Multi-Group Regularized Gaussian Variational Estimation: Fast Detection of DIF

Weicong Lyu
Faculty of Education, University of Macau, Macau, China
Chun Wang
College of Education, University of Washington, Seattle, WA, USA
Gongjun Xu
Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Correspondence concerning this manuscript should be addressed to Chun Wang at:
312E Miller Hall
2012 Skagit Ln, Seattle, WA 98105
Email: wang4066@uw.edu
Correspondence concerning this manuscript should be addressed to Gongjun Xu at:
456 West Hall
1085 South University, Ann Arbor, MI 48109
Email: gongjun@umich.edu

Competing interest and Data Availability statement:
The authors declare no competing interests. The code that supports the findings of this study will be available on the project webpage (https://sites.uw.edu/pmetrics/projects/) shortly as we are still working on creating user-friendly R package and Shiny App. The real data will be available upon request.

Author contributions: Weicong Lyu performed the formal analysis and initial draft writing; Chun Wang contributed the original idea, partial writing, reviewing and editing, and obtaining funding; Gongjun Xu contributed the original idea, partial writing, reviewing and editing, and obtaining funding.

Citation
Lyu, W., Wang, C. & Xu, G. (2024, accepted in October). Multi-Group Regularized Gaussian Variational Estimation: Fast Detection of DIF. *Psychometrika*.

Abstract

Data harmonization is an emerging approach to strategically combining data from multiple independent studies, enabling addressing new research questions that are not answerable by a single contributing study. A fundamental psychometric challenge for data harmonization is to create commensurate measures for the constructs of interest across studies. In this study, we focus on a regularized explanatory multidimensional item response theory model (re-MIRT) for establishing measurement equivalence across instruments and studies, where regularization enables the detection of items that violate measurement invariance, also known as differential item functioning (DIF). Because the MIRT model is computationally demanding, we leverage the recently developed Gaussian Variational Expectation-Maximization (GVEM) algorithm to speed up the computation. In particular, the GVEM algorithm is extended to a more complicated and improved multi-group version with categorical covariates and Lasso penalty for re-MIRT, namely, the importance weighted GVEM with one additional maximization step (IW-GVEMM). This study aims to provide empirical evidence to support feasible uses of IW-GVEMM for re-MIRT DIF detection, providing a useful tool for integrative data analysis. Our results show that IW-GVEMM accurately estimates the model, detects DIF items, and finds a more reasonable number of DIF items in a real world dataset. The proposed method has been integrated into R package VEMIRT (`https://map-lab-uw.github.io/VEMIRT`).

Key words: latent variable modeling, variational estimation, regularization, differential item functioning

1 Introduction

Addressing broad scope research questions, such as the impact of medical, behavioral, and psycho-social interventions, is typically beyond the scope of a single research project and requires data from multiple studies to build a more cumulative science. Integrative data analysis (IDA) is a novel framework for conducting simultaneous analysis of raw data pooled from different studies. It offers many advantages, including increased power due to larger sample sizes, enhanced external validity and generalizability due to greater heterogeneity in demographic and psycho-social characteristics, cost-effectiveness due to the reuse of extant data, and potential to address new research questions not feasible by a single study, among others (Curran & Hussong, 2009; Curran, Obeidat, & Losardo, 2010). However, significant methodological challenges must be addressed when pooling data from independent studies, and one such challenge is to establish commensurate measures for the constructs of interest (e.g., Nance et al., 2017). When data from different yet overlapping instruments and diverse samples are pooled, the assumption of measurement invariance, often required by existing methods, would likely be violated.

Procedures for evaluating and establishing measurement equivalence across samples are well developed from both factor analysis and item response theory frameworks. These traditional methods focus on comparing independent groups defined by a single categorical covariate to determine if any items display differential item functioning (DIF, also known as item-level measurement non-invariance). More recently, Bauer (2017) proposed the moderated nonlinear factor analysis (MNLFA), a unified flexible model that can handle different types of study-specific covariates simultaneously, such as gender (categorical) and age (continuous), and can handle different types of responses. The cost of this generalization is the drastically increased model complexity that prohibits the adoption of conventional DIF detection methods simply because the resulting number of potential model comparisons would be huge. To overcome this problem, Bauer, Belzak, and Cole (2020) proposed a regularized MNLFA by using a penalized likelihood function that imposes a Lasso (i.e., least absolute

shrinkage and selection operator) penalty on DIF parameters. This procedure obviates the reliance on statistical hypothesis testing for DIF, but instead, the penalty term shrinks small DIF parameters directly toward zero, indicating that DIF is not detected on these item parameters. Although the Lasso penalty has been proven to have good performance under some conditions (van de Geer, 2008; Zhao & Yu, 2006), the theoretical guarantee of Lasso (such as oracle property) in item response theory models has yet to be established.

The current regularized MNLFA is only restricted to unidimensional constructs, while this work aims to expand the methodology to accommodate multidimensional constructs. This is an important step forward as many theoretical constructs in behavioral and health measurement in general are related, complex, and multifaceted (Fayers, 2007; Michel et al., 2018; Zheng, Chang, & Chang, 2013). For instance, HIV stigma, a barrier to HIV testing and counseling, status disclosure, partner notification, and antiretroviral theory (ART) access and adherence, is found to have at least two dimensions: emotional stigma and physical stigma (Carrasco, Arias, & Figueroa, 2017). In addition, clinical patient-reported outcome measures (PROMs) have been increasingly endorsed, or even mandated by policymakers and payers as a means of gauging not only a treatment's benefits, but also its appropriateness. Since multi-trait assessment has emerged as a fundamental requirement for patient-centered decision making, the methodology also needs to advance on par with the demand. From a statistical perspective, using a multivariate approach would also produce more accurate factor scores with reduced standard errors of measurement by borrowing information from correlated scales.

In this study, we focus on a regularized explanatory multidimensional IRT (re-MIRT) model that handles potential item measurement non-invariance (i.e., DIF), thereby adjusting for, for instance, between-study heterogeneity. With proper penalty such as Lasso, fitting re-MIRT on the integrated data will output a commensurate scale for multidimensional constructs (e.g., depression, anxiety, alcohol use) that well accounts for study-specific idiosyncrasy resulting from the diversity of study populations and the use of different instru-

3

ments. In addition, for the common items shared among studies, re-MIRT automatically tests for measurement invariance and corrects for non-invariance when spotted. Hence, the final factor scores from re-MIRT are cleaned from the contamination of DIF and they can be readily used in subsequent statistical analyses for addressing critical research questions.

In recent literature, C. Wang, Zhu, and Xu (2023) first used the Lasso-type penalty with the two-dimensional two-parameter logistic model, and their proposed methods outperform the likelihood ratio test approach, especially when the proportion of DIF items is high. However, the regularization method can be slow because it requires a full estimation for each candidate tuning parameter value. When a large grid of tuning parameters is considered, the entire algorithm may take hours to finish. In this study, we aim to overcome these difficulties by leveraging the recently developed Gaussian Variational Expectation-Maximization (GVEM) algorithm (Cho, Wang, Zhang, & Xu, 2021) for MIRT models, which relies on a variational lower bound to approximate the true marginal likelihood, to speed up the computation. We generalize the GVEM algorithm to the more complicated DIF analysis setting with categorical covariates. To obtain a tighter lower bound for more accurate DIF detection, we further incorporate the importance sampling approach as an additional step after GVEM estimation to reduce the estimation bias (Ma, Ouyang, Wang, & Xu, 2023). Compared to existing DIF detection methods, our proposed method is more efficient and scalable to higher dimensions and large-scale data, while still performing well in DIF detection. In addition, the source code for the proposed method is made available in R package VEMIRT, which can be accessed at `https://map-lab-uw.github.io/VEMIRT`.

The rest of the paper is organized as follows. We first introduce the re-MIRT model for binary responses, followed by the regularized GVEM algorithm and bias reduction methods. Then we present two simulation studies and a real data analysis to evaluate the performance of the proposed algorithm. This paper ends with some final discussions.

