528	0.508
		2000	33%	0.93	0.88	0.94	0.833
			66%	0.82	0.687	0.813	0.592
	Correlated Binary+ Continuous	500	33%	0.35	0.34	0.3	0.575
			66%	0.26	0.307	0.193	0.409
		2000	33%	0.65	0.65	0.56	0.675
			66%	0.67	0.673	0.56	0.704

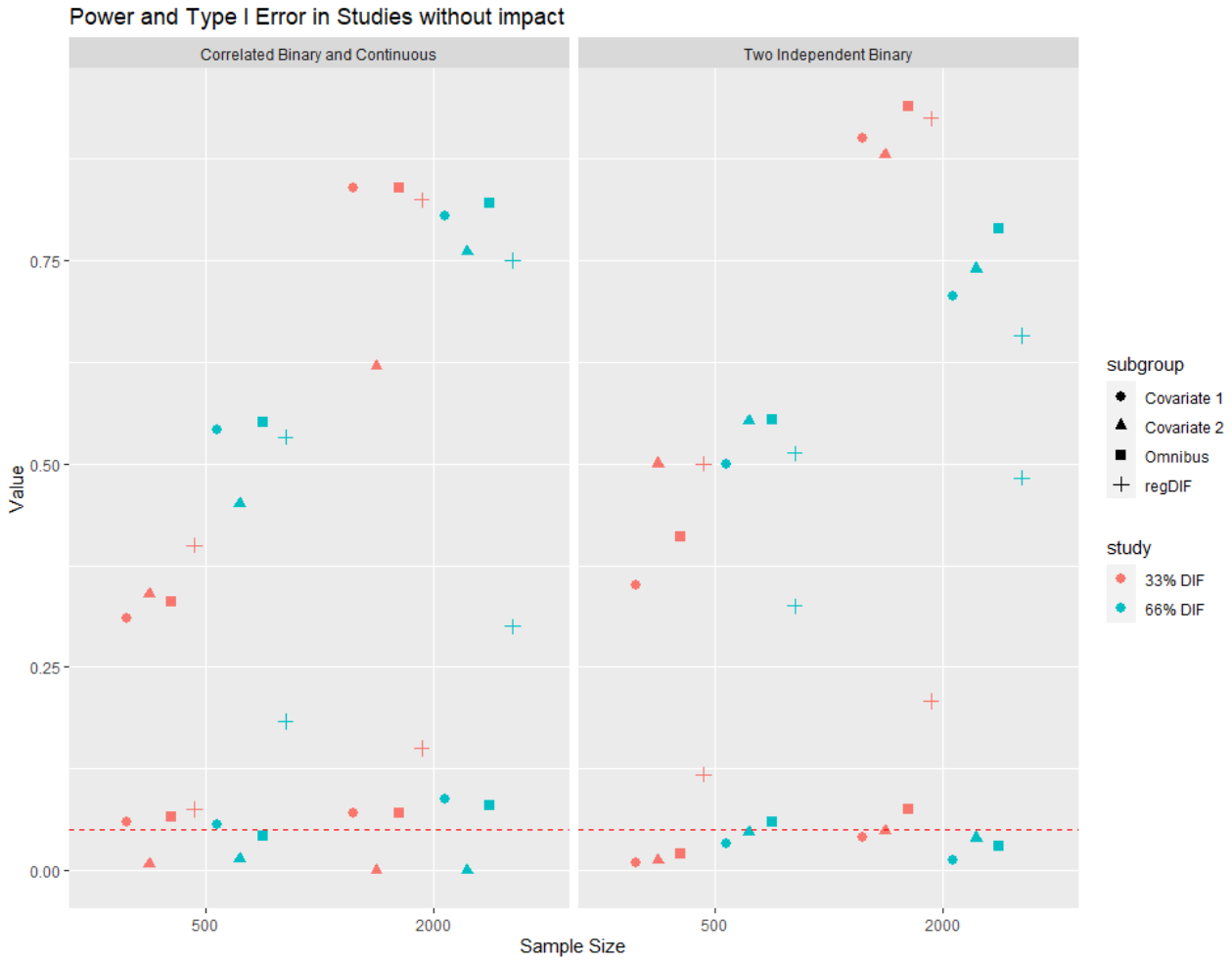

Figure 4.1: Study I: Power and Type I Error in Conditions without impact

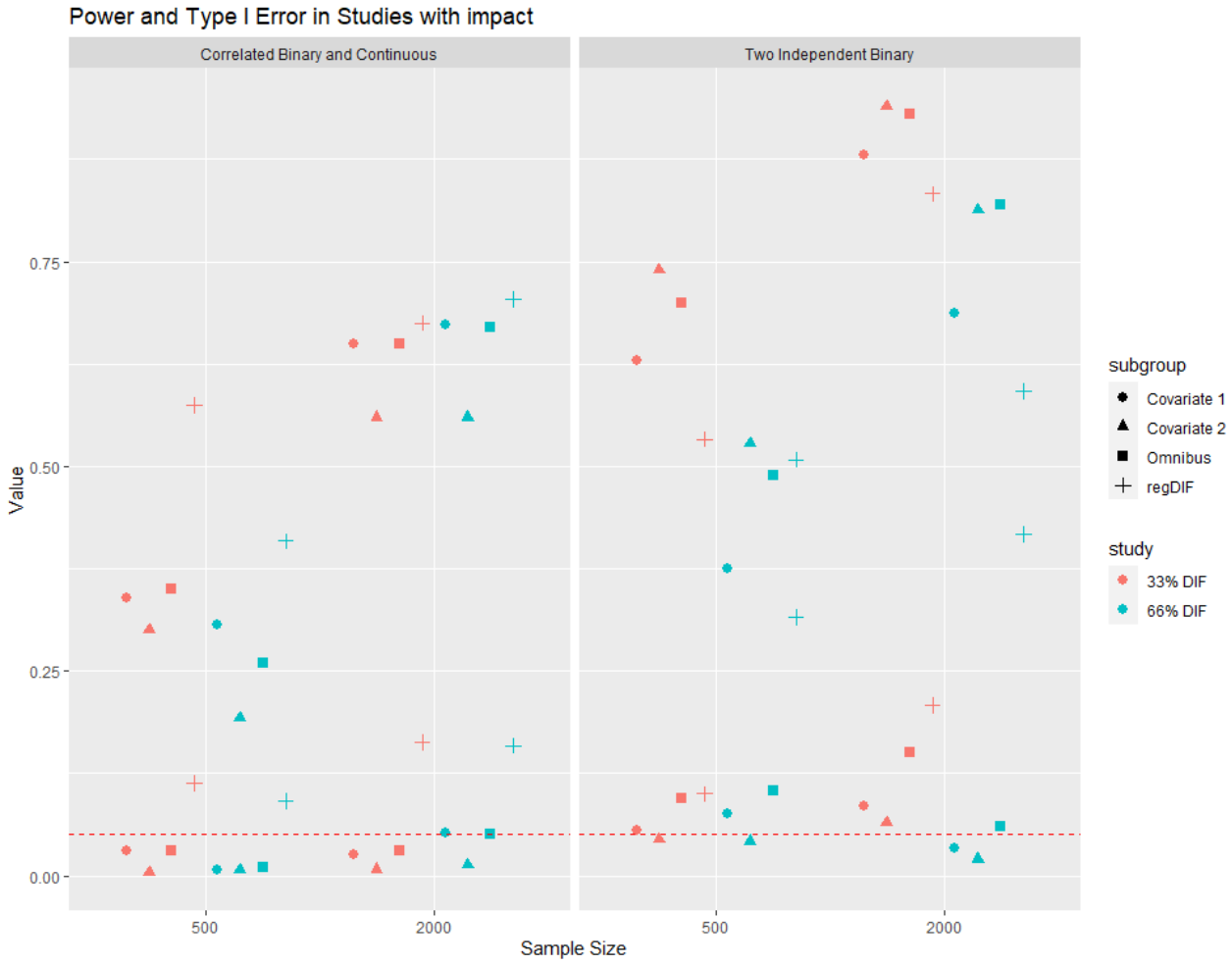

Figure 4.2: Study I: Power and Type I Error in Conditions with impact

the first and last six items possess different DIF magnitudes, the DIF parameter recovery is reported separately for each group. The values are calculated exclusively for items that are correctly identified as having DIF. A missing value indicates that no items were successfully detected as DIF across all 25 replications. Tables 4.9 and 4.10 detail the MAE for Impact Parameter Recovery for our two impact estimation methods.

Table 4.7: Study I Mean Absolute Error for lasso EMM Item Parameter Recovery

Impact	Covariates	N	DIF%	a_1	a_2	d_1	d_2	d_3
None	Two Independent Binary	500	33%	0.114	0.142	0.194	0.126	0.132
			66%	0.145	0.204	0.303	0.203	0.221
		2000	33%	0.044	0.082	0.175	0.073	0.094
			66%	0.062	0.094	0.251	0.119	0.146
	Correlated Binary+ Continuous	500	33%	0.112	0.131	0.167	0.146	0.185
			66%	0.166	0.187	0.209	0.154	0.155
		2000	33%	0.050	0.070	0.105	0.071	0.093
			66%	0.088	0.097	0.173	0.101	0.122
Impact	Two Independent Binary	500	33%	0.172	0.191	0.255	0.215	0.223
			66%	0.208	0.229	0.265	0.245	0.244
		2000	33%	0.115	0.148	0.139	0.107	0.154
			66%	0.134	0.162	0.186	0.127	0.131
	Correlated Binary+ Continuous	500	33%	0.122	0.134	0.105	0.128	0.109
			66%	0.159	0.168	0.215	0.145	0.202
		2000	33%	0.11	0.119	0.146	0.114	0.094
			66%	0.097	0.108	0.113	0.115	0.123

From the results, it can be observed that parameter recovery is generally good, and an increase in sample size is seen to significantly reduce estimation bias for all parameters. Estimation of the impact effect on the covariance matrix is challenging. Different impact sizes are generated for our two impact estimation methods, making it difficult to directly compare the results from Table 4.9 and 4.10 and conclude which method performs better. For this reason, I performed a separate simulation study for the second impact estimation method, where $\phi = \begin{pmatrix} 0.3 & 0.3 \\ 0.3 & 0.3 \\ 0.3 & 0.3 \end{pmatrix}$ was used to achieve a similar impact size as that obtained

Table 4.8: Study I Mean Absolute Error for DIF Parameter Recovery

Impact	Covariates	N	DIF%	Small DIF (item 1-6)		Large DIF (item 7-12)	
				$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
None	Two Independent Binary	500	33%	-	-	0.259	0.319
			66%	0.045	0.061	0.219	0.216
		2000	33%	0.091	0.089	0.113	0.092
			66%	0.058	0.075	0.132	0.111
	Correlated Binary+ Continuous	500	33%	0.086	0.063	0.222	0.155
			66%	0.130	0.048	0.233	0.194
		2000	33%	0.067	0.049	0.108	0.109
			66%	0.091	0.056	0.121	0.116
Impact	Two Independent Binary	500	33%	0.093	0.190	0.255	0.215
			66%	0.012	0.036	0.204	0.188
		2000	33%	0.092	0.109	0.148	0.105
			66%	0.030	0.036	0.099	0.093
	Correlated Binary+ Continuous	500	33%	0.018	-	0.151	0.161
			66%	-	-	0.128	0.083
		2000	33%	0.047	0.030	0.120	0.115
			66%	0.082	-	0.113	0.106

Table 4.9: Study I Mean Absolute Error for Approach 1 Impact Parameter Recovery

Covariates	N	DIF%	κ_1	κ_2	ω_1	ω_2	ω_3
Two Independent Binary	500	33%	0.099	0.028	0.101	0.118	0.094
		66%	0.109	0.059	0.098	0.110	0.113
	2000	33%	0.065	0.010	0.098	0.073	0.099
		66%	0.054	0.026	0.105	0.100	0.108

Table 4.10: Study I Mean Absolute Error for Approach 2 Impact Parameter Recovery

Covariates	N	DIF%	κ_1	κ_2	ϕ_1	ϕ_2	ϕ_3
Correlated Binary + Continuous	500	33%	0.139	0.116	0.152	0.139	0.163
		66%	0.119	0.173	0.137	0.176	0.146
	2000	33%	0.090	0.076	0.138	0.134	0.131
		66%	0.076	0.092	0.141	0.140	0.139

using $\omega = \begin{pmatrix} 0.15 & 0.15 & 0.15 \\ 0.15 & 0.15 & 0.15 \end{pmatrix}$ in the first method. We fix the item and DIF paramteres and estimate the impact paramaters using two methods. Only condition $N = 2000$ in table 4.10 was studied. The results of 25 replications are reported below.