4

2 Method

2.1 Regularized Explanatory MIRT

Let N, J, K, and G denote the numbers of persons, items, latent dimensions, and groups, respectively. For a dichotomously scored item j, the probability that person i with a latent trait vector $\boldsymbol{\theta}_i$ gives a correct response to item j is modeled as

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i) = \frac{1}{1 + \mathrm{e}^{-\left[(\boldsymbol{a}_j + \boldsymbol{\gamma}_j^\mathsf{T} \boldsymbol{X}_i)^\mathsf{T} \boldsymbol{\theta}_i - (b_j - \boldsymbol{\beta}_j^\mathsf{T} \boldsymbol{X}_i)\right]}}, \tag{1}$$

where $\boldsymbol{a}_j \in \mathbb{R}^K$ is a vector of discrimination parameters of item j, b_j is a difficulty parameter of item j, and $\boldsymbol{\theta}_i \in \mathbb{R}^K$ is a vector of latent traits for person i. The explanatory feature of the model is reflected by the inclusion of person level covariates, $\boldsymbol{X}_i \in \mathbb{R}^P$, which includes all the grouping information related to DIF (Wilson, De Boeck, & Carstensen, 2008). In this study, we focus on a simpler case where person level covariates uniquely determine the group membership, i.e., $\boldsymbol{X}_i \equiv \bar{\boldsymbol{X}}_g$ for all $i \in I_g$ where I_g is the set of all persons in group g. $\boldsymbol{\beta}_j \in \mathbb{R}^P$ is a vector of regression coefficients implying the effect of grouping variables on the probability of correct response on item j. Similarly, $\boldsymbol{\gamma}_j \in \mathbb{R}^{P \times K}$ is a matrix of regression coefficients denoting the interaction effects of $\boldsymbol{\theta}_i$ and grouping variables on item responses. By this way of parameterization, $\boldsymbol{\gamma}_j = \boldsymbol{0}$ and $\boldsymbol{\beta}_j = \boldsymbol{0}$ if item j does not have DIF, while $\boldsymbol{\gamma}_j = \boldsymbol{0}$ and $\boldsymbol{\beta}_j \neq \boldsymbol{0}$ if item j has uniform DIF. Similar to the multiple-group IRT approach, the distribution of $\boldsymbol{\theta}_i$ is allowed to differ across groups, i.e., $\boldsymbol{\theta}_i \sim \mathcal{N}_K(\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)$ for all $i \in I_g$, which is known as impact.

Let $K_j \subseteq \{1, \ldots, K\}$ denote the set of latent dimensions that item j loads on, $|K_j|$ denote the cardinality of the set, and define $-K_j \equiv \{1, \ldots, K\} \setminus K_j$. For any $\boldsymbol{a} \in \mathbb{R}^K$, $\boldsymbol{\gamma} \in \mathbb{R}^{P \times K}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$, let $\{\boldsymbol{a}\}_{K_j} \in \mathbb{R}^{|K_j|}$, $\{\boldsymbol{\gamma}\}_{K_j} \in \mathbb{R}^{P \times |K_j|}$ and $\{\boldsymbol{\Sigma}\}_{K_j} \in \mathbb{R}^{|K_j| \times |K_j|}$ be the slices of $\boldsymbol{a}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ that keep the rows and/or columns indicated by K_j. As explained in C. Wang et al. (2023), in a confirmatory MIRT model, if $k \notin K_j$, then $a_{jk} = 0$ and $\{\boldsymbol{\gamma}_j\}_{\{k\}} = \boldsymbol{0}$, i.e.,

5

the kth column of $\boldsymbol{\gamma}_j$ is a zero vector. Hence for each item j, we have $\{\boldsymbol{a}_j\}_{-K_j} = \boldsymbol{0}$ and $\{\boldsymbol{\gamma}_j\}_{-K_j} = \boldsymbol{0}$.

Denoting all model parameters by $\boldsymbol{\Delta} = \{\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g\}_{g=1}^G \bigcup \{\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j\}_{j=1}^J$ and the latent traits of all persons by $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_i\}_{i=1}^N$, the marginal likelihood of all the responses is

$$
\begin{aligned}
L(\boldsymbol{\Delta}) &\equiv \int_{\mathbb{R}^{N \times K}} P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \\
&= \prod_{g=1}^G \prod_{i \in I_g} \int_{\mathbb{R}^K} \left[\prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) \right] p_i(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i,
\end{aligned}
$$

where

$$
P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) = [P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i)]^{y_{ij}} [1 - P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i)]^{1-y_{ij}}
$$

is the conditional likelihood,

$$
p_i(\boldsymbol{\theta}_i) = \mathcal{N}_K(\boldsymbol{\theta}_i \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g) \tag{2}
$$

is the K-dimensional Gaussian density of $\boldsymbol{\theta}_i$, and $\bar{\boldsymbol{\mu}}_g$ and $\bar{\boldsymbol{\Sigma}}_g$ are the corresponding group-level population mean and covariance matrix respectively.

Since persons i and j are in the same group if and only if $\boldsymbol{X}_i = \boldsymbol{X}_j$, we only need to consider the case where each $\bar{\boldsymbol{X}}_g \in \mathbb{R}^{G-1}$ consists of $G-1$ dummy variables indicating the group membership, that is,

$$
\begin{bmatrix} \bar{\boldsymbol{X}}_1 & \bar{\boldsymbol{X}}_2 & \cdots & \bar{\boldsymbol{X}}_G \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I}_{G-1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.
$$

For groups $g = 2, \ldots, G$ let $\bar{\boldsymbol{\gamma}}_{gj} = \boldsymbol{\gamma}_j^\mathsf{T} \bar{\boldsymbol{X}}_g$ and $\bar{\beta}_{gj} = \boldsymbol{\beta}_j^\mathsf{T} \bar{\boldsymbol{X}}_g$ denote the DIF slope and intercept parameters of group g against group 1 on item j, respectively. As the reference group, fix $\bar{\boldsymbol{\gamma}}_{1j} = \boldsymbol{0}$ and $\bar{\beta}_{1j} = 0$ for $j = 1, \ldots, J$. Since $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$ only contain group-level dummy variables, estimating $\boldsymbol{\gamma}_j$ and $\boldsymbol{\beta}_j$ is equivalent to estimating $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$. To simplify notations,

6

we let $\boldsymbol{\gamma}_{ij} \equiv \boldsymbol{\gamma}_j^\mathsf{T} \boldsymbol{X}_i = \bar{\gamma}_{gj}$ and $\beta_{ij} \equiv \boldsymbol{\beta}_j^\mathsf{T} \boldsymbol{X}_i = \bar{\beta}_{gj}$ for $g = 1, \ldots, G$ and $i \in I_g$. Throughout this paper we will gradually add more parameters to the model, and for simplicity we always let $\boldsymbol{\Delta}$ denote the current set of all the parameters to be estimated.

The "regularized" feature of the model is reflected by the Lasso or L_1-penalized marginal log-likelihood function

$$\ell^*(\boldsymbol{\Delta}) = \log L(\boldsymbol{\Delta}) - \lambda \left(\|\bar{\boldsymbol{\gamma}}\|_1 + \|\bar{\boldsymbol{\beta}}\|_1 \right), \tag{3}$$

where

$$
\begin{aligned}
\log L(\boldsymbol{\Delta}) &= \log \int_{\mathbb{R}^{N \times K}} P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \\
&= \sum_{g=1}^{G} \sum_{i \in I_g} \log \int_{\mathbb{R}^K} \left[\prod_{j=1}^{J} P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) \right] p_i(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i, \\
\|\bar{\boldsymbol{\gamma}}\|_1 &= \sum_{g=1}^{G} \sum_{j=1}^{J} \sum_{k=1}^{K} |\bar{\gamma}_{gjk}|, \\
\|\bar{\boldsymbol{\beta}}\|_1 &= \sum_{g=1}^{G} \sum_{j=1}^{J} |\bar{\beta}_{gj}|,
\end{aligned}
\tag{4}
$$

and $\lambda > 0$ is a prespecified regularization parameter that controls sparsity (C. Wang et al., 2023). We will discuss a data-driven method which selects λ using information criteria in Section 2.4. Since (4) involves K-dimensional integrals which are intractable when K is large, directly maximizing (3) is challenging and approximation methods are needed.

2.2 Regularized Multi-Group GVEM

2.2.1 Variational Estimation

We generalize the Gaussian variational EM algorithm for MIRT models in Cho et al. (2021) to the more complex multiple-group scenario. Variational approximation methods are emerging approaches in modern statistics and machine learning for large-scale data analysis (Blei, Kucukelbir, & McAuliffe, 2017). The primary idea of GVEM is to approximate the original

7

marginal likelihood that involves intractable integrals with a computationally feasible form known as the variational lower bound.

Traditional EM algorithms require finding the posterior distributions of latent variables, $p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)$ for each person i, by Bayes' theorem within the E-step, which is intractable for large K due to the high-dimensional integral needed for computing marginal distributions. Variational EM algorithms, by contrast, approximate this unknown posterior distribution $p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)$ by a variational distribution q_i whose density is $q_i(\boldsymbol{\theta}_i)$. Then the logarithm of the integral in (4) can be written as

$$
\begin{aligned}
&\log \int_{\mathbb{R}^K} \left\{ \frac{\left[\prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) \right] p_i(\boldsymbol{\theta}_i)}{q_i(\boldsymbol{\theta}_i)} \right\} q_i(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i \\
&= \log \mathbb{E}_{\boldsymbol{\theta}_i \sim q_i} \left\{ \frac{\left[\prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) \right] p_i(\boldsymbol{\theta}_i)}{q_i(\boldsymbol{\theta}_i)} \right\},
\end{aligned}
\tag{5}
$$

where the expectation in (5) is taken with respect to the variational distribution. By Jensen's inequality, we obtain the evidence lower bound (ELBO) of (5) by switching the order of logarithm and expectation, i.e.,

$$
\begin{aligned}
&\log \mathbb{E}_{\boldsymbol{\theta}_i \sim q_i} \left\{ \frac{\left[\prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) \right] p_i(\boldsymbol{\theta}_i)}{q_i(\boldsymbol{\theta}_i)} \right\} \\
&\geq \mathbb{E}_{\boldsymbol{\theta}_i \sim q_i} \log \left\{ \frac{\left[\prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) \right] p_i(\boldsymbol{\theta}_i)}{q_i(\boldsymbol{\theta}_i)} \right\} \\
&= \sum_{j=1}^J \mathbb{E}_{\boldsymbol{\theta}_i \sim q_i} \log P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) + \mathbb{E}_{\boldsymbol{\theta}_i \sim q_i} \log p_i(\boldsymbol{\theta}_i) - \mathbb{E}_{\boldsymbol{\theta}_i \sim q_i} \log q_i(\boldsymbol{\theta}_i).
\end{aligned}
\tag{6}
$$

By carefully choosing the family of distributions where q_i is from, we hope all terms in (6) have analytical solutions such that numerical integration is not necessary. Then, we estimate parameters by maximizing the ELBO instead of the intractable marginal log-likelihood function. The performance of this strategy depends heavily on how tight this lower bound is. Actually, it can be shown that the equality holds if and only if the Kullback–Leibler (KL) divergence $\mathrm{KL}\left(q_i(\boldsymbol{\theta}_i) \,\|\, p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)\right)$ is zero, or equivalently $q_i(\boldsymbol{\theta}_i) \equiv p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)$. Therefore, the

8

key of the GVEM algorithm is to find a suitable q_i which not only approximates the posterior distribution well so that the KL divergence above is small but also leads to an ELBO that is easy to maximize. Following Cho et al. (2021), we choose q_i from the K-dimensional Gaussian family $\mathcal{N}_K(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ the mean vector and the covariance matrix of the Gaussian variational distribution respectively, and intend to maximize a variational lower bound of ELBO,

$$
\begin{aligned}
Q(\boldsymbol{\Delta}) = \sum_{i=1}^{N} \sum_{j=1}^{J} \Bigg\{ & \log \frac{\mathrm{e}^{\xi_{ij}}}{1 + \mathrm{e}^{\xi_{ij}}} + \left(\frac{1}{2} - Y_{ij} \right) \left[(b_j - \beta_{ij}) - (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\mathsf{T} \boldsymbol{\mu}_i \right] - \frac{1}{2} \xi_{ij} \\
& - \eta(\xi_{ij}) \Big[(b_j - \beta_{ij})^2 - 2(b_j - \beta_{ij})(\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\mathsf{T} \boldsymbol{\mu}_i \\
& \qquad\quad + (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\mathsf{T} \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\mathsf{T} \right) (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij}) - \xi_{ij}^2 \Big] \Bigg\} \\
& - \frac{1}{2} \sum_{g=1}^{G} \left[N_g \log \left| \bar{\boldsymbol{\Sigma}}_g \right| + \sum_{i \in I_g} \mathrm{tr} \left\{ \bar{\boldsymbol{\Sigma}}_g^{-1} \left[\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)^\mathsf{T} \right] \right\} \right],
\end{aligned}
\tag{7}
$$

where $N_g = |I_g|$ is the size of group g, ξ_{ij} is a local variational parameter that helps simplify the estimation procedure (Cho et al., 2021), and

$$
\eta(\xi) = \begin{cases} \dfrac{1}{2\xi} \left(\dfrac{1}{1 + \mathrm{e}^{-\xi}} - \dfrac{1}{2} \right), & \xi \neq 0, \\[2ex] \dfrac{1}{8}, & \xi = 0. \end{cases}
$$

In the E-step we maximize (7) with respect to each variational distribution $q_i = \mathcal{N}_K(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, which is equivalent to minimize $\mathrm{KL}\left(q_i(\boldsymbol{\theta}_i) \parallel p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)\right)$ so that $q_i(\boldsymbol{\theta}_i)$ is the best approximation of $p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)$ within the Gaussian family. This is different from the E-step in traditional EM algorithms where we let $q_i(\boldsymbol{\theta}_i) \leftarrow p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)$ be the true posterior distribution such that $\mathrm{KL}\left(q_i(\boldsymbol{\theta}_i) \parallel p(\boldsymbol{\theta}_i \mid \boldsymbol{y}_i)\right) = 0$, which is ideal but leads to difficulty in computation. In the M-step we maximize (7) with respect to model parameters, including ξ_{ij}, $\bar{\boldsymbol{\mu}}_g$, $\bar{\boldsymbol{\Sigma}}_g$, $\boldsymbol{a}_j$, b_j, $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$. In summary, the E-step and the M-step of GVEM are both "maximization" steps, but with respect to variational parameters and model parameters, respectively. Hence

9

Rijmen and Jeon (2013) referred to it as the Maximization-Maximization (MM) algorithm. For GVEM, the two steps can be combined into one joint maximization step with respect to all the parameters in $\boldsymbol{\Delta}$.

Maximizing (7) is straightforward for all the parameters except L_1-penalized $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ because we end up with a closed form updating formula for each parameter by letting the partial derivative of $Q(\boldsymbol{\Delta})$ with respect to it be zero. There are no closed form updating formulas for $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$, so we adopt a quadratic approximation approach similar to C. Wang et al. (2023): the closed form update rule for entry δ with respect to objective function f is

$$\delta \leftarrow -\frac{\mathcal{S}_\lambda\left(\dfrac{\partial Q}{\partial \delta} - \delta\dfrac{\partial^2 Q}{\partial \delta^2}\right)}{\dfrac{\partial^2 Q}{\partial \delta^2}},$$

where $\mathcal{S}_\lambda(z) = \text{sign}(z)\max(|z| - \lambda, 0)$ is a soft thresholding operator (Donoho & Johnstone, 1995). The updating formulas derived for all the parameters are shown in the supplementary material.

2.2.2   Model Identification

We need to fix $\bar{\boldsymbol{\mu}}_1 = \boldsymbol{0}$ and $\text{diag}(\bar{\boldsymbol{\Sigma}}_1) = \boldsymbol{1}$ for model identification. Here the subscript "1" denotes the reference group, and users are free to define any group as the reference. However, the model is still not identified even with these two constraints when impact is present because any group $g' \in \{2, \ldots, G\}$ can be rescaled without affecting other groups. Fixing any $\boldsymbol{u} \in \mathbb{R}_+^K$ and $\boldsymbol{v} \in \mathbb{R}^K$, for all $g = 1, \ldots, G$ and $i \in I_g$, we have the equality

$$(\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj})^\mathsf{T}\boldsymbol{\theta}_i - (b_j - \bar{\beta}_{gj}) = (\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}'_{gj})^\mathsf{T}\boldsymbol{\theta}'_i - (b_j - \bar{\beta}'_{gj}),$$

where

$$\boldsymbol{\theta}'_i = \boldsymbol{\theta}_i, \quad \bar{\boldsymbol{\gamma}}'_{gj} = \bar{\boldsymbol{\gamma}}_{gj}, \quad \bar{\beta}'_{gj} = \bar{\beta}_{gj}$$

10

when $g \neq g'$, and

$$\boldsymbol{\theta}_i' = \text{diag}(\boldsymbol{u})\boldsymbol{\theta}_i + \boldsymbol{v},$$

$$\bar{\boldsymbol{\gamma}}_{gj}' = [\text{diag}(\boldsymbol{u})]^{-1}(\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj}) - \boldsymbol{a}_j,$$

$$\bar{\beta}_{gj}' = \bar{\beta}_{gj} - (\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj})^{\mathsf{T}}[\text{diag}(\boldsymbol{u})]^{-1}\boldsymbol{v}$$

when $g = g'$. For example, under uniform DIF such that $\bar{\gamma}_{gj} = 0$, consider the unidimensional case $K = 1$ where

$$\bar{\mu}_1 = 0, \quad \bar{\mu}_2 = -1, \quad \bar{\Sigma}_1 = \bar{\Sigma}_2 = 1, \quad \bar{\beta}_{1j} = 0, \quad \bar{\beta}_{2j} = a_j.$$

For any $\theta_1 \sim \mathcal{N}(\bar{\mu}_1, \bar{\Sigma}_1)$ from group 1, consider corresponding $\theta_2 = \theta_1 - 1 \sim \mathcal{N}(\bar{\mu}_2, \bar{\Sigma}_2)$ from group 2. Since

$$a_j\theta_1 - (b_j - \bar{\beta}_{1j}) = a_j(\theta_2 + 1) - b_j = a_j\theta_2 - (b_j - \bar{\beta}_{2j}),$$

we cannot statistically distinguish between group 1 and group 2: group 2 has a lower mean θ level, but its DIF in the intercept offsets this difference. However, although the two groups are statistically equivalent, group 1 does not have DIF since it is the reference group, while group 2 has DIF in the intercepts. Intuitively, all the items are more difficult to group 2 than to group 1 (i.e., $\bar{\beta}_{2j} = a_j > 0$ for all j) is equivalent to that group 2 has a lower mean latent trait level (i.e., $\bar{\mu}_2 = \bar{\mu}_1 - 1$).