Table 4.11: Mean Absolute Error for Large Impact Parameters Estimated by Approach 2

Covariates	N	DIF%	κ_1	κ_2	ϕ_1	ϕ_2	ϕ_3
Correlated Binary + Continuous	2000	33%	0.095	0.075	0.239	0.224	0.183
		66%	0.070	0.077	0.223	0.229	0.173

Comparing Table 4.11 and the lower half of Table 4.10, it can be observed that the MAEs of parameter $\boldsymbol{\phi}$ are larger when the true parameter $\boldsymbol{\phi}$ becomes larger. Additionally, comparing Table 4.11 and the lower half of Table 4.9, it can be noticed that the MAEs of parameter $\boldsymbol{\phi}$ are larger than the MAEs of parameter $\boldsymbol{\omega}$. To determine if this means that impact estimation approach I outperforms approach II, we calculate the mean absolute bias of each element in the reconstructed covariance matrix by rebuilding the covariance matrix at four different levels using the estimated impact parameters. We summarize this result in three tables corresponding to the conditions and methods in Tables 4.9, 4.10, and 4.11, respectively.

Looking at table 4.12 and 4.14, it can be concluded that the accuracy of the two impact methods is comparable under the condition $N = 2000$, 33% and 66% DIF. Although neither method produced very accurate estimates of the impact on the covariance matrix, this did not significantly affect the DIF detection results. I also selected one condition—$N = 2000$, 33% DIF, with correlated covariates and impact—where I reran the simulation for 10 replications by fixing the impact parameters at their true values. The resulting power and Type I error rates for the DIF detection were (0.038, 0.7) and (0.025, 0.675) respectively for the two covariates, which are comparable to those obtained with biased impact parameter estimates.

Covariate Combination	True Covariance Matrix	N=500 33% DIF	N=500 66% DIF	N=2000 33% DIF	N=2000 66% DIF
(0,0)	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.004 \\ 0.004 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.004 \\ 0.004 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.004 \\ 0.004 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 0.003 \\ 0.003 & 0 \end{pmatrix}$
(1,0)	$\begin{pmatrix} 1.35 & 0.675 \\ 0.675 & 1.35 \end{pmatrix}$	$\begin{pmatrix} 0.271 & 0.125 \\ 0.125 & 0.177 \end{pmatrix}$	$\begin{pmatrix} 0.259 & 0.151 \\ 0.151 & 0.261 \end{pmatrix}$	$\begin{pmatrix} 0.249 & 0.129 \\ 0.129 & 0.208 \end{pmatrix}$	$\begin{pmatrix} 0.273 & 0.140 \\ 0.140 & 0.266 \end{pmatrix}$
(0,1)	$\begin{pmatrix} 1.35 & 0.675 \\ 0.675 & 1.35 \end{pmatrix}$	$\begin{pmatrix} 0.215 & 0.107 \\ 0.107 & 0.210 \end{pmatrix}$	$\begin{pmatrix} 0.222 & 0.133 \\ 0.133 & 0.225 \end{pmatrix}$	$\begin{pmatrix} 0.223 & 0.119 \\ 0.119 & 0.211 \end{pmatrix}$	$\begin{pmatrix} 0.235 & 0.111 \\ 0.111 & 0.220 \end{pmatrix}$
(1,1)	$\begin{pmatrix} 1.822 & 0.911 \\ 0.911 & 1.822 \end{pmatrix}$	$\begin{pmatrix} 0.593 & 0.237 \\ 0.237 & 0.485 \end{pmatrix}$	$\begin{pmatrix} 0.585 & 0.280 \\ 0.280 & 0.593 \end{pmatrix}$	$\begin{pmatrix} 0.583 & 0.256 \\ 0.256 & 0.518 \end{pmatrix}$	$\begin{pmatrix} 0.624 & 0.292 \\ 0.292 & 0.599 \end{pmatrix}$

Table 4.12: Mean absolute bias on each element of reconstructed covariance matrix corresponding to Table 4.9

Covariate Combination	True Covariance Matrix	N=500 33% DIF	N=500 66% DIF	N=2000 33% DIF	N=2000 66% DIF
(0,0)	$\begin{pmatrix} 1.023 & 0.523 \\ 0.523 & 1.023 \end{pmatrix}$	$\begin{pmatrix} 0.143 & 0.172 \\ 0.172 & 0.143 \end{pmatrix}$	$\begin{pmatrix} 0.023 & 0.113 \\ 0.113 & 0.023 \end{pmatrix}$	$\begin{pmatrix} 0.023 & 0.113 \\ 0.113 & 0.023 \end{pmatrix}$	$\begin{pmatrix} 0.023 & 0.119 \\ 0.119 & 0.023 \end{pmatrix}$
(1,0)	$\begin{pmatrix} 1.09 & 0.59 \\ 0.59 & 1.09 \end{pmatrix}$	$\begin{pmatrix} 0.200 & 0.146 \\ 0.146 & 0.216 \end{pmatrix}$	$\begin{pmatrix} 0.083 & 0.065 \\ 0.065 & 0.085 \end{pmatrix}$	$\begin{pmatrix} 0.088 & 0.049 \\ 0.049 & 0.088 \end{pmatrix}$	$\begin{pmatrix} 0.089 & 0.053 \\ 0.053 & 0.090 \end{pmatrix}$
(0,1)	$\begin{pmatrix} 1.09 & 0.59 \\ 0.59 & 1.09 \end{pmatrix}$	$\begin{pmatrix} 0.192 & 0.155 \\ 0.155 & 0.205 \end{pmatrix}$	$\begin{pmatrix} 0.081 & 0.103 \\ 0.103 & 0.070 \end{pmatrix}$	$\begin{pmatrix} 0.085 & 0.052 \\ 0.052 & 0.085 \end{pmatrix}$	$\begin{pmatrix} 0.088 & 0.054 \\ 0.054 & 0.089 \end{pmatrix}$
(1,1)	$\begin{pmatrix} 1.202 & 0.703 \\ 0.703 & 1.202 \end{pmatrix}$	$\begin{pmatrix} 0.302 & 0.189 \\ 0.189 & 0.349 \end{pmatrix}$	$\begin{pmatrix} 0.166 & 0.104 \\ 0.104 & 0.169 \end{pmatrix}$	$\begin{pmatrix} 0.192 & 0.063 \\ 0.063 & 0.192 \end{pmatrix}$	$\begin{pmatrix} 0.198 & 0.058 \\ 0.058 & 0.200 \end{pmatrix}$

Table 4.13: Mean absolute bias on each element of reconstructed covariance matrix corresponding to Table 4.10

Covariate Combination	True Covariance Matrix	N=2000 33% DIF	N=2000 66% DIF
(0,0)	$\begin{pmatrix} 1.09 & 0.59 \\ 0.59 & 1.09 \end{pmatrix}$	$\begin{pmatrix} 0.090 & 0.102 \\ 0.102 & 0.090 \end{pmatrix}$	$\begin{pmatrix} 0.079 & 0.089 \\ 0.089 & 0.079 \end{pmatrix}$
(1,0)	$\begin{pmatrix} 1.36 & 0.86 \\ 0.86 & 1.36 \end{pmatrix}$	$\begin{pmatrix} 0.337 & 0.146 \\ 0.146 & 0.340 \end{pmatrix}$	$\begin{pmatrix} 0.294 & 0.123 \\ 0.123 & 0.291 \end{pmatrix}$
(0,1)	$\begin{pmatrix} 1.36 & 0.86 \\ 0.86 & 1.36 \end{pmatrix}$	$\begin{pmatrix} 0.321 & 0.134 \\ 0.134 & 0.329 \end{pmatrix}$	$\begin{pmatrix} 0.284 & 0.116 \\ 0.116 & 0.287 \end{pmatrix}$
(1,1)	$\begin{pmatrix} 1.81 & 1.31 \\ 1.31 & 1.81 \end{pmatrix}$	$\begin{pmatrix} 0.698 & 0.521 \\ 0.521 & 0.722 \end{pmatrix}$	$\begin{pmatrix} 0.618 & 0.445 \\ 0.445 & 0.607 \end{pmatrix}$

Table 4.14: Mean absolute bias on each element of reconstructed covariance matrix corresponding to Table 4.11

4.2 Simulation Study II

In the second simulation study, the algorithm's performance in detecting non-uniform DIF associated with two different types of demographic variables within a 2DGRM is assessed. The test length remains consistent with Simulation Study I, and the item parameters for the graded response model are identical, as indicated in Table 4.24.

The covariate design follows the same structure as Simulation Study I. Within each design, I manipulated two levels of sample size (500 and 2000) and two levels of DIF item proportion (33% and 66%), mirroring the approach in Study I. Additionally, I simulated two levels of DIF magnitude for each condition.

In line with our previous experience documented in C. Wang et al. (2023), a slightly smaller DIF magnitude was simulated for the non-uniform study. The details of these DIF parameters are provided in Table 4.15.

Two levels of impact were generated using the previously described impact parameterization methods, and the impact parameters were estimated using each method accordingly.

In the second simulation study, I continue to examine the wABC for the same two covariate designs, now detailed in Tables 4.16 and 4.17 with updated values. As in the first study, Table 4.16 outlines the wABC results for a configuration of two binary covariates. The first row documents the differential item functioning (DIF) size between the reference

Table 4.15: Study II Simulated True DIF Parameters

	DIF %		γ_1		γ_2		β_1		β_2		β_3	
Item	33%	66%	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
1												
2												
3	√	√	-.4				.25		.25		.25	
4	√	√	-.4	-.4			.25	.25	.25	.25	.25	.25
5		√		-.4				.25		.25		.25
6		√	-.4	-.4			.25	.25	.25	.25	.25	.25
7												
8												
9	√	√			-.6		.6		.6		.6	
10	√	√			-.6	-.6	.6	.6	.6	.6	.6	.6
11		√				-.6		.6		.6		.6
12		√			-.6	-.6	.6	.6	.6	.6	.6	.6

and focal groups when individuals exhibit $X = (1, 0)$—where the first binary covariate is set to 1 and the second to 0. The second row again provides the DIF size for the focal group configuration of $X = (0, 1)$, with reversed values of the covariates. Lastly, the third row details the DIF outcomes when both covariates are at their maximal value, $X = (1, 1)$.

In Table 4.16, some units in the first row are identical to those in the third row (units for items 3, 5, 9, 11). This occurs because these items exhibit DIF on only one covariate (either covariate 1 or covariate 2). The third row can be understood as representing the combined DIF effect of covariate 1 and covariate 2. The DIF magnitude for these items is slightly different in the fourth and sixth rows. For example, with item 3, covariate 2 does not introduce additional DIF beyond what is already contributed by the combination of covariate 1 and covariate 2, but it does impact the item, resulting in a slightly different DIF size (0.134 for covariate 1 only and 0.126 for covariate 1 + covariate 2).