Note that the possibility that DIF in item parameters is absorbed into differences in the distributions of latent traits across groups (i.e., impact) is not a problem if we have prior information about which items are DIF-free and hence can work as anchor items; any group differences detected on these items are attributed to impact rather than DIF (J.-H. Chen, Chen, & Shih, 2014). Even without such information, identifiability is still not an issue if we can safely assume that the proportion of DIF items is not too high. Since the regularization method penalizes non-zero DIF parameters, it favors sparse models with fewer DIF items and automatically lets non-DIF items be the anchors (Y. Chen, Li, Ouyang, & Xu, 2023;

11

C. Wang et al., 2023). The only difficult case is when there are too many DIF items (e.g., DIF proportion exceeds 50%). In the simulation study below we will consider the case where 60% of the items have DIF. To distinguish between DIF and impact under such a challenging scenario, we generate balanced DIF effects (Debelak & Strobl, 2019), that is, there are both positive and negative DIF parameters that cancel out on average. Our proposed method turns out to work well on detecting such balanced DIF effects. Only if DIF occurs uniformly in one direction and DIF prevalence is higher than 50% that the method will not work well. It is worth emphasizing that the identification constraints are required by the model implied by (1) and (2), not the estimation algorithm. Regularization methods are already an improvement over other approaches that require pre-specified DIF-free items because they automatically look for them and set them as anchors.

2.2.3 Debiasing Lasso

After the EM algorithm converges, DIF parameters are determined because they have not been shrunk to exactly zero. No DIF is detected in item j if all entries in $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ have been shrunk to zero for every group g, while any non-zero entry in $\bar{\boldsymbol{\gamma}}_{gj}$ or $\bar{\beta}_{gj}$ indicates DIF in item j. However, although DIF items have been determined, Lasso penalty is known to result in biased estimators for non-zero entries (Hastie, Tibshirani, & Wainwright, 2015). To better estimate model parameters and conduct model comparison for finding the best tuning parameter λ, it is necessary to re-estimate all the non-zero entries in $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$. Following C. Wang et al. (2023), debiasing can be done by running the EM algorithm again without a penalty (i.e., $\lambda = 0$) while fixing current zero entries in $\bar{\boldsymbol{\gamma}}$'s and $\bar{\boldsymbol{\beta}}$'s at zero. Our final regularized GVEM algorithm for DIF detection is as follows:

1. Set initial values: $\boldsymbol{\mu}_i \leftarrow \boldsymbol{0}, \boldsymbol{\Sigma}_i \leftarrow \boldsymbol{I}, \xi_{ij} \leftarrow 0, \bar{\boldsymbol{\mu}}_g \leftarrow \boldsymbol{0}, \bar{\boldsymbol{\Sigma}}_g \leftarrow \boldsymbol{I}, \{\boldsymbol{a}_j\}_{K_j} \leftarrow \boldsymbol{1}, \{\boldsymbol{a}_j\}_{-K_j} \leftarrow \boldsymbol{0}, b_j \leftarrow 0, \bar{\boldsymbol{\gamma}}_{gj} \leftarrow \boldsymbol{0}$ and $\bar{\beta}_{gj} \leftarrow 0$.

2. Repeat until convergence with $\lambda = 0$ to find better initial values: in each iteration, update all the parameters using the closed form formulas.

12

3. For each λ, start from the initial values obtained from Steps 1 and 2:

   a. Repeat until convergence with the λ given: in each iteration, update all the parameters using the closed form formulas.

   b. Repeat until convergence with $\lambda = 0$ and current zero entries of $\bar{\gamma}_{gj}$ and $\bar{\beta}_{gj}$ fixed at zero: in each iteration, update all the parameters using the closed form formulas.

Note that zero entries in $\bar{\gamma}_{gj}$ and $\bar{\beta}_{gj}$ are determined by Step 3a, and Step 3b is for debiasing non-zero entries only. The choice of initial values in Step 1 is arbitrary here, and researchers are encouraged to choose initial values based on their prior information. Step 2 is not necessary but helps speed up Step 3a by starting from better initial values.

2.3 Bias Reduction via Importance Sampling

The multi-group GVEM algorithm is known to have a large bias in discrimination parameters, especially when the latent traits of different dimensions are highly correlated and the sample size is not large enough (Cho et al., 2021), so it may not perform as well when detecting non-uniform DIF. To reduce bias in model parameter estimates, we employ an additional importance sampling step after GVEM converges to find a better approximation of the marginal log-likelihood function than the variational lower bound. This idea has recently been used in Ma et al. (2023) to reduce the estimation bias of GVEM for (single-group) MIRT models.

Recall that we obtained a lower bound of the marginal log-likelihood function (4) by Jensen's inequality,

$$\mathbb{E} \log X \leq \log \mathbb{E} X = \mathbb{E} \log \mathbb{E} X,$$

where the last equality holds because $\log \mathbb{E} X$ is a constant. To obtain a tighter bound, we hope to find a random variable Y such that

$$\mathbb{E} \log X \leq \mathbb{E} \log Y \leq \mathbb{E} \log \mathbb{E} X,$$

13

so a natural choice is the empirical mean, i.e.,

$$Y = \frac{1}{M} \sum_{m=1}^{M} X^{(m)},$$

where $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$ are independent and have the same distribution as X. Applying this idea to the marginal log-likelihood function $\log L(\boldsymbol{\Delta})$ in (4), we can sample from the estimated variational distributions of latent traits and use the importance sampling weighted samples to approximate a tighter variational lower bound than (7). More specifically, after the GVEM algorithm converges, in an additional E-step we draw $S \times M$ samples $\{\boldsymbol{\theta}_i^{(s,m)}\}_{s=1\ m=1}^{S\ \ M}$ from estimated $q_i(\boldsymbol{\theta}_i) = \mathcal{N}_K(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ for each person i. With these samples, we have the following improved variational lower bound:

$$
\begin{aligned}
\log L(\boldsymbol{\Delta}) &= \sum_{g=1}^{G} \sum_{i \in I_g} \log \mathbb{E}_{\boldsymbol{\theta}_i \sim q_i} \left\{ \frac{\left[\prod_{j=1}^{J} P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i)\right] p_i(\boldsymbol{\theta}_i)}{q_i(\boldsymbol{\theta}_i)} \right\} \\
&\geq \sum_{g=1}^{G} \sum_{i \in I_g} \mathbb{E}_{\{\boldsymbol{\theta}_i^{(m)}\}_{m=1}^{M} \sim q_i} \log \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{\left[\prod_{j=1}^{J} P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i^{(m)})\right] p_i(\boldsymbol{\theta}_i^{(m)})}{q_i(\boldsymbol{\theta}_i^{(m)})} \right\} \\
&\approx \sum_{g=1}^{G} \sum_{i \in I_g} \frac{1}{S} \sum_{s=1}^{S} \log \left\{ \frac{1}{M} \sum_{m=1}^{M} \frac{\left[\prod_{j=1}^{J} P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i^{(s,m)})\right] p_i(\boldsymbol{\theta}_i^{(s,m)})}{q_i(\boldsymbol{\theta}_i^{(s,m)})} \right\} \\
&\equiv \hat{Q}(\boldsymbol{\Delta}).
\end{aligned}
\tag{8}
$$

When $S \to \infty$ and $M \to \infty$, it can be shown that $\hat{Q}(\boldsymbol{\Delta})$ converges in probability to the marginal log-likelihood function $\log L(\boldsymbol{\Delta})$ (Burda, Grosse, & Salakhutdinov, 2016). In the simulation study we will show that even small values like $S = M = 10$ can lead to huge improvement and satisfactory performance. The new objective function to be maximized in the two additional M-steps now becomes

$$\hat{Q}^*(\boldsymbol{\Delta}) = \hat{Q}(\boldsymbol{\Delta}) - \lambda \left(\|\bar{\boldsymbol{\gamma}}\|_1 + \|\bar{\boldsymbol{\beta}}\|_1 \right). \tag{9}$$

Due to its complexity, there are no closed form updating formulas as in GVEM, so instead

14

we employ gradient-based optimization algorithms.