The analysis of wABC for a model with one binary and one continuous covariate is presented again in Table 4.17, mirroring the approach from study I. For the standardized continuous covariate (covariate 2), the median position (50th percentile) represents a zero value due to the symmetrical distribution of covariate 2. Values for covariate 2 are negative at lower percentiles and positive at higher ones. Table 4.17 documents the DIF sizes across

Table 4.16: Study II Non-uniform DIF Magnitude for Two Binary Covariates

Condition	Item	3	4	5	6	9	10	11	12
No Impact	Covariate 1	0.134	0.083		0.109	0.276	0.369		0.292
	Covariate 2		0.083	0.154	0.109		0.369	0.288	0.292
	Covariate 1 + 2	0.134	0.170	0.154	0.229	0.276	0.772	0.288	0.640
Impact 1	Covariate 1	0.124	0.079		0.099	0.247	0.341		0.272
	Covariate 2		0.084	0.158	0.108		0.372	0.295	0.295
	Covariate 1 + 2	0.126	0.170	0.153	0.211	0.248	0.747	0.276	0.618

different quantiles—specifically at the 0.1%, 10%, 25%, 75%, 90%, and 99.9% levels.

Table 4.17: Study II Non-uniform DIF Magnitude for One Binary and One Continuous Covariate

Condition	Item	3	4	5	6	9	10	11	12
No Impact	Covariate 1	0.134	0.083		0.109	0.276	0.369		0.292
	Covariate 2 (0.1%)		0.099	0.174	0.125		0.347	0.327	0.301
	Covariate 2 (10%)		0.051	0.091	0.064		0.193	0.175	0.162
	Covariate 2 (25%)		0.026	0.047	0.033		0.102	0.090	0.084
	Covariate 2 (75%)		0.027	0.049	0.034		0.113	0.094	0.090
	Covariate 2 (90%)		0.054	0.100	0.070		0.235	0.190	0.186
	Covariate 2 (99.9%)		0.111	0.208	0.147		0.503	0.380	0.401
Impact 2	Covariate 1	0.124	0.079		0.099	0.247	0.341		0.272
	Covariate 2 (0.1%)		0.099	0.175	0.125		0.348	0.329	0.302
	Covariate 2 (10%)		0.052	0.093	0.066		0.197	0.179	0.166
	Covariate 2 (25%)		0.026	0.047	0.034		0.103	0.091	0.085
	Covariate 2 (75%)		0.026	0.048	0.034		0.111	0.093	0.089
	Covariate 2 (90%)		0.053	0.099	0.069		0.231	0.186	0.183
	Covariate 2 (99.9%)		0.111	0.208	0.147		0.504	0.381	0.401

The data generation process in this study is closely mirrored to that implemented in the first study. However, when the same method was used to determine starting values for the algorithm—specifically, using the differences between the single model for each group as initial values for the DIF parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$—convergence issues were encountered when all DIF parameters on slope and intercept were non-zero. To address this, I set all $\boldsymbol{\beta}$ parameters initially to zero, while non-zero values and a relatively small tuning parameter, $\eta = 5$, were assigned to all $\boldsymbol{\gamma}$ parameters. This model configuration is referred to as fit1.

This approach, which penalizes non-zero $\boldsymbol{\gamma}$ values, helps identify some anchor items for the $\boldsymbol{\gamma}$ parameter. The small value of the tuning parameter helps minimize Type II errors, meaning that true non-zero values of $\boldsymbol{\gamma}$ are less likely to be incorrectly shrunk to zero. These anchor items are then used in subsequent estimations with varying tuning values. For DIF detection, non-zero values are set for all $\boldsymbol{\beta}$ (specifically $\boldsymbol{\beta} = 0.4$), and the initial $\boldsymbol{\gamma}$ values calculated in fit1 are used. This strategy not only resolves the convergence issue but also prevents a loss of power in DIF detection due to the mis-specification of anchor items. Starting values for other parameters, including the item parameters $\boldsymbol{a}$ and $\boldsymbol{d}$, and the impact parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\omega}/\ \boldsymbol{\phi}$, are calculated in the same manner as in simulation study I.

The selection of tuning parameters for the algorithm mirrors that of Simulation Study I, using the same number of tuning parameters (10) for regDIF as in Study I. Type I error rates and power are once again reported to evaluate the performance of the proposed method and the regDIF comparison method in Tables 4.18 and 4.19, as well as in Figures 4.3 and 4.4. I report the results for power and Type I error in this non-uniform DIF study on an omnibus level: regardless of whether DIF is detected on the slope, intercept, or both, the item is marked as a DIF item.

From the results tables and figures, it can be observed that the type I error remains well-controlled, while good power is maintained by the proposed method in comparison with regDIF. However, in this non-uniform DIF study, a reduction in power compared to the previously uniform DIF study is noted, particularly in conditions involving impact. This is evident when comparing Table 4.6 with Table 4.19, where a significant reduction in power is observed in the lower half of the table, representing the conditions with impact.

The reduction in power is attributed to the smaller DIF magnitude simulated in this study, as observed through a comparison of the wABC tables for Study I and Study II, specifically concerning the intercept. The detection of DIF on the slope is found to be more challenging than on the intercept, with only 25-50% of DIF on the slope detectable in most replications, which means Power in detecting non-uniform DIF is primarily derived from successfully identifying DIF on the intercept. Therefore, a smaller DIF magnitude on the intercept results in a reduction in power in this study.

Table 4.18: Study II Type I error

Impact	Covariates	N	DIF%	Omnibus	Covariate 1	Covariate 2	regDIF
None	Two Independent Binary	500	33%	0.025	0.025	0.000	0.083
			66%	0.004	0.03	0.01	0.295
		2000	33%	0.08	0.050	0.040	0.208
			66%	0.13	0.11	0.03	0.341
	Correlated Binary+ Continuous	500	33%	0.025	0.010	0.015	0.025
			66%	0.05	0.01	0.04	0.068
		2000	33%	0.06	0.045	0.025	0.225
			66%	0.080	0.080	0.05	0.455
Impact	Two Independent Binary	500	33%	0.11	0.04	0.07	0.065
			66%	0.104	0.076	0.042	0.29
		2000	33%	0.14	0.115	0.035	0.235
			66%	0.17	0.13	0.04	0.36
	Correlated Binary+ Continuous	500	33%	0.075	0.035	0.040	0.015
			66%	0.010	0.01	0.00	0.05
		2000	33%	0.06	0.055	0.005	0.18
			66%	0.17	0.12	0.05	0.39

Table 4.19: Study II Power

Impact	Covariates	N	DIF%	Omnibus	Covariate 1	Covariate 2	regDIF
None	Two Independent Binary	500	33%	0.51	0.39	0.41	0.283
			66%	0.51	0.38	0.42	0.477
		2000	33%	0.98	0.73	0.76	0.767
			66%	0.955	0.690	0.870	0.711
	Correlated Binary+ Continuous	500	33%	0.41	0.25	0.25	0.13
			66%	0.285	0.140	0.20	0.25
		2000	33%	0.84	0.65	0.69	0.767
			66%	0.735	0.635	0.530	0.680
Impact	Two Independent Binary	500	33%	0.39	0.26	0.35	0.34
			66%	0.405	0.3	0.335	0.50
		2000	33%	0.79	0.68	0.70	0.61
			66%	0.76	0.575	0.685	0.515
	Correlated Binary+ Continuous	500	33%	0.29	0.15	0.15	0.13
			66%	0.18	0.10	0.15	0.155
		2000	33%	0.68	0.54	0.58	0.642
			66%	0.75	0.59	0.62	0.54

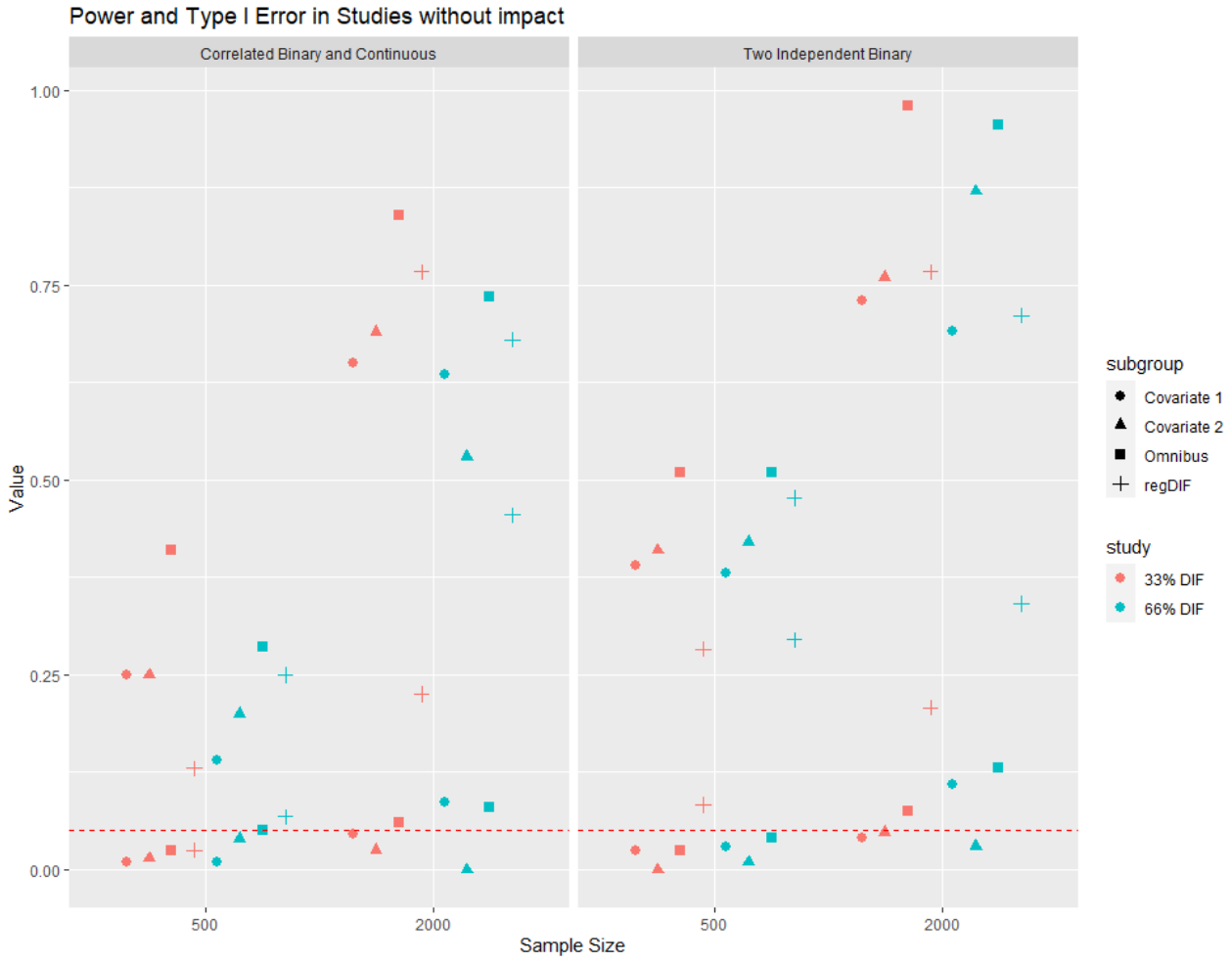

Figure 4.3: Study II: Power and Type I Error in Conditions without impact

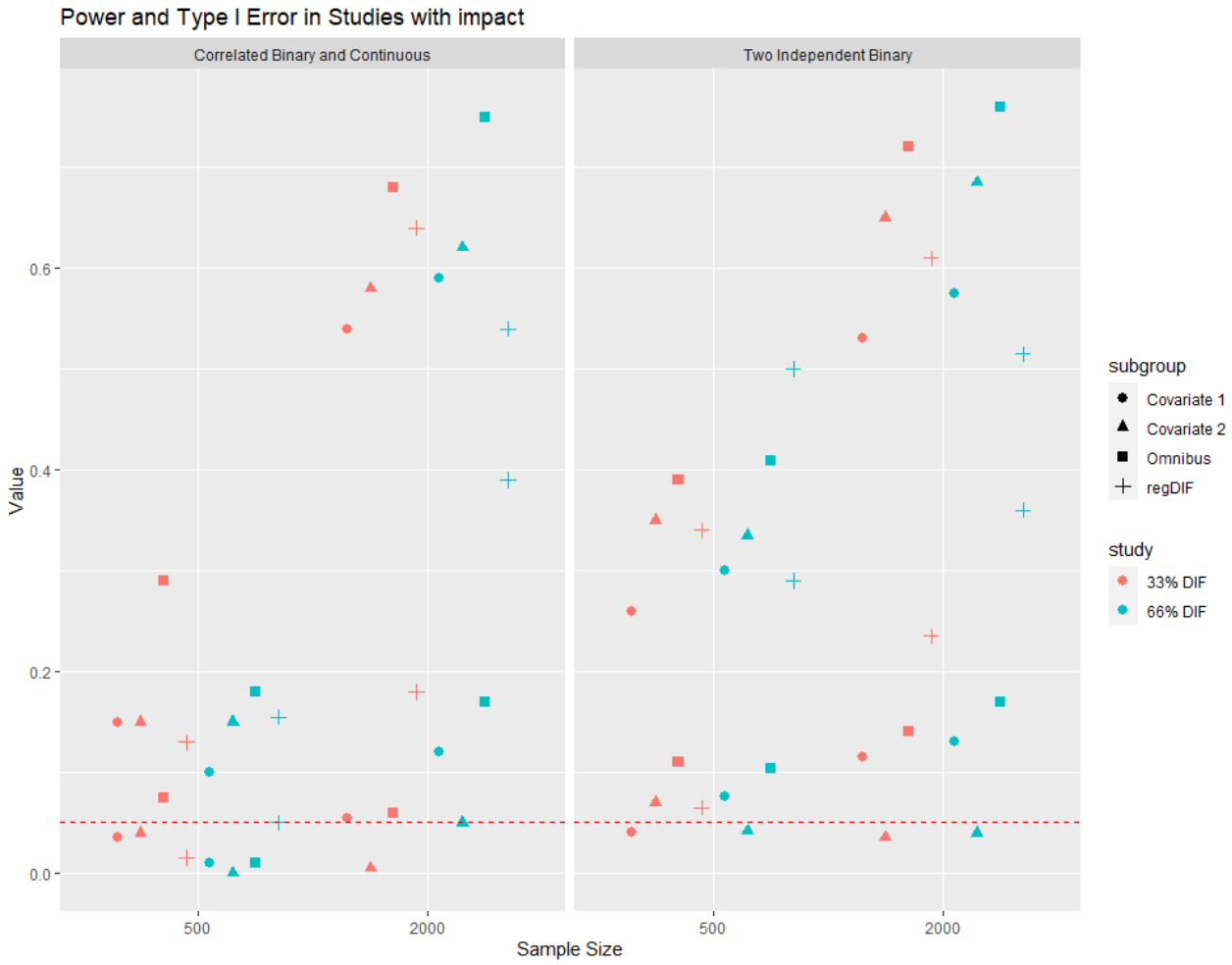

Figure 4.4: Study II: Power and Type I Error in Conditions with impact

In the presence of impact, the reduction in DIF size leads to a significant decrease in power, even when impact parameters are accurately predicted. In contrast, without impact, the reduction in power in this simulation is not significant. However, the upper half of Table 4.5 shows smaller overall values compared to Table 4.18, indicating that a reduction in power could be observed if the type I error in Table 4.18 is controlled to levels similar to Table 4.5 by employing different information criteria for model selection.

In the last column of Table 4.19, despite the suboptimal detection results of regDIF, a small reduction in power is observable between conditions without impact and those with impact. For instance, in the condition with two independent binary covariates, N = 2000, and 33% DIF, the power and type I error are 0.825 and 0.14 in the absence of impact, and 0.78 and 0.22 in its presence.

Same as before, I report the MAE for item parameter and DIF Parameter Recovery of our method (in Tables 4.20 and 4.21, respectively). In Table 4.8, no significant discrepancy was observed between the estimation bias for DIF on the two covariates, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Therefore, in Table 4.21, the results for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have been combined, and $\boldsymbol{\beta}$ is used to represent the two DIF parameters corresponding to both covariates by averaging $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for each condition. The same approach is applied to the parameter $\boldsymbol{\gamma}$. Both $\boldsymbol{\gamma}_1$ (for covariate 1) and $\boldsymbol{\gamma}_2$ (for covariate 2) are merged into a single $\boldsymbol{\gamma}$. Since the first six items with small DIF load exclusively on the first trait dimension, and items 7-12 with large DIF load exclusively on the second trait dimension, only one $\boldsymbol{\gamma}$ is reported for each DIF size group. Thus, in Table 4.21, $\boldsymbol{\gamma}_1$ represents the DIF parameter for both covariates on the first latent trait dimension, and $\boldsymbol{\gamma}_2$ represents the DIF parameter for both covariates on the second latent trait dimension.

From the results, it can be observed that parameter recovery is generally good, and as usual larger sample size yield better estimation. In Table 4.21, it can be observed that the bias for larger DIF is smaller than those for small DIF. This is because there are few DIF successully detected on item 1-6, and the MAEs of DIF parameters are only calculate for successfuly detected item, the DIF parameter estimates for small DIF are easily to be overestimated in our method. Not only DIF detection is harder on the discrimination parameters, the parameter estimation for $\boldsymbol{\gamma}$ is also harder. It can be observed from Table

Table 4.20: Study II Mean Absolute Error for Item Parameter Recovery

Impact	Covariates	N	DIF%	a_1	a_2	d_1	d_2	d_3
None	Two Independent Binary	500	33%	0.173	0.178	0.199	0.133	0.138
			66%	0.197	0.218	0.239	0.181	0.201
		2000	33%	0.087	0.104	0.115	0.078	0.102
			66%	0.073	0.097	0.121	0.109	0.117
	Correlated Binary+ Continuous	500	33%	0.166	0.179	0.174	0.152	0.191
			66%	0.178	0.197	0.218	0.152	0.161
		2000	33%	0.077	0.089	0.111	0.078	0.089
			66%	0.081	0.093	0.123	0.109	0.114
Impact	Two Independent Binary	500	33%	0.163	0.181	0.205	0.157	0.174
			66%	0.176	0.187	0.203	0.166	0.182
		2000	33%	0.145	0.107	0.122	0.109	0.114
			66%	0.102	0.118	0.133	0.104	0.118
	Correlated Binary+ Continuous	500	33%	0.177	0.189	0.172	0.156	0.178
			66%	0.167	0.175	0.205	0.154	0.191
		2000	33%	0.092	0.119	0.137	0.103	0.126
			66%	0.103	0.115	0.145	0.118	0.124

Table 4.21: Study II Mean Absolute Error for DIF Parameter Recovery

Impact	Covariates	N	DIF%	Small DIF (item 1-6)		Large DIF (item 7-12)	
				$\boldsymbol{\gamma}_1$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}_2$	$\boldsymbol{\beta}$
None	Two Independent Binary	500	33%	0.390	0.193	0.23	0.219
			66%	0.303	0.125	0.204	0.216
		2000	33%	0.191	0.189	0.113	0.129
			66%	0.158	0.175	0.12	0.111
	Correlated Binary+ Continuous	500	33%	0.224	0.24	0.155	0.163
			66%	0.325	0.34	0.166	0.157
		2000	33%	0.267	0.149	0.108	0.109
			66%	0.291	0.256	0.121	0.116
Impact	Two Independent Binary	500	33%	0.293	0.290	0.155	0.135
			66%	0.212	0.236	0.134	0.138
		2000	33%	0.192	0.139	0.108	0.105
			66%	0.130	0.136	0.099	0.093
	Correlated Binary+ Continuous	500	33%	0.218	0.129	0.151	0.161
			66%	0.276	0.278	0.148	0.133
		2000	33%	0.147	0.130	0.120	0.115
			66%	0.182	0.167	0.113	0.106

4.21 that the MAE of $\boldsymbol{\gamma}$ is larger than the MAE of $\boldsymbol{\beta}$ in the same condition, and this also cause bias in the estimation in $\boldsymbol{a}$ for the item with non-zero DIF parameter estimates. But overall the MAE of $\boldsymbol{a}$ is not affected a lot because the MAE of $\boldsymbol{a}$ is average out all 12 items, so the increased bias caused by DIF items are deluted.

From the results, it is clear that parameter recovery is generally accurate, and larger sample sizes, as expected, lead to better estimation. Table 4.21 shows that the bias for larger DIF is smaller than that for small DIF. This is because only a few DIFs were detected on items 1-6, and the MAEs of DIF parameters were only calculated for successfully detected items. The DIF parameter estimates for small DIF are easily overestimated and also have high variance.

Not only is DIF detection harder on the discrimination parameters, but parameter estimation for $\boldsymbol{\gamma}$ is also more difficult. Table 4.21 reveals that the MAE for $\boldsymbol{\gamma}$ is greater than the MAE for $\boldsymbol{\beta}$ under the same conditions. This discrepancy introduces bias in the estimation of $\boldsymbol{a}$ for items with non-zero DIF parameter estimates. However, the overall MAE for $\boldsymbol{a}$ remains relatively unaffected, as it represents the average across all 12 items, thereby diluting the increased bias caused by DIF items.

Table 4.22: Study II Mean Absolute Error for Approach 1 Impact Parameter Recovery

Covariates	N	DIF%	κ_1	κ_2	ω_1	ω_2	ω_3
Two Independent Binary	500	33%	0.196	0.212	0.099	0.115	0.091
		66%	0.206	0.192	0.101	0.114	0.117
	2000	33%	0.096	0.102	0.095	0.071	0.096
		66%	0.107	0.108	0.101	0.102	0.105

Table 4.23: Study II Mean Absolute Error for Approach 2 Impact Parameter Recovery

Covariates	N	DIF%	κ_1	κ_2	ϕ_1	ϕ_2	ϕ_3
Correlated Binary + Continuous	500	33%	0.206	0.192	0.152	0.139	0.163
		66%	0.223	0.178	0.143	0.181	0.149
	2000	33%	0.086	0.072	0.134	0.129	0.133
		66%	0.073	0.096	0.137	0.137	0.135

Introducing DIF on the slope also complicates the estimation of impact parameters. In

Tables 4.22 and 4.23, the MAE of impact parameter estimates is higher than in the uniform study. This is expected, as the relationship between $\boldsymbol{a\theta} + \boldsymbol{d}$ shows that the estimation of $\boldsymbol{a}$ and $\boldsymbol{\theta}$ influences one another. Thus, biases in estimating $\boldsymbol{a}$ and $\boldsymbol{\gamma}$ will lead to biases in estimating the distribution of the latent trait $\boldsymbol{\theta}$, especially in the mean impact, $\boldsymbol{\kappa}$. As a result, a significant increase in bias for the parameters κ_1 and κ_2 is evident.

As before, the estimation of impact on the covariance matrix remains not good. Despite neither method producing highly accurate impact parameter estimates, the DIF detection results were largely unaffected.

4.3 Simulation Study III

In the third simulation study, I examined the performance of the group lasso EMM algorithm by comparing it with the lasso EMM algorithm described in C. Wang et al. (2023). Specifically, I performed uniform DIF detection on simulated data generated in Simulation Study I in C. Wang et al. (2023) using the group lasso EMM algorithm, and the results were compared to those of the lasso EMM method presented in C. Wang et al. (2023).