To ensure the positive definiteness of $\bar{\boldsymbol{\Sigma}}_g$, we conduct Cholesky decomposition $\bar{\boldsymbol{\Sigma}}_g = \bar{\boldsymbol{L}}_g \bar{\boldsymbol{L}}_g^\mathsf{T}$ and maximize (9) with respect to $\bar{\boldsymbol{L}}_g$ instead of $\bar{\boldsymbol{\Sigma}}_g$. Furthermore, we let $\bar{\boldsymbol{\mu}}_1 = \boldsymbol{0}$ and fix $\mathrm{diag}(\bar{\boldsymbol{\Sigma}}_1) = \mathrm{diag}(\bar{\boldsymbol{L}}_1 \bar{\boldsymbol{L}}_1^\mathsf{T}) = \boldsymbol{1}$ by utilizing the transformation

$$
\bar{\boldsymbol{L}}_1 = \begin{bmatrix}
1 & & & & \\
u_{21} & \sqrt{1 - u_{21}^2} & & & \\
u_{31} & u_{32}\sqrt{1 - u_{31}^2} & \sqrt{(1 - u_{31})^2(1 - u_{32})^2} & & \\
\vdots & \vdots & \vdots & \ddots & \\
u_{K1} & u_{K2}\sqrt{1 - u_{K1}^2} & u_{K3}\sqrt{(1 - u_{K1})^2(1 - u_{K2})^2} & \cdots & \sqrt{\prod_{k=1}^K (1 - u_{Kk})^2}
\end{bmatrix},
$$

where $u_{k\ell} = \tanh v_{k\ell} \in (-1, 1)$, and $v_{k\ell} \in \mathbb{R}$ for $k = 2, \ldots, K$ and $\ell = 1, \ldots, i - 1$ (Lewandowski, Kurowicka, & Joe, 2009). Gradients can be computed implicitly by automatic differentiation, such as using the torch package in R (Falbel & Luraschi, 2023).

We apply the Adam optimization algorithm by Kingma and Ba (2014), a popular optimizer in deep learning, to minimize $\hat{Q}(\boldsymbol{\Delta})$ with respect to all the model parameters: $\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{L}}_g, v_{k\ell}, \boldsymbol{a}_j, b_j, \bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$. Adam computes the adaptive learning rate for each parameter based on moving averages of the first and second moments of the gradient, which helps avoid the difficulty in choosing a single proper learning rate for all the parameters. Since $\hat{Q}^*(\boldsymbol{\Delta})$ has additional penalty terms that are not differentiable but convex, we apply the proximal gradient method (PGM; Hastie et al., 2015) to $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$: each entry δ with penalty $\lambda|\delta|$ is updated by

$$
\delta \leftarrow \mathcal{S}_{s\lambda}(\delta), \tag{10}
$$

where s is the adaptive learning rate used to update δ in Adam.

Similar to the regularized GVEM algorithm, we maximize $\hat{Q}^*(\boldsymbol{\Delta})$ twice in two consecutive M-steps, the first one with a penalty and the second one without a penalty but fixing current zero entries in $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$, determined by the first one, at zero. Moreover, we found through

15

simulation that the GVEM algorithm without penalty provides better initial values to the following importance sampling procedure because regularized GVEM does not detect DIF items well and hence gives inaccurate variational distributions of latent traits that importance sampling is based on. Our final algorithm, named "importance-weighted Gaussian variational expectation-maximization-maximization" (IW-GVEMM), is as follows:

1. Obtain initial values: run Steps 1 and 2 of the GVEM algorithm.

2. Conduct Cholesky decomposition and inversely transform parameters: $\bar{\boldsymbol{\Sigma}}_g = \bar{\boldsymbol{L}}_g \bar{\boldsymbol{L}}_g^{\mathsf{T}}$ and compute $v_{k\ell}$ from $\bar{\boldsymbol{L}}_1$.

3. Draw random samples: draw $\boldsymbol{\theta}_i^{(s,m)} \sim \mathcal{N}_K(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

4. For each λ, start from the initial values obtained from Steps 1 to 3:

   a. Repeat until convergence with the λ given: in each iteration, update $\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{L}}_g, v_{k\ell}, \boldsymbol{a}_j, b_j, \bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ using Adam, and then update $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ using (10).

   b. Repeat until convergence with $\lambda = 0$ and current zero entries of $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ fixed at zero: in each iteration, update $\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{L}}_g, v_{k\ell}, \boldsymbol{a}_j, b_j, \bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ using Adam.

To avoid randomness in the E-step of the importance sampling, we only sample $\boldsymbol{\theta}_i^{(s,m)}$'s once in Step 3 and then fix them for all values of λ when running Step 4. Since Steps 1 to 3 do not depend on λ, we only need to run them once for each dataset.

Compared to the Lasso EMM method proposed by C. Wang et al. (2023) which maximizes the objective function (3) on K-dimensional Gaussian quadrature using the Newton-Raphson method, the main advantage of this regularized IW-GVEMM method is that it better handles higher dimensional latent traits because it does not need to compute K-dimensional numerical integrals or invert K-dimensional matrices, which become very slow and numerically unstable for large K.

2.4 Information Criteria for Tuning Parameter Selection

We use information criteria to find the best tuning parameter λ in this study. The marginal log-likelihood $\log L(\boldsymbol{\Delta})$ in (4) is difficult to compute because it involves K-dimensional numerical integration, but its (approximate) variational lower bounds $Q(\boldsymbol{\Delta})$ in (7) and $\hat{Q}(\boldsymbol{\Delta})$ in (8) are by-products of our proposed algorithms. Consequently, we modify the generalized information criterion (GIC; Zhang, Li, & Tsai, 2012)

$$
\begin{aligned}
\mathrm{GIC} = & -2\log L(\boldsymbol{\Delta}) + k_N \cdot \ell_0(\boldsymbol{\Delta}) \\
= & -2\sum_{g=1}^{G}\sum_{i \in I_g} \log \int \prod_{j=1}^{J} P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i) p_i(\boldsymbol{\theta}_i) \mathrm{d}\boldsymbol{\theta}_i + k_N \cdot \ell_0(\boldsymbol{\Delta})
\end{aligned}
$$

by replacing $\log L(\boldsymbol{\Delta})$ with $Q(\boldsymbol{\Delta})$ as

$$
\begin{aligned}
\mathrm{GIC} = & -2Q(\boldsymbol{\Delta}) + k_N \left(\|\bar{\boldsymbol{\gamma}}\|_0 + \|\bar{\boldsymbol{\beta}}\|_0 \right) \\
= & -2\sum_{i=1}^{N}\sum_{j=1}^{J} \left\{ \log \frac{\mathrm{e}^{\xi_{ij}}}{1+\mathrm{e}^{\xi_{ij}}} + \left(\frac{1}{2} - Y_{ij}\right)\left[(b_j - \beta_{ij}) - (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^{\mathsf{T}}\boldsymbol{\mu}_i\right] - \frac{1}{2}\xi_{ij} \right\} \\
& + \sum_{g=1}^{G} \left[N_g \log \left|\bar{\boldsymbol{\Sigma}}_g\right| + \sum_{i \in I_g} \mathrm{tr}\left\{ \bar{\boldsymbol{\Sigma}}_g^{-1}\left[\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)^{\mathsf{T}}\right]\right\}\right] + k_N \cdot \ell_0(\boldsymbol{\Delta})
\end{aligned}
$$

for GVEM, and by replacing $\log L(\boldsymbol{\Delta})$ with $\hat{Q}(\boldsymbol{\Delta})$ as

$$
\begin{aligned}
\mathrm{GIC} = & -2\hat{Q}(\boldsymbol{\Delta}) + k_N \cdot \ell_0(\boldsymbol{\Delta}) \\
= & -\frac{2}{S}\sum_{g=1}^{G}\sum_{i \in I_g}\sum_{s=1}^{S} \log \left\{ \frac{1}{M}\sum_{m=1}^{M} \frac{\left[\prod_{j=1}^{J} P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}_i^{(s,m)})\right] p_i(\boldsymbol{\theta}_i^{(s,m)})}{q_i(\boldsymbol{\theta}_i^{(s,m)})} \right\} + k_N \cdot \ell_0(\boldsymbol{\Delta})
\end{aligned}
$$

for IW-GVEMM, where

$$
\begin{aligned}
\ell_0(\boldsymbol{\Delta}) = & \|\bar{\boldsymbol{\gamma}}\|_0 + \|\bar{\boldsymbol{\beta}}\|_0 \\
= & \sum_{g=1}^{G}\sum_{j=1}^{J}\sum_{k=1}^{K} \mathbb{1}\{\bar{\gamma}_{gjk} \neq 0\} + \sum_{g=1}^{G}\sum_{j=1}^{J} \mathbb{1}\{\bar{\beta}_{gj} \neq 0\}
\end{aligned}
$$

17

is the number of non-zero DIF parameters and k_N is an increasing function of N. In particular, GIC becomes BIC by taking $k_N = \log N$. Our simulation study shows that BIC does not penalize DIF parameters strongly enough and leads to too many false positives under some scenarios. Therefore, we also use $k_N = c \log N \log \log N$ where $c > 0$ is a prespecified constant that controls the magnitude of penalty, i.e., larger c indicates a higher penalty and shrinks more parameters toward zero.

We first apply the GVEM and the IW-GVEMM algorithms with different values of λ, and after all the estimation is done, we choose the λ with the lowest information criteria (BIC or GIC). Note that c is a constant for model comparison using GIC rather than a model parameter that affects the estimation. Our simulation study shows that $\hat{Q}(\boldsymbol{\Delta})$ works as a good proxy of $\log L(\boldsymbol{\Delta})$ for selecting the best tuning parameter for IW-GVEMM that helps detect DIF.

3   Simulation

Two simulation studies are conducted to examine the performance of GVEM and IW-GVEMM algorithms for DIF detection in two-parameter re-MIRT models. Study I focuses on uniform DIF detection and study II focuses on non-uniform DIF detection. In both studies, we set $G = 3$ groups, one reference group and two focal groups, where the first focal group has low DIF and the second has high DIF. The latent traits $\boldsymbol{\theta}$ of all the three groups are generated from

$$\mathcal{N}_2\left(\mathbf{0}, \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}\right)$$

when $K = 2$ and from

$$\mathcal{N}_3\left(\mathbf{0}, \begin{bmatrix} 1 & 0.85 & 0.85 \\ 0.85 & 1 & 0.85 \\ 0.85 & 0.85 & 1 \end{bmatrix}\right)$$

18

when $K = 3$, i.e., there is no impact for the two focal groups. Such high correlations among latent dimensions are not uncommon in practice (W.-C. Wang, Chen, & Cheng, 2004), and prior studies showed that high correlations like 0.85 result in more difficulty in model estimation and DIF detection compared to low correlations like 0.25 (Cho et al., 2021; C. Wang et al., 2023). Our pilot study suggests that the correlation has little effect on the running time and the DIF detection accuracy of the proposed approaches. The test length is fixed at $J = 10K$, and each dimension corresponds to 10 items that load solely on this dimension. In both studies, for the reference group, slopes $\boldsymbol{a}$ are generated from $\mathcal{U}(1.5, 2.5)$ and intercepts $\boldsymbol{b}$ are generated from $\mathcal{N}(0, 1)$. To evaluate the magnitude of DIF, we compute wABC, the area between expected item score curves for the reference and the focal groups (Edelen, Stucky, & Chandra, 2015). Two factors, sample size n of each group (500 and 1000) and proportion of DIF items (20% and 60%), are manipulated. For each simulation condition we run 100 replications. Our convergence criterion is that the absolute difference of every entry δ of all the parameters between consecutive iterations (i.e., $|\delta^{(t)} - \delta^{(t-1)}|$ at the tth iteration) is less than 0.001.

For comparison, we also apply the Lasso EMM method proposed by C. Wang et al. (2023) to the same simulated data. Same as GVEM and IW-GVEMM, EMM also tries to maximize the objective function in (3), but uses Gaussian quadrature rather than Gaussian variational approximation to deal with the integrals. Denser quadrature approximates the integral better and results in higher accuracy and longer computation time. We construct multidimensional Gaussian quadrature using the nested Gauss-Hermite rule (Genz & Keister, 1996) and the sparse combination technique (Heiss & Winschel, 2008). As a result, there are 133 and 703 grid points for $K = 2$ and $K = 3$ respectively. More grid points help achieve higher accuracy, but as will be shown later, even with such small numbers of grid points, EMM is much slower than our proposed methods. It is worth noting that since latent dimensions are highly correlated with each other, in several replications the estimates of group-level covariance matrices can be nearly singular after the M-steps of EMM, which

19

makes estimation difficult. For such replications, we have to increase the numbers of grid points to 445 for $K = 2$ and 4191 for $K = 3$ to make EMM work, and they require much longer running time.

Following Cho, Xiao, Wang, and Xu (2024), we set $c = 1$ for GIC. We will consider some other possible strategies for choosing c in the next section on real data analysis. Within each replication, we estimate the model with eight λ values from $\left\{0.1\sqrt{N}, 0.2\sqrt{N}, \ldots, 0.8\sqrt{N}\right\}$ first. If the best model corresponds to $\lambda = 0.8\sqrt{N}$, then we additionally estimate the model with larger tuning parameters, $0.9\sqrt{N}, 1.0\sqrt{N}, 1.1\sqrt{N}, \ldots$, until the best model does not correspond to the largest λ. Under the simplest condition ($n = 500$ and 20% DIF), the mean running times (in seconds) of the first five replications over the eight λ's on a MacBook Pro with M3 Max are shown in Table 1, where the numbers of grid points for EMM are 133 and 703 for $K = 2$ and $K = 3$. The two proposed methods show a clear advantage in efficiency compared to the EMM method, especially when the dimension of the latent traits grows. GVEM is faster than IW-GVEMM, but as will be shown later, IW-GVEMM is more accurate. Given the long running time of EMM, we use multiple computers to run the remaining replications, so their running times are not comparable.

Table 1: Mean Running Times (in Seconds) of the First Five Replications

K	GVEM	IW-GVEMM	EMM
2	5.61	66.98	168.13
3	8.44	85.94	2288.70

3.1 Simulation I: Uniform DIF

Under the uniform DIF condition, the slopes for the two focal groups are equal to those for the reference group even for DIF items. Table 2 shows the DIF parameters and the mean wABCs of each condition.

Tables 3 and 4 show the true and the false positive rates of DIF detection across 100 replications, where standard deviations are shown in parentheses. Besides low and high

20

Table 2: DIF Parameters in Simulation Study I

	First Half of DIF Items			Second Half of DIF Items		
Group	$\bar{\gamma}_{gj}$	$\bar{\beta}_{gj}$	Mean wABC	$\bar{\gamma}_{gj}$	$\bar{\beta}_{gj}$	Mean wABC
Low DIF	0	0.5	0.070	0	-0.5	0.070
High DIF	0	1	0.138	0	-1	0.138

DIF groups, we also show whether items are marked as DIF regardless of low or high DIF group as "Total". It turns out that importance sampling leads to a huge improvement: true positive rates are much higher, and false positive rates are similar or lower except for BIC with 60% DIF. Moreover, IW-GVEMM has a similar performance to EMM but runs much faster. EMM is better at detecting low DIF under 20% DIF conditions, but this pattern is reversed under 60% DIF. One possible reason is that due to long running time we do not use large numbers of grid points for EMM except for several replications where group-level covariance matrices become singular. With more grid points, EMM is expected to be more accurate, but still it is unlikely that EMM will show very obvious advantages in accuracy over IW-GVEMM. It is worth noting that since IW-GVEMM works on the Cholesky factors of the covariance matrices, it is more robust to high correlations among latent dimensions than EMM. This also helps explain why the performance of EMM is not consistently better than its approximation IW-GVEMM. We found in the pilot study that EMM tends to have better performance than IW-GVEMM when the correlations among the latent traits are lower, which agrees with our explanation here. GIC leads to both lower true positive rates and lower false positive rates in all conditions than BIC, which is expected because GIC penalizes more severely and shrinks more DIF parameters to zero. BIC generally works well for 20% DIF but leads to high false positive rates for 60% DIF; GIC with $c = 1$ controls false positive rates in all conditions but has difficulty detecting low DIF. In practice, researchers may apply the methods proposed in the next section to find a better c for GIC that achieves balance between the true and the false positive rates. Unsurprisingly, the larger sample size leads to higher true positive rates but also slightly higher false positive rates with GIC. Both the proportions of DIF items and the numbers of latent dimensions result in mixed

differences.