The generated data in C. Wang et al. (2023) were as follows: The latent trait had a two-dimensional simple structure, and the test length was set at 20. Items 1-10 measured the first trait dimension, while items 11-20 measured the second. Two discrimination parameters, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, were drawn from Uniform(1.5, 2.5), and the boundary parameters, $\boldsymbol{d}$, were drawn from N(0,1). The true item parameters can be found in Table 4.24.

Table 4.24: Simulated True Item Parameters

Item	1	2	3	4	5	6	7	8	9	10
$\boldsymbol{a}_1$	2.17	0	2.41	2.45	2.34	1.84	1.85	1.92	1.94	1.90
$\boldsymbol{a}_2$	0	2.46	0	0	0	0	0	0	0	0
$\boldsymbol{d}$	0.03	-1.28	0.58	-2.06	0.12	3.25	-0.41	-0.51	0.89	1.33
Item	11	12	13	14	15	16	17	18	19	20
$\boldsymbol{a}_1$	1.92	0	0	0	0	0	0	0	0	0
$\boldsymbol{a}_2$	0	2.43	1.82	2.22	1.93	1.88	1.84	2.12	2.42	2.15
$\boldsymbol{d}$	0.85	0.82	-0.37	-0.99	-0.27	0.19	1.73	0.05	-1.86	-0.63

In C. Wang et al. (2023), we manipulated three factors. Two levels of total sample

size were considered: 1500 and 3000, and two levels of DIF proportion: 20% and 60%. Additionally, two levels of correlation between the factors were tested: low (0.25) and high (0.85) (Jiang, Wang, & Weiss, 2016). As no significant difference was noted between the two correlation levels, I only analyzed one level of trait correlation, fixing the correlation between the two latent traits at 0.85. A new sample size level was added, giving three levels: 750, 1500, and 3000. As before, the sample size was divided evenly among three groups: one reference group and two focal groups.

I simulated three covariate groups: one reference group and two focal groups. The first focal group showed a small DIF magnitude ($\beta_{j1} = 0.5$), while the second focal group had a larger DIF magnitude ($\beta_{j2} = 1$). For the 20% DIF condition, items 4, 5, 12, and 13 exhibited DIF. For the 60% condition, items 4-9 and 12-17 had DIF. The wABC values of the true DIF items are presented in Table 4.25 to illustrate the DIF magnitude. The latent trait $\boldsymbol{\theta}$ followed a multivariate normal distribution with zero means and unit variances. For each condition, I generated 50 independent datasets using the 2PL model.

Table 4.25: Uniform DIF magnitude measured by wABC

Item	4	5	6	7	8	9
Focal 1	0.06	0.07	0.03	0.08	0.08	0.07
Focal 2	0.12	0.13	0.05	0.16	0.15	0.13
Item	12	13	14	15	16	17
Focal 1	0.06	0.08	0.07	0.08	0.08	0.06
Focal 2	0.12	0.16	0.14	0.15	0.15	0.11

The evaluation criteria include Type I error and power at the omnibus level. In the lasso EMM method described in C. Wang et al. (2023), a separate DIF parameter is assigned to each item and each focal group. An item is considered to have omnibus DIF if at least one focal group displays DIF for that item, meaning that item j is considered to have uniform DIF at the omnibus level if any non-zero element exists in $\boldsymbol{\beta}_j$. In the group lasso EMM method, the DIF parameters for the two focal groups are grouped and hence penalized together for a single item, providing an omnibus-level outcome for all groups. The power and Type I error results are presented in the tables below.

Table 4.26: Study III Type I error of detecting uniform DIF

Correlation	N	DIF%	Lasso EMM	Group Lasso EMM
0.85	750	20%	0.028	0.007
		60%	0.045	0.017
	1500	20%	0.021	0
		60%	0.035	0.002
	3000	20%	0.026	0.002
		60%	0.021	0.002

Table 4.27: Study III Power of detecting uniform DIF

Correlation	N	DIF%	Lasso EMM	Group Lasso EMM
0.85	750	20%	0.75	0.57
		60%	0.712	0.197
	1500	20%	0.96	0.945
		60%	0.885	0.790
	3000	20%	1	1
		60%	0.943	0.998

From Tables 4.26 and 4.27, it can be observed that: 1. Power increases as the sample size becomes larger; 2. The low DIF proportion conditions have higher power compared to the high DIF proportion conditions with the same sample size; and 3. The group lasso EMM method controls Type I error very well, resulting in lower power compared to the lasso EMM method. Adjusting the information criteria to reduce the penalty for model complexity, thereby allowing more complex models, could increase both Type I error and power. However, the current Type I error for the group lasso EMM algorithm is very low, suggesting that power could be substantially improved with only a slight increase in Type I error, which would still remain below 0.05.

In C. Wang et al. (2023), group-level results were also reported alongside the omnibus-level results. Here, the group-level results for the lasso EMM method are extracted and presented in Table 4.28. The previous results indicate that the lasso EMM algorithm has significantly higher power for the large DIF group than for the small DIF group. This observation raises the question of whether the group lasso EMM can outperform the lasso EMM algorithm when all covariate groups have small magnitude DIF. To investigate, an

additional simulation study was conducted in which the DIF size for the second focal group was reduced to 0.5, matching the small DIF size of the first focal group. This adjustment meant that both focal groups had a DIF magnitude of 0.5. The two methods were then retested with variations in sample size (1500 vs. 3000) and DIF proportion (20% vs. 60%). The omnibus-level Type I error and power results are shown in Table 4.29, and the results are averaged across 50 replications.

Table 4.28: Group Level Type I Error and Power for Lasso EMM Algorithm

Corr	N	DIF%	Group	LASSO EMM Type I Error	LASSO EMM Power
0.85	1500	20%	Omnibus DIF	0.021 (0.005)	0.96 (0.017)
			Low DIF group	0.013 (0.004)	0.55 (0.043)
			High DIF group	0.011 (0.003)	0.96 (0.017)
		60%	Omnibus DIF	0.035 (0.011)	0.885 (0.019)
			Low DIF group	0.025 (0.009)	0.208 (0.024)
			High DIF group	0.013 (0.005)	0.885 (0.019)
	3000	20%	Omnibus DIF	0.026 (0.006)	1.000 (0)
			Low DIF group	0.021 (0.005)	0.84 (0.029)
			High DIF group	0.006 (0.003)	1.000 (0)
		60%	Omnibus DIF	0.060 (0.015)	0.998 (0.002)
			Low DIF group	0.058 (0.015)	0.632 (0.032)
			High DIF group	0.008 (0.004)	0.998 (0.002)

Table 4.29: Type I Error and Power for Small DIF Magnitude Detection

Result	N	DIF%	Lasso EMM	Group Lasso EMM
Type I error	1500	20%	0.020	0.001
		60%	0.015	0.018
	3000	20%	0.008	0.002
		60%	0.021	0.008
Power	1500	20%	0.335	0.320
		60%	0.080	0.802
	3000	20%	0.710	0.683
		60%	0.205	0.998

Analyzing the results, it is clear that the group lasso EMM algorithm tends to perform much better than the lasso EMM algorithm when all focal groups have a small DIF

magnitude. Both methods yield acceptable Type I error rates across all conditions. When the DIF proportion is low, the lasso EMM method sometimes outperforms the group lasso EMM method. This occurs because the group lasso EMM algorithm maintains a very low Type I error rate. One potential solution, as mentioned before, is to adopt different information criteria with a less stringent penalty on model complexity would substantially enhance power while keeping the Type I error rate similar to that of the lasso EMM method.

With a high DIF proportion, the group lasso EMM method significantly outperforms the lasso EMM method. Thus, based on this simulation study, it can be concluded that when biased items impact all focal groups, the group lasso EMM method is a strong alternative to the lasso EMM method, especially when the bias for all groups is relatively small and the proportion of biased items is high.

Chapter 5

EMPIRICAL DATA ANALYSIS

A real dataset from the Patient-Reported Outcome Measures (PROMIS) was used to show the effectiveness of the updated lasso EMM algorithm compared to the earlier version detailed in C. Wang et al. (2023). As before, age-related DIF was specifically targeted in the analysis. Previously, as described in C. Wang et al. (2023), age was divided into three distinct groups, and polytomous item responses were condensed into binary outcomes. In this study, however, I considered age as a continuous variable while retaining the original polytomous format for item responses. A comparative analysis was performed between scenarios where a continuous covariate is categorized, and polytomous item responses were collapsed into binary outcomes, and where a continuous covariate remains continuous and is estimated with the graded response model.

The sample includes responses from 5,219 cancer patients to the depression and anxiety PROMIS scales. The covariate age is a continuous integer ranging from 21 to 84. In the previous study, where age was a categorical covariate, the reference group was "Age 21-49," with a sample size of $n = 1,143$, and the two focal groups were "Age 50-64" ($n = 1,935$) and "Age 65-84" ($n = 2,141$). In the current analysis, where age is treated as a continuous variable, it was standardized. After standardization, age 60 is set as the reference level (standardized age 0). The data contain 129 individuals of this age.

The dataset includes 21 polytomous items, each with five response categories: 1 for "Never," 2 for "Rarely," 3 for "Sometimes," 4 for "Often," and 5 for "Always." For the analysis using the algorithm proposed in C. Wang et al. (2023), these response categories were combined into a dichotomous dataset. The "Never" response was coded as '0', while the other four categories were grouped together and coded as '1'. The data has a simple structure, with the first 10 items designed to measure depression and the remaining 11 to assess anxiety. Table 5.1 provides the details of each item's content.

Table 5.1: PROMIS depression and anxiety imputed data set: Item description

1	I felt worthless
2	I felt that I had nothing to look forward to
3	I felt helpless
4	I felt sad
5	I felt like a failure
6	I felt depressed
7	I felt unhappy
8	I felt hopeless
9	I felt discouraged about the future
10	I felt disappointed in myself
11	I felt fearful
12	I felt anxious
13	I felt worried
14	I found it hard to focus on anything other than my anxiety
15	I felt nervous
16	I felt uneasy
17	I felt tense
18	My worries overwhelmed me
19	I felt like I needed help for my anxiety
20	Many situations made me worry
21	I had difficulty calming down

In the analysis, I used two methods to detect DIF in the intercept. With the previous lasso EMM method in C. Wang et al. (2023), two DIF parameters, β_1 and β_2, were estimated, each representing a focal group. Specifically, β_1 indicated the DIF effect in focal group 1 (ages 50-64), and β_2 represented the DIF effect in focal group 2 (ages 65-84). For the results obtained using the proposed lasso EMM algorithm in this dissertation, four DIF parameters—β_1, β_2, β_3, and β_4—were estimated, each reflecting the DIF effect of the continuous covariate age on the four difficulty parameters d_1, d_2, d_3, and d_4.

As seen in Table 5.2, the results from both methods were consistent, with items 4, 6, 7, 11, 13, and 18 showing DIF across both methods. The new algorithm also identified DIF in four additional items: 12, 16, 17, and 19, compared to the old method. However, the magnitude of DIF for these items was relatively small, biasing only two of the five response categories.