Table 3: Means (Standard Deviations) of True Positive Rates across Replications of Simulation Study I

K	n	DIF	Group	GVEM BIC	GVEM GIC	IW-GVEMM BIC	IW-GVEMM GIC	EMM BIC	EMM GIC
2	500	20%	Total	0.715 (0.304)	0.642 (0.314)	0.973 (0.079)	0.882 (0.157)	0.967 (0.092)	0.865 (0.172)
			Low	0.235 (0.321)	0.100 (0.216)	0.340 (0.285)	0.112 (0.182)	0.368 (0.267)	0.132 (0.202)
			High	0.715 (0.304)	0.642 (0.314)	0.973 (0.079)	0.882 (0.157)	0.967 (0.092)	0.865 (0.172)
		60%	Total	0.686 (0.280)	0.517 (0.261)	0.992 (0.028)	0.905 (0.119)	0.983 (0.046)	0.828 (0.166)
			Low	0.169 (0.256)	0.039 (0.124)	0.475 (0.164)	0.123 (0.131)	0.486 (0.193)	0.108 (0.128)
			High	0.686 (0.280)	0.517 (0.261)	0.992 (0.028)	0.905 (0.119)	0.983 (0.046)	0.828 (0.166)
	1000	20%	Total	0.887 (0.234)	0.848 (0.266)	0.998 (0.025)	0.990 (0.049)	0.993 (0.043)	0.988 (0.055)
			Low	0.450 (0.364)	0.333 (0.343)	0.733 (0.259)	0.452 (0.260)	0.767 (0.217)	0.472 (0.280)
			High	0.887 (0.234)	0.848 (0.266)	0.998 (0.025)	0.990 (0.049)	0.993 (0.043)	0.988 (0.055)
		60%	Total	0.867 (0.216)	0.811 (0.261)	1.000 (0.000)	0.999 (0.008)	1.000 (0.000)	0.998 (0.014)
			Low	0.449 (0.402)	0.285 (0.330)	0.843 (0.120)	0.583 (0.153)	0.754 (0.152)	0.482 (0.163)
			High	0.867 (0.216)	0.811 (0.261)	1.000 (0.000)	0.999 (0.008)	1.000 (0.000)	0.998 (0.014)
3	500	20%	Total	0.725 (0.302)	0.565 (0.333)	0.982 (0.062)	0.877 (0.155)	0.972 (0.082)	0.877 (0.149)
			Low	0.217 (0.284)	0.078 (0.168)	0.358 (0.252)	0.070 (0.130)	0.367 (0.225)	0.127 (0.161)
			High	0.725 (0.302)	0.565 (0.333)	0.982 (0.062)	0.877 (0.155)	0.970 (0.083)	0.877 (0.149)
		60%	Total	0.714 (0.244)	0.516 (0.266)	0.993 (0.018)	0.909 (0.116)	0.964 (0.084)	0.833 (0.174)
			Low	0.151 (0.240)	0.045 (0.127)	0.468 (0.125)	0.104 (0.110)	0.438 (0.171)	0.100 (0.109)
			High	0.713 (0.244)	0.516 (0.266)	0.993 (0.018)	0.909 (0.116)	0.953 (0.112)	0.833 (0.173)
	1000	20%	Total	0.857 (0.212)	0.805 (0.273)	1.000 (0.000)	0.995 (0.029)	1.000 (0.000)	0.992 (0.037)
			Low	0.378 (0.363)	0.290 (0.319)	0.735 (0.198)	0.452 (0.231)	0.732 (0.198)	0.452 (0.259)
			High	0.857 (0.212)	0.805 (0.273)	1.000 (0.000)	0.995 (0.029)	1.000 (0.000)	0.992 (0.037)
		60%	Total	0.839 (0.219)	0.783 (0.270)	0.999 (0.006)	0.996 (0.015)	0.999 (0.008)	0.995 (0.021)
			Low	0.343 (0.348)	0.202 (0.243)	0.819 (0.092)	0.500 (0.166)	0.671 (0.139)	0.429 (0.131)
			High	0.839 (0.219)	0.783 (0.270)	0.999 (0.006)	0.996 (0.015)	0.999 (0.008)	0.994 (0.022)

23

Table 4: Means (Standard Deviations) of False Positive Rates in Simulation Study I

K	n	DIF	Group	GVEM BIC	GVEM GIC	IW-GVEMM BIC	IW-GVEMM GIC	EMM BIC	EMM GIC
2	500	20%	Total	0.064 (0.125)	0.016 (0.048)	0.047 (0.059)	0.002 (0.011)	0.051 (0.063)	0.002 (0.012)
			Low	0.036 (0.082)	0.007 (0.026)	0.024 (0.037)	0.001 (0.009)	0.026 (0.042)	0.001 (0.009)
			High	0.036 (0.085)	0.009 (0.032)	0.024 (0.044)	0.001 (0.006)	0.025 (0.044)	0.001 (0.009)
		60%	Total	0.041 (0.117)	0.004 (0.028)	0.106 (0.104)	0.005 (0.025)	0.164 (0.189)	0.014 (0.061)
			Low	0.019 (0.065)	0.000 (0.000)	0.064 (0.076)	0.001 (0.013)	0.111 (0.143)	0.004 (0.021)
			High	0.025 (0.077)	0.004 (0.028)	0.045 (0.078)	0.004 (0.021)	0.064 (0.117)	0.010 (0.049)
	1000	20%	Total	0.087 (0.131)	0.024 (0.052)	0.064 (0.065)	0.009 (0.024)	0.058 (0.069)	0.006 (0.018)
			Low	0.047 (0.091)	0.010 (0.028)	0.029 (0.043)	0.006 (0.021)	0.031 (0.047)	0.004 (0.016)
			High	0.044 (0.079)	0.015 (0.039)	0.035 (0.051)	0.003 (0.014)	0.027 (0.049)	0.001 (0.009)
		60%	Total	0.084 (0.179)	0.015 (0.045)	0.085 (0.108)	0.015 (0.041)	0.141 (0.156)	0.020 (0.049)
			Low	0.052 (0.131)	0.010 (0.038)	0.039 (0.073)	0.005 (0.025)	0.101 (0.150)	0.013 (0.042)
			High	0.050 (0.126)	0.005 (0.025)	0.049 (0.087)	0.010 (0.034)	0.045 (0.082)	0.007 (0.030)
3	500	20%	Total	0.055 (0.085)	0.013 (0.038)	0.068 (0.071)	0.003 (0.011)	0.056 (0.056)	0.005 (0.016)
			Low	0.023 (0.045)	0.005 (0.023)	0.030 (0.045)	0.001 (0.006)	0.033 (0.044)	0.003 (0.012)
			High	0.035 (0.058)	0.010 (0.024)	0.040 (0.048)	0.002 (0.009)	0.025 (0.036)	0.002 (0.009)
		60%	Total	0.036 (0.089)	0.006 (0.024)	0.123 (0.106)	0.008 (0.025)	0.258 (0.260)	0.024 (0.060)
			Low	0.022 (0.058)	0.003 (0.016)	0.071 (0.075)	0.006 (0.021)	0.154 (0.178)	0.005 (0.020)
			High	0.020 (0.059)	0.003 (0.016)	0.060 (0.079)	0.003 (0.014)	0.133 (0.227)	0.019 (0.058)
	1000	20%	Total	0.032 (0.073)	0.011 (0.030)	0.047 (0.047)	0.010 (0.020)	0.056 (0.080)	0.003 (0.011)
			Low	0.019 (0.053)	0.005 (0.021)	0.025 (0.031)	0.005 (0.013)	0.028 (0.042)	0.001 (0.006)
			High	0.016 (0.034)	0.005 (0.015)	0.023 (0.029)	0.005 (0.015)	0.030 (0.052)	0.002 (0.009)
		60%	Total	0.059 (0.155)	0.011 (0.039)	0.108 (0.090)	0.018 (0.042)	0.180 (0.209)	0.027 (0.053)
			Low	0.034 (0.112)	0.004 (0.022)	0.055 (0.073)	0.011 (0.031)	0.150 (0.210)	0.021 (0.046)
			High	0.034 (0.098)	0.007 (0.023)	0.058 (0.065)	0.007 (0.023)	0.038 (0.059)	0.008 (0.028)

24

3.2 Simulation II: Non-Uniform DIF

In the second simulation study, there are DIF effects on both intercepts and slopes, which are shown in Table 5. The true and false positive rates of DIF detection across replications are shown in Tables 6 and 7. All three algorithms perform worse due to the more complex model setting, but the general patterns are largely similar to the uniform DIF simulation study: IW-GVEMM and EMM perform similarly, and both have better performance than GVEM, GIC penalizes more than BIC, and true positive rates increase with larger sample sizes. Besides, although IW-GVEMM and EMM are still good at detecting high DIF, DIF is mostly detected on the intercept $\bar{\boldsymbol{\beta}}$ rather than the slope $\bar{\boldsymbol{\gamma}}$. It is less of a problem in practice because DIF in slopes usually comes with DIF in intercepts.

Table 5: DIF Parameters in Simulation Study II

Group	First Half of DIF Items			Second Half of DIF Items		
	$\bar{\gamma}_{gj}$	$\bar{\beta}_{gj}$	Mean wABC	$\bar{\gamma}_{gj}$	$\bar{\beta}_{gj}$	Mean wABC
Low DIF	-0.4	0.5	0.079	0.4	-0.5	0.065
High DIF	-0.8	1	0.170	0.8	-1	0.117

25

Table 6: Means (Standard Deviations) of True Positive Rates in Simulation Study II

K	n	DIF	Group	GVEM BIC	GVEM GIC	IW-GVEMM BIC	IW-GVEMM GIC	EMM BIC	EMM GIC
2	500	20%	Total	0.683 (0.288)	0.603 (0.291)	0.913 (0.148)	0.752 (0.218)	0.905 (0.150)	0.743 (0.202)
			Low	0.145 (0.264)	0.062 (0.160)	0.380 (0.285)	0.105 (0.164)	0.377 (0.245)	0.117 (0.168)
			High	0.683 (0.288)	0.603 (0.291)	0.913 (0.148)	0.752 (0.218)	0.902 (0.150)	0.743 (0.202)
		60%	Total	0.658 (0.215)	0.521 (0.206)	0.940 (0.073)	0.760 (0.132)	0.928 (0.088)	0.680 (0.171)
			Low	0.128 (0.239)	0.037 (0.129)	0.491 (0.165)	0.102 (0.132)	0.469 (0.184)	0.089 (0.138)
			High	0.657 (0.214)	0.518 (0.203)	0.939 (0.073)	0.760 (0.132)	0.925 (0.088)	0.680 (0.171)
	1000	20%	Total	0.800 (0.241)	0.755 (0.256)	0.975 (0.083)	0.928 (0.130)	0.978 (0.080)	0.935 (0.121)
			Low	0.325 (0.392)	0.198 (0.332)	0.678 (0.226)	0.410 (0.274)	0.702 (0.218)	0.447 (0.267)
			High	0.800 (0.241)	0.755 (0.256)	0.975 (0.083)	0.928 (0.130)	0.978 (0.080)	0.935 (0.121)
		60%	Total	0.756 (0.227)	0.686 (0.215)	0.980 (0.049)	0.923 (0.096)	0.978 (0.054)	0.923 (0.096)
			Low	0.312 (0.389)	0.113 (0.234)	0.759 (0.142)	0.409 (0.163)	0.735 (0.159)	0.432 (0.185)
			High	0.756 (0.227)	0.686 (0.215)	0.979 (0.049)	0.923 (0.096)	0.978 (0.058)	0.923 (0.096)
3	500	20%	Total	0.732 (0.237)	0.603 (0.259)	0.947 (0.097)	0.760 (0.182)	0.932 (0.106)	0.742 (0.186)
			Low	0.195 (0.304)	0.067 (0.177)	0.385 (0.244)	0.110 (0.181)	0.368 (0.214)	0.122 (0.164)
			High	0.730 (0.235)	0.602 (0.257)	0.938 (0.102)	0.758 (0.181)	0.927 (0.109)	0.740 (0.184)
		60%	Total	0.649 (0.204)	0.517 (0.207)	0.939 (0.059)	0.746 (0.134)	0.911 (0.100)	0.674 (0.150)
			Low	0.111 (0.213)	0.020 (0.079)	0.482 (0.118)	0.093 (0.123)	0.431 (0.172)	0.081 (0.105)
			High	0.649 (0.203)	0.516 (0.206)	0.937 (0.059)	0.746 (0.134)	0.898 (0.107)	0.673 (0.151)
	1000	20%	Total	0.760 (0.244)	0.732 (0.246)	0.982 (0.052)	0.932 (0.106)	0.968 (0.070)	0.917 (0.112)
			Low	0.307 (0.392)	0.162 (0.279)	0.702 (0.237)	0.378 (0.251)	0.708 (0.204)	0.390 (0.237)
			High	0.760 (0.244)	0.732 (0.246)	0.982 (0.052)	0.932 (0.106)	0.967 (0.071)	0.917 (0.112)
		60%	Total	0.746 (0.201)	0.668 (0.206)	0.979 (0.034)	0.935 (0.068)	0.976 (0.054)	0.918 (0.106)
			Low	0.229 (0.341)	0.072 (0.156)	0.764 (0.109)	0.443 (0.163)	0.721 (0.145)	0.431 (0.157)
			High	0.746 (0.201)	0.668 (0.205)	0.978 (0.035)	0.935 (0.068)	0.973 (0.055)	0.918 (0.106)

Table 7: Means (Standard Deviations) of False Positive Rates in Simulation Study II

K	n	DIF	Group	GVEM BIC	GVEM GIC	IW-GVEMM BIC	IW-GVEMM GIC	EMM BIC	EMM GIC
2	500	20%	Total	0.031 (0.080)	0.008 (0.035)	0.056 (0.068)	0.002 (0.011)	0.060 (0.070)	0.002 (0.011)
			Low	0.013 (0.041)	0.004 (0.026)	0.025 (0.042)	0.000 (0.000)	0.031 (0.047)	0.001 (0.006)
			High	0.018 (0.050)	0.004 (0.016)	0.031 (0.045)	0.002 (0.011)	0.030 (0.048)	0.001 (0.009)
		60%	Total	0.041 (0.122)	0.007 (0.053)	0.123 (0.112)	0.002 (0.018)	0.145 (0.158)	0.005 (0.025)
			Low	0.021 (0.073)	0.004 (0.021)	0.061 (0.090)	0.001 (0.013)	0.086 (0.120)	0.001 (0.013)
			High	0.026 (0.098)	0.004 (0.038)	0.064 (0.086)	0.001 (0.013)	0.066 (0.110)	0.004 (0.021)
	1000	20%	Total	0.071 (0.160)	0.016 (0.064)	0.065 (0.077)	0.010 (0.026)	0.058 (0.070)	0.007 (0.020)
			Low	0.042 (0.116)	0.007 (0.031)	0.035 (0.058)	0.004 (0.018)	0.035 (0.056)	0.004 (0.016)
			High	0.039 (0.093)	0.009 (0.040)	0.031 (0.044)	0.006 (0.020)	0.023 (0.040)	0.003 (0.014)
		60%	Total	0.106 (0.247)	0.009 (0.041)	0.136 (0.134)	0.010 (0.034)	0.151 (0.185)	0.013 (0.038)
			Low	0.071 (0.176)	0.002 (0.025)	0.065 (0.098)	0.002 (0.018)	0.110 (0.159)	0.006 (0.027)
			High	0.060 (0.152)	0.006 (0.033)	0.078 (0.088)	0.007 (0.030)	0.050 (0.110)	0.006 (0.027)
3	500	20%	Total	0.054 (0.129)	0.007 (0.028)	0.063 (0.061)	0.005 (0.014)	0.064 (0.075)	0.005 (0.018)
			Low	0.031 (0.084)	0.003 (0.017)	0.035 (0.043)	0.003 (0.012)	0.032 (0.066)	0.001 (0.006)
			High	0.032 (0.087)	0.005 (0.018)	0.031 (0.042)	0.002 (0.009)	0.032 (0.043)	0.004 (0.017)
		60%	Total	0.032 (0.118)	0.003 (0.019)	0.135 (0.106)	0.010 (0.030)	0.198 (0.216)	0.018 (0.045)
			Low	0.017 (0.080)	0.002 (0.017)	0.060 (0.069)	0.003 (0.016)	0.129 (0.165)	0.005 (0.023)
			High	0.019 (0.070)	0.001 (0.008)	0.078 (0.077)	0.007 (0.023)	0.087 (0.160)	0.013 (0.036)
	1000	20%	Total	0.062 (0.131)	0.011 (0.045)	0.065 (0.064)	0.004 (0.015)	0.057 (0.057)	0.005 (0.014)
			Low	0.028 (0.063)	0.006 (0.027)	0.035 (0.041)	0.003 (0.012)	0.031 (0.036)	0.002 (0.010)
			High	0.037 (0.084)	0.006 (0.026)	0.032 (0.041)	0.002 (0.010)	0.028 (0.041)	0.003 (0.011)
		60%	Total	0.072 (0.180)	0.008 (0.032)	0.153 (0.118)	0.023 (0.042)	0.259 (0.234)	0.047 (0.073)
			Low	0.051 (0.142)	0.004 (0.022)	0.073 (0.077)	0.010 (0.030)	0.192 (0.209)	0.037 (0.067)
			High	0.048 (0.137)	0.004 (0.022)	0.087 (0.085)	0.013 (0.033)	0.094 (0.147)	0.011 (0.033)

27

4 Real Data Analysis

To demonstrate the feasibility of the IW-GVEMM algorithm for detecting DIF in real data, we apply it to a dataset from the Patient-Reported Outcomes Measurement Information System (PROMIS) depression and anxiety subscales, which includes responses to 21 items of 5219 cancer patients. The two subscales measure depressive (10 items) and anxiety (11 items) symptoms respectively, and item content can be found in Table 11 of C. Wang et al. (