To validate which approach produces results that are closer to the unknown truth, I

Table 5.2: Uniform DIF Detection by Two Algorithm

	Previous lasso EMM		New lasso EMM			
Item	β_1	β_2	β_1	β_2	β_3	β_4
1						
2						
3						
4		-0.55	-0.34	-0.408	-0.371	
5						
6		-0.41	-0.25	-0.282	-0.289	
7		-0.33	-0.194	-0.250	-0.334	
8						
9						
10						
11	-0.35	-0.67	-0.314	-0.314	-0.284	
12					-0.304	
13		-0.47	-0.365	-0.297	-0.368	-0.403
14						
15						
16					-0.343	
17					-0.425	
18		-0.34	-0.213		-0.261	
19					-0.293	
20						
21						

compared the observed response proportions with the model-predicted response proportions. Specifically, I calculated the group-specific probability of response $u = 1$ using the estimated item parameters and latent traits for both approaches, and then I compared each set of estimated response proportions to the observed response proportions to identify the approach with the smaller discrepancy. For the first approach, I calculated the group-specific probability of response $u = 1$ for item j using the following equation:

$$\hat{O}_{jp} = \sum_{m=1}^{M} \frac{1}{1 + e^{-(\boldsymbol{\alpha}_{jp}\boldsymbol{q}_m + \delta_{jp})}} f(\boldsymbol{q}_m | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \tag{5.1}$$

where $\boldsymbol{\alpha}jp$ represents the group-specific discrimination parameter. For the first group (ages 21-49, the reference group), $\boldsymbol{\alpha}j1 = \boldsymbol{a}j$. For the second group (ages 50-64, the first focal group), $\boldsymbol{\alpha}j2 = \boldsymbol{a}j + \boldsymbol{\gamma}j1$. For the third group (ages 65-84, the second focal group), $\boldsymbol{\alpha}j3 = \boldsymbol{a}j + \boldsymbol{\gamma}j2$. Similarly, δjp is the group-specific boundary parameter: $\delta_{j1} = d_j$, $\delta_{j2} = d_j + \beta_{j1}$, and $\delta_{j3} = d_j + \beta_{j2}$. The group-specific mean vector and covariance matrix are denoted as $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$, respectively. The combination of Gaussian rectangular quadrature nodes across M_0 nodes for each dimension of the latent variable $\boldsymbol{\theta}$ is represented as $\boldsymbol{q}_m$ $(m = 1, ..., M;\ M = (M_0)^K)$.

To ensure comparability between the two approaches, I categorized the results of the second approach into three age groups and the response categories were converted into dichotomous data. Formula 5.1 was then used to calculate the group-specific response probability. In the original data, the mean standardized age for the three age groups is -1.435, -0.201, and 0.947. To dichotomize the estimated probability, only the first boundary parameter $\boldsymbol{d}_1$ and its corresponding DIF parameter $\boldsymbol{\beta}_1$ were used, ignoring the remaining three boundary parameters $\boldsymbol{d}_2$, $\boldsymbol{d}_3$, and $\boldsymbol{d}_4$. For the first group (ages 21-49), $\boldsymbol{\alpha}_{j1} = \boldsymbol{a}_j + (-1.435) * \boldsymbol{\gamma}_{j1}$. For the second group (ages 50-64), $\boldsymbol{\alpha}_{j2} = \boldsymbol{a}_j + (-0.201) * \boldsymbol{\gamma}_{j2}$. For the third group (ages 65-84), $\boldsymbol{\alpha}_{j3} = \boldsymbol{a}_j + 0.947 * \boldsymbol{\gamma}_{j3}$. Similarly, $\delta_{j1} = d_{j1} + (-1.435) * \beta_{j1}$, $\delta_{j2} = d_{j1} + (-0.201) * \beta_{j1}$, and $\delta_{j3} = d_{j1} + 0.947 * \beta_{j1}$.

The results are presented in Tables 5.3 through 5.7. Table 5.3 displays the proportion of observed responses greater than or equal to 1 for the three age groups. Tables 5.4 and 5.5 show the proportion of predicted responses greater than or equal to 1 for the three age

groups using the previous and new approaches, respectively. Tables 5.6 and 5.7 detail the absolute differences between the observed and predicted response proportions for the two approaches, showing the differences of the predicted results from the truth.

Table 5.3: Observed (response≥1) proportion for different age groups.

Item	Age 21-49	Age 50-64	Age 65-84
1	0.416	0.594	0.337
2	0.400	0.390	0.334
3	0.467	0.413	0.347
4	0.696	0.623	0.524
5	0.412	0.368	0.290
6	0.585	0.535	0.432
7	0.647	0.601	0.505
8	0.437	0.380	0.301
9	0.563	0.503	0.432
10	0.497	0.459	0.374
11	0.595	0.506	0.391
12	0.659	0.577	0.483
13	0.770	0.690	0.584
14	0.466	0.429	0.342
15	0.593	0.548	0.462
16	0.586	0.553	0.468
17	0.634	0.579	0.493
18	0.529	0.459	0.357
19	0.421	0.393	0.297
20	0.628	0.571	0.496
21	0.440	0.427	0.349

From the results presented, it can be seen that the recovered response proportion matrix for both methods is satisfactory, and the new method exhibits a slightly smaller MAE, suggesting a better fit to the original data. As observed in Table 5.3, all items except the first show a reduction in response proportion with increasing age (ages 21-49 have the highest response proportion, ages 50-64 have the second highest, and ages 65-84 have the lowest). From Tables 5.6 and 5.7, it can be seen that both model fits can capture this response pattern by accurately estimating the impact parameters. However, patterns like those observed in the first item could not be correctly recovered by the model; therefore, item 1 shows the largest MAE for both methods, as seen in the last column of Tables 5.6 and 5.7. Fortunately,

Table 5.4: Previous Method predicted response proportion.

Item	Age 21-49	Age 50-64	Age 65-84
1	0.504	0.391	0.307
2	0.505	0.387	0.300
3	0.533	0.415	0.327
4	0.737	0.624	0.501
5	0.478	0.362	0.276
6	0.652	0.531	0.412
7	0.709	0.593	0.483
8	0.502	0.381	0.291
9	0.620	0.502	0.414
10	0.562	0.447	0.362
11	0.652	0.507	0.387
12	0.711	0.598	0.513
13	0.806	0.701	0.592
14	0.539	0.426	0.333
15	0.657	0.540	0.450
16	0.662	0.543	0.452
17	0.688	0.571	0.483
18	0.599	0.484	0.365
19	0.488	0.382	0.292
20	0.681	0.568	0.481
21	0.529	0.420	0.330

Table 5.5: New Method predicted response proportion.

Item	Age 21-49	Age 50-64	Age 65-84
1	0.474	0.388	0.313
2	0.497	0.407	0.328
3	0.497	0.409	0.331
4	0.742	0.629	0.511
5	0.445	0.358	0.283
6	0.627	0.512	0.404
7	0.699	0.596	0.493
8	0.462	0.37	0.291
9	0.591	0.504	0.421
10	0.535	0.449	0.371
11	0.631	0.503	0.384
12	0.659	0.575	0.494
13	0.803	0.699	0.583
14	0.503	0.413	0.333
15	0.627	0.539	0.454
16	0.632	0.542	0.457
17	0.659	0.574	0.49
18	0.56	0.447	0.345
19	0.454	0.368	0.293
20	0.652	0.569	0.488
21	0.496	0.409	0.332

there is only one item with this pattern, leading to the conclusion that the overall model fit for the PROMIS data is good. The new lasso EMM method, which uses continuous age as a covariate and retains polytomous response data, results in a better fit. However, it must be acknowledged that categorizing the continuous covariate and collapsing the polytomous responses can significantly reduce computational effort. Therefore, in practical applications, categorizing the continuous covariate and collapsing the polytomous responses may be a viable approach, offering a slightly more biased but much more efficient analysis.

Table 5.6: Absolute Difference for previous method.

Item	Age21-49	Age50-64	Age65-84	Mean
1	0.088	0.203	0.030	0.107
2	0.105	0.003	0.034	0.047
3	0.066	0.002	0.020	0.029
4	0.041	0.001	0.023	0.022
5	0.066	0.006	0.014	0.029
6	0.067	0.004	0.020	0.030
7	0.062	0.008	0.022	0.031
8	0.065	0.001	0.010	0.025
9	0.057	0.001	0.018	0.025
10	0.065	0.012	0.012	0.030
11	0.057	0.001	0.004	0.021
12	0.052	0.021	0.030	0.034
13	0.036	0.011	0.008	0.018
14	0.073	0.003	0.009	0.028
15	0.064	0.008	0.012	0.028
16	0.076	0.010	0.016	0.034
17	0.054	0.008	0.010	0.024
18	0.070	0.025	0.008	0.034
19	0.067	0.011	0.005	0.028
20	0.053	0.003	0.015	0.024
21	0.089	0.007	0.019	0.038
Mean	0.062	0.009	0.016	0.029

Table 5.7: Absolute Difference for new method.

Item	Age21-49	Age50-64	Age65-84	Mean
1	0.058	0.206	0.024	0.096
2	0.097	0.017	0.006	0.04
3	0.03	0.004	0.016	0.016
4	0.046	0.006	0.013	0.022
5	0.033	0.01	0.007	0.017
6	0.042	0.023	0.028	0.031
7	0.052	0.005	0.012	0.023
8	0.025	0.01	0.01	0.015
9	0.028	0.001	0.011	0.013
10	0.038	0.010	0.003	0.017
11	0.036	0.003	0.007	0.015
12	0.000	0.002	0.011	0.004
13	0.033	0.009	0.001	0.014
14	0.037	0.016	0.009	0.021
15	0.034	0.009	0.008	0.017
16	0.046	0.011	0.011	0.023
17	0.025	0.005	0.003	0.011
18	0.031	0.012	0.012	0.018
19	0.033	0.025	0.004	0.021
20	0.024	0.002	0.008	0.011
21	0.056	0.018	0.017	0.030
Mean	0.038	0.019	0.010	0.023

Chapter 6

DISCUSSION

The main objectives of this dissertation were to develop a computationally more efficient algorithm for evaluating DIF associated with multiple covariates, polytomous responses, multidimensional latent traits, and situations where no anchor items are available. Additionally, it aimed to introduce two methods that can accurately and effectively recover the impact of covariates on latent traits. Another goal was to evaluate and compare the new EMM algorithms with the existing regularized DIF package (i.e., regDIF) and our previous lasso EMM algorithm in C. Wang et al. (2023). Lastly, the dissertation validated the proposed methods using an empirical example, thus providing insight into their generalizability. Below, I summarize the progress made on these objectives, conclude the findings, and discuss future directions for research.

In our previous work in C. Wang et al. (2023), we developed three lasso-based regularization DIF detection algorithms and compared them with the IRT-LR-DIF. The lasso EMM algorithm was identified as having the best performance after rigorous simulation examinations. However, several limitations exist in the previous version of the method.

Firstly, the old version of the lasso EMM algorithm could only handle a limited number of categorical covariates. This limitation arose because the method estimated separate parameters for each focal group. For example, if there was one categorical covariate with 10 categories, it would result in 10 covariate groups, including 9 focal groups. Consequently, there would be nine sets of impact and DIF parameters that needed to be estimated in addition to the item parameters. This made it impossible for the algorithm to handle continuous covariates.

In the proposed algorithm in this dissertation, the inclusion of $\boldsymbol{X}$ in equation 2.2 enables us to model the continuous covariate effect. This approach allows for the detection of DIF associated with multiple covariates, where each covariate can be binary or continuous.

The second improvement made here is the introduction of two impact estimation methods for multidimensional IRT models. In the unidimensional case, only one mean impact parameter and one variance impact parameter need to be estimated for each covariate. In a multidimensional model with K dimensions, it is assumed that latent traits follow a conditional multivariate normal distribution. Therefore, there are $\frac{K(K+1)}{2}$ correlations between latent variables to be estimated. In addition to these correlations, to model how covariates influence the distribution of latent traits, I need to estimate a P-by-K matrix for mean impact and a P-by-$\frac{K(K+1)}{2}$ or $(P+1)$-by-K matrix for covariance impact for the two approaches (Approach I and Approach II), respectively, where P is the number of covariates. Modeling the impact on the mean is relatively straightforward. I calculate the multiplication of the mean impact parameter $\boldsymbol{\kappa}$ and the covariate variable $\boldsymbol{X}$. For the reference group, where $\boldsymbol{X}_i = 0$, the mean vector will be $\boldsymbol{\mu} = \boldsymbol{\kappa}^T \boldsymbol{X}^T = \mathbf{0}$.

To estimate the impact on the covariance matrix, I need to ensure the covariance matrix remains positive definite during the iterative estimation process. To address this, in Approach I, the impact estimation method in MNLFA is expanded. A P-by-$\frac{K(K+1)}{2}$ matrix ω is used to model the impact on the covariance matrix. Our approach then performs a Cholesky decomposition of the covariance matrix and incorporates covariate effects into the lower triangular part of the decomposed matrix. For the reference group, covariate values are set to zero, which fixes the diagonal of the baseline covariance matrix to all ones.

In Approach II, a similar mean structure as in Approach I is used, but the parameterization of the covariance structure from Hoff and Niu (2012) is introduced to capture the effect of covariates on the covariance matrix. In this method, a $(P+1)$-by-K matrix ϕ is used for covariance impact. This allows us to estimate a much smaller number of covariance impact parameters compared to P-by-$\frac{K(K+1)}{2}$ when K is large. Additionally, instead of continuing to use $\boldsymbol{X}$ as in Approach I, an intercept is added to the covariates to form $\boldsymbol{X}^* = (1, \boldsymbol{X})$ for both mean and covariance impact effects. This ensures flexibility and avoids assumptions about the variance being smallest for the reference group. The most important advantage of this method is that it is more computationally efficient because there is a closed-form solution for the three matrices to be estimated ($\boldsymbol{\kappa}$, $\boldsymbol{\phi}$, and $\boldsymbol{\Sigma}_0$). Combining these advantages, this approach is more suitable for cases with many latent traits and covariates compared to

Approach I.

Other generalization work I have done includes generalizing the 2PL model from C. Wang et al. (2023) to the GRM, enabling the method to handle different forms of items and tests. Additionally, I derived and examined the group-lasso penalty in addition to the lasso penalty. I believe that this group-level DIF detection can serve as an initial step for identifying DIF. When an item is found to have DIF, the lasso method can then be used to determine if the DIF is on a specific item parameter (whether it is uniform or non-uniform DIF) or to identify which specific focal group, trait dimension, or response category has DIF.

The simulation studies demonstrate the accuracy of our proposed methods under various conditions and their ability to handle different types of covariates and DIF. In the first two studies, I evaluated the performance of our method with both uniform and non-uniform DIF. In the non-uniform DIF study, we simulated small DIF on the intercept, thus increasing the difficulty of DIF detection. The results showed that our method still performs well under these challenging conditions. In each study, I simulated different types of covariates and their impact; these effects on the latent trait distribution interact with the DIF effect on discrimination parameters, posing additional challenges in model estimation. Despite these challenges, the proposed method demonstrated good power in DIF detection with well-controlled Type I errors under both uniform and non-uniform DIF conditions.

The most notable finding from the first two simulation studies is the robustness of the EMM algorithm when the DIF proportion is high, as we previously observed in C. Wang et al. (2023). Typically, using regularization methods to introduce sparsity in the parameter sets also causes bias in parameter estimation, necessitating a re-estimate step without regularization to obtain unbiased estimates for all parameters. In previous implementations, I found that biased parameter estimates from each EM cycle could adversely affect the subsequent iteration of the EM algorithms. Consequently, the cumulative bias from multiple EM cycles could cause the final estimation to converge at an incorrect point. To address this issue, I implemented a re-estimate step immediately after each EM cycle. Thus, each iteration consists of one E-step followed by two M-steps: the first M-step includes the regularization penalty, while the second M-step incorporates the sparsity introduced in the first M-step and performs a re-estimation without the penalty. This process allows us to input

unbiased estimates for the next E-step, thereby avoiding cumulative bias and ensuring the accuracy of our method.

The MSE for item, DIF, and impact parameters were reported for both uniform and non-uniform studies. In both studies, the MAE for item parameters was good, while the MAE for DIF parameters was slihgtly larger in the non-uniform study. Nonetheless, the impact parameters recovery show that the two impact estimation approaches have comparable accuracy.

The third simulation study was relatively simple. I did not use the proposed general framework to model the covariate effect on item and latent trait parameters. Instead, I used the DIF model from C. Wang et al. (2023), which estimates a set of impact and item parameters for each focal group. Except for η, all other tuning parameters for the group-lasso algorithm in (3.52) were fixed. The results showed that the group-lasso has good power and very low Type I error for DIF detection when DIF affects all covariate groups.

The real data analysis of the PROMIS dataset served as an excellent example to demonstrate the effects of performing DIF detection by collapsing ordinal response data into binary data and transforming continuous covariates into several discrete groups. The results suggest that it was worthwhile to generalize the method to handle multiple and continuous covariates as well as the Graded Response Model (GRM). By comparing the model-predicted response proportions with the observed response proportions, I found that the results generated by the new algorithm were slightly closer to the observed data. However, it must be admitted that the computing time required for the new algorithm is significantly greater than that for the old algorithm. Therefore, in practice, collapsing ordinal response data into binary data and transforming continuous covariates into discrete groups might still be a practical approach.

There are several directions for future research to advance the development and evaluation of the proposed method, which may also apply to other regularization methods in general. First, different methods for calculating starting values can be explored. Starting values are crucial for the convergence efficiency of complex EM algorithms with many parameters. The current algorithm uses the `mirt` package to estimate an initial set of starting values, and then uses these values, along with our function and a zero or very small

penalty, to estimate another set of starting values. This second set of starting values serves as the initial points in the early computation for different tuning parameters in the later estimation.

The procedure for finding starting values was even more complicated for non-uniform DIF since our algorithm would not converge with $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ with all large non-zero values. In this case, I used our algorithm to calculate starting values for $\boldsymbol{\gamma}$, and set the starting values for $\boldsymbol{\beta}$ to 0.5 for all items. This method can be further improved by first estimating starting values for $\boldsymbol{\gamma}$ while constraining $\boldsymbol{\beta}$, and then estimating starting values for $\boldsymbol{\beta}$ while constraining $\boldsymbol{\gamma}$. Additionally, our algorithm could be used to calculate starting values for the impact parameters by constraining the DIF parameters. Overall, using the parameter estimates obtained from the lasso EMM algorithm with a zero or small tuning parameter as starting values is expected to speed up the convergence of our algorithm for later tuning parameter values.

Second, there are different ways to select the regularization parameter other than BIC, such as GIC proposed by Zhang, Li, and Tsai (2010). The GIC is defined as

$$GIC_{\hat{\boldsymbol{\Delta}}_\lambda} = -2 \max_{\hat{\boldsymbol{\Delta}}_\lambda} \log M + k_n(\|\hat{\boldsymbol{\beta}}\|_0 + \|\hat{\boldsymbol{\gamma}}\|_0). \tag{6.1}$$

Acording to the theorem 1 in Zhang et al. (2010), $k_n = c \log\log(N) \log(N)$ for GIC. And when $k_n = \log(N)$, GIC is equal to BIC. We can change the weight of the penalty term by adjusting the value of c for different conditions. In conditions where we have low power and very small Type I error, we might adopt a value of $c < 1$ to reduce the weight of the penalty term, thus increasing the power (with a slight increase in Type I error) of the final selected results. Most of the time, we just need to control the Type I error to be less than 0.05, so we can achieve as large power as possible while keeping the Type I error below 0.05.

Third, we use the Gaussian quadrature method for estimating integrals in the marginal likelihood function, and we only explored a two-dimensional model in this dissertation. Although the proposed algorithms are generalized enough to handle latent traits with more than two dimensions, the quadrature-based method may not be efficient for three-dimensional models and may struggle with four or higher dimensions. An alternative method

for integral approximation could be the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010).

Fourth, we use the Expectation-Maximization (EM) algorithm combined with a cyclical coordinate descent algorithm for parameter estimation. This algorithm is not computationally efficient for two reasons. First, in the E-step, when calculating the complete-data likelihood function, we need to compute the exact posterior distributions of latent variables for each individual. This involves a nested for loop over the number of quadrature points and the sample size, resulting in a loop over $N * G$, which is computationally intensive. Second, the cyclical coordinate descent is necessary because there is no closed-form solution for the MLE of item and DIF parameters in Equation (3.5). This requires iterative optimization for item and DIF parameter estimations in each M-step. Since the EM algorithm is already iterative, adding cyclical coordinate descent to each M-step makes the entire algorithm highly inefficient.

A possible solution to speed up the current algorithm is the use of variational methods. Cho, Wang, Zhang, and Xu (2021) developed a Gaussian Variational Expectation-Maximization (GVEM) algorithm for MIRT models. The GVEM algorithm approximates the intractable integrals involved in the marginal likelihood with a computationally feasible form known as the variational lower bound. This significantly reduces the complexity of calculations compared to traditional EM algorithms. Additionally, model parameters can be updated using closed-form solutions during the M-step of the GVEM algorithm. This means that instead of iterative numerical optimization for each parameter, the GVEM employs direct analytical updates, which are more efficient.

The GVEM algorithm can be generalized for DIF analysis. For example, in the 2PL model, DIF parameters for an item are introduced as

$$P_j(\boldsymbol{\theta}_i) = \frac{1}{1 + e^{-[(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\boldsymbol{\theta}_i - (b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)]}} \quad (i = 1, ..., N; j = 1, 2, ..., J), \tag{6.2}$$

where X_i is the covariate group indicator, and X can take G different values indicating G different covariate groups.

The marginal likelihood approximated by the variational lower bound for item j is

$$
\begin{aligned}
L_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j) = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\Big(&\log\frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + (\frac{1}{2} - Y_{ij})(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j) + (Y_{ij} - \frac{1}{2})(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i^{(t)} \\
&- \frac{1}{2}\xi_{i,j} - \eta(\xi_{i,j})\{(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)^2 - 2(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i^{(t)} \\
&+(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T](\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j) - \xi_{i,j}^2\}\Big) \\
&+ \sum_{g=1}^{G}\frac{N_g}{2}\log|\Sigma_{\theta g}^{-1}| - \sum_{g=1}^{G}\sum_{i=1}^{N_g}\frac{1}{2}Tr(\Sigma_{\theta g}^{-1}[\Sigma_i^{(t)} + (\mu_i^{(t)} - \mu_{\theta g})(\mu_i^{(t)} - \mu_{\theta g})^T]),
\end{aligned}
\tag{6.3}
$$

where $\xi_{i,j}$ is the variational parameter, and $\eta(\xi_{i,j}) = (2\xi_{i,j})^{-1}[e^{\xi_{i,j}}/(1+e^{\xi}_{i,j}) - 1/2]$. $\mu_i^{(t)}$ and $\Sigma_i^{(t)}$ are the mean vector and covariance matrix of the variational density $\hat{q}_i(\hat{\boldsymbol{\theta}}_i) \sim N(\hat{\mu}_i, \hat{\Sigma}_i)$ at the tth EM iteration. $\boldsymbol{\theta} \sim N(\mu_{\theta g}, \Sigma_{\theta g})$ is the distribution of the latent trait $\boldsymbol{\theta}$ in group g, and N_g is the sample size of group g.

Then penalties on DIF parameters can be added to the approximated marginal likelihood which is our objective function for DIF detection. Specifically, we have

$$
Q(\boldsymbol{a}, b, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{j=1}^{J}(L_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j) - \lambda|\boldsymbol{\gamma}_j| - \lambda|\boldsymbol{\beta}_j|), \tag{6.4}
$$

where λ is the regularization tuning parameter, $|\boldsymbol{\gamma}_j|$ is the L_1 penalty on DIF effect on slope, and $|\boldsymbol{\beta}_j|$ is the L_1 penalty on DIF effect on intercept. The L_1 estimators are denoted by $\hat{\boldsymbol{\Delta}}_\lambda = (\hat{\boldsymbol{a}}_\lambda, \hat{b}_\lambda, \hat{\boldsymbol{\gamma}}_\lambda, \hat{\boldsymbol{\beta}}_\lambda) = \arg\max_{\hat{\boldsymbol{\Delta}}_\lambda} Q(\boldsymbol{a}, b, \boldsymbol{\gamma}, \boldsymbol{\beta})$. A details for the update for each parameter are given in appendix.

Beyond computational advances, future research on model violations will also be crucial. In this dissertation, several assumptions underpin the proposed method. Firstly, it is assumed that the latent traits of the test-takers are multivariate normally distributed. If the trait distribution significantly deviates from normality, the estimates of item parameters and trait levels might be biased, necessitating adjustments to the likelihood function. Additionally, other assumptions such as local independence and monotonicity must be verified to prevent biased estimates and incorrect conclusions. Furthermore, although the data used in this study are complete, missing data represent another important area for future research

focus.

In summary, this dissertation introduces an innovative, generalized, and accurate framework for DIF detection. The framework is innovative because our algorithm leverages advanced regularization techniques and bridges a gap in the literature by contributing two methods to estimate covariate impact on multidimensional latent traits. It is generalized, as it can handle multiple types of covariates, items with various response categories, and multidimensional traits without needing anchor items. It is accurate because our algorithm effectively detects DIF in various simulation conditions and accurately recovers the parameters.

The advancements presented in this dissertation significantly improve DIF detection methodologies, providing researchers with a more powerful and versatile tool for assessing differential item functioning in diverse contexts. These contributions enhance the precision and applicability of DIF detection and pave the way for future innovations in the field.

Appendix A

DERIVATIVES FOR THE GVEM PARAMETER ESTIMATES

Variational parameters μ_i, Σ_i, and $\xi_{i,j}$ and latent trait distribution parameters μ_θ and Σ_θ are updated by

$$\Sigma_i^{-1} = \Sigma_{\theta g}^{-1} + 2\sum_{j=1}^{J} \eta(\xi_{i,j})(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T \tag{A.1}$$

$$\mu_i = \Sigma_i \times \{\sum_{j=1}^{J}[2\eta(\xi_{i,j})(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j) + Y_{ij} - \frac{1}{2}](\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T + \Sigma_{\theta g}^{-1}\mu_{\theta g}\} \tag{A.2}$$

$$\begin{aligned}\xi_{i,j}^2 = {} & (b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)^2 - 2(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i \\ & + (\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T[\Sigma_i + \mu_i\mu_i^T](\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)\end{aligned} \tag{A.3}$$

We fix $\mu_{\theta 1} = \mathbf{0}$, and for g=2,...,G we have

$$\mu_{\theta g} = \frac{1}{N_g}\sum_{i=1}^{N_g}\mu_i. \tag{A.4}$$

$\Sigma_{\theta g}$ is updated by

$$\Sigma_{\theta g} = \frac{1}{N_g}\sum_{i=1}^{N_g}[\Sigma_i + (\mu_i - \mu_\theta)(\mu_i - \mu_\theta)^T] \tag{A.5}$$

To make the reference group has varaince 1, we rescale $\Sigma_{\theta g}$ by

$$\Sigma_{\theta g}^* = ((\sqrt{diag(\Sigma_{\theta 1})})^{-1})^T\Sigma_{\theta g}(\sqrt{diag(\Sigma_{\theta 1})})^{-1} \tag{A.6}$$

The discrimination parameter $\boldsymbol{a}_j$ is updated by

$$\frac{\partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}{\partial \boldsymbol{a}_j} = \sum_{i=1}^{N} \Big((Y_{ij} - \frac{1}{2})\mu_i^{(t)} - \eta(\xi_{i,j})\{-2(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)\mu_i^{(t)} + 2(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T]\}\Big) = 0 \tag{A.7}$$

$$\hat{\boldsymbol{a}}_j = \frac{\sum_{i=1}^{N}(Y_{ij} - \frac{1}{2})\mu_i^{(t)} + 2\eta(\xi_{i,j})(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)\mu_i^{(t)} - 2\eta(\xi_{i,j})(\boldsymbol{X}_i\boldsymbol{\gamma}_j)^T[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T]}{\sum_{i=1}^{N} 2\eta(\xi_{i,j})[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T]} \tag{A.8}$$

Since the discrimination parameter interacts with the latent trait and we scale the latent trait in each EM cycle before updating the item parameters, we need to perform a rescale when updating the item discrimination parameter as follows:

$$\hat{\boldsymbol{a}}_j = \hat{\boldsymbol{a}}_j \sqrt{diag(\Sigma_\theta)} \tag{A.9}$$

The difficulty parametere b_j is updated by

$$\frac{\partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}{\partial b_j} = \sum_{i=1}^{N} \Big((\frac{1}{2} - Y_{ij}) - \eta(\xi_{i,j})\{(2b_j - 2\boldsymbol{X}_i\boldsymbol{\beta}_j) - 2(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i^{(t)}\}\Big) = 0. \tag{A.10}$$

$$\hat{b}_j = \frac{\sum_{i=1}^{N}\Big((\frac{1}{2} - Y_{ij}) + 2\eta(\xi_{i,j})\boldsymbol{X}_i\boldsymbol{\beta}_j + 2\eta(\xi_{i,j})(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i^{(t)}\Big)}{\sum_{i=1}^{N} 2\eta(\xi_{i,j})}. \tag{A.11}$$

Denote $\boldsymbol{\gamma}_{jg} = \boldsymbol{X}_i\boldsymbol{\gamma}_j$ as the DIF on slope for the covariate group g (g=1,...,G), and let N_g be the sample size of group g. $\boldsymbol{\gamma}_{jg}$ is updated by

$$\frac{\partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}{\partial \boldsymbol{\gamma}_{jg}} = \sum_{i=1}^{N_g} \Big((Y_{ij} - \frac{1}{2})\mu_i^{(t)} - \eta(\xi_{i,j})\{-2(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)\mu_i^{(t)} + 2(\boldsymbol{a}_j + \boldsymbol{\gamma}_{jg})^T[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T]\}\Big) \tag{A.12}$$

$$\frac{\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}{\partial \boldsymbol{\gamma}_{jg}^2} = \sum_{i=1}^{N_g} \Big(-2\eta(\xi_{i,j})[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T]\Big) \tag{A.13}$$

$$\hat{\gamma}_{jg} = -\frac{S\left(-\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j) * \boldsymbol{\gamma}^*_{jg} + \partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j), \lambda\right)}{\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}$$
$$= \frac{S\left(\sum_{i=1}^{N_g}(Y_{ij} - \frac{1}{2})\mu_i^{(t)} + 2\eta(\xi_{i,j})(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)\mu_i^{(t)} - 2\eta(\xi_{i,j})(\boldsymbol{a}_j)^T[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T]\right)}{\sum_{i=1}^{N_g} 2\eta(\xi_{i,j})[\Sigma_i^{(t)} + (\mu_i^{(t)})(\mu_i^{(t)})^T]} \tag{A.14}$$

Similar to the parameter $\boldsymbol{a}$, since $\hat{\boldsymbol{\gamma}}_{jg}$ also interacts with the latent trait, we need to rescale it as follows:

$$\hat{\boldsymbol{\gamma}}_{jg} = \hat{\boldsymbol{\gamma}}_{jg}\sqrt{diag(\Sigma_{\theta 1})} \tag{A.15}$$

DIF on the difficulty parameter is denoted as $\beta_{jg} = \boldsymbol{X}_i\boldsymbol{\beta}_j$ for covariate group g and is updated by

$$\frac{\partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}{\partial \beta_{jg}} = \sum_{i=1}^{N_g}\left(-(\frac{1}{2} - Y_{ij}) - \eta(\xi_{i,j})\{(-2b_j + 2\beta_{jg}) + 2(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i^{(t)}\}\right) \tag{A.16}$$

$$\frac{\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}{\partial \beta_{jg}^2} = \sum_{i=1}^{N_g} -2\eta(\xi_{i,j}) \tag{A.17}$$

$$\hat{\beta}_{jg} = -\frac{S\left(-\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j) * \beta^*_{jg} + \partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j), \lambda\right)}{\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}$$
$$= \frac{S\left(\sum_{i=1}^{N_g} -(\frac{1}{2} - Y_{ij}) - \eta(\xi_{i,j})\{-2b_j + 2(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i^{(t)}\}, \lambda\right)}{\sum_{i=1}^{N_g} 2\eta(\xi_{i,j})} \tag{A.18}$$

REFERENCES

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55.

Bauer, D. J., & Hussong, A. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125.

Bechger, T. M., Maris, G., Verstralen, H. H., & Béguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied psychological measurement*, *27*(5), 319–334.

Belzak, W. C., & Bauer, D. J. (2024). Using regularization to identify measurement bias across multiple background characteristics: A penalized expectation–maximization algorithm. *Journal of Educational and Behavioral Statistics*, 10769986231226439.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a metropolis–hastings robbins–monro algorithm. *Psychometrika*, *75*, 33–57.

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

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2022). Advantages of spike and slab priors for detecting differential item functioning relative to other bayesian regularizing priors and frequentist lasso. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(1), 122–139.

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

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). Lordif: An r package for detect-

ing differential item functioning using iterative hybrid ordinal logistic regression/item response theory and monte carlo simulations. *Journal of statistical software*, *39*(8), 1.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied psychological measurement*, *17*(4), 335–350.

Curran, P. J., Cole, V., Bauer, F. J., Hussong, A., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 827–844.

Edelen, M. O., Stucky, B., & Chandra, A. (2015). Quantifying 'problematic' dif within an irt framework: application to a cancer stigma index. *Qual Life Res*, *24*(1), 95–103.

Fidalgo, Á. M. (2011). A new approach for differential item functioning detection using mantel-haenszel methods. the gmhdif program. *The Spanish journal of psychology*, *14*(2), 1018–1022.

Finch, H. (2005). The mimic model as a method for detecting dif: Comparison with mantel-haenszel, sibtest, and the irt likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295.

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

Gierl, M. J., Gotzmann, A., & Boughton, K. A. (2004). Performance of sibtest when the percentage of dif items is large. *Applied Measurement in Education*, *17*(3), 241–264.

Hoff, P. D., & Niu, X. (2012). A covariance regression model. *Statistica Sinica*, 729–753.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the mantel-haenszel procedure. *ETS Research Report Series*, *1986*(2), i–24.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, *7*, 179786.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American statistical Association*, *70*(351a), 631–639.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning

using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, *58*(303), 690–700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719–748.

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(1), 53–71.

Oshima, T., Raju, N. S., & Flowers, C. P. (1997). Development and demonstration of multidimensional irt-based internal measures of differential functioning of ltems and tests. *Journal of Educational Measurement*, *34*(3), 253–272.

Roju, N. S., Van der Linden, W. J., & Fleer, P. F. (1995). Irt-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368.

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*(52), 279–294.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/dif from group ability differences and detects test bias/dtf as well as item bias/dif. *Psychometrika*, *58*(2), 159–194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361–370.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in rasch models. *Psychometrika*, *80*, 21–43.

Wallin, G., Chen, Y., & Moustaki, I. (2024). Dif analysis with unknown groups and anchor items. *Psychometrika*, 1–29.

Wang, C., Zhu, R., & Xu, G. (2023). Using lasso and adaptive lasso to identify dif in multidimensional 2pl models. *Multivariate behavioral research*, *58*(2), 387–407.

Wang, W., Liu, Y., & Liu, H. (2022). Testing differential item functioning without predefined anchor items using robust regression. *Journal of Educational and Behavioral Statistics*, *47*(6), 666–692.

Woods, C. M. (2009). Evaluation of mimic-model methods for dif testing with comparison to two-group analysis. *Multivariate behavioral research*, *44*(1), 1–27.

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential item functioning analysis without a priori information on anchor items: Qq plots and graphical test. *psychometrika*, *86*, 345–377.

Zhang, Y., Li, R., & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*, 312-323.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (dif). *Ottawa: National Defense Headquarters*, *160*.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233–251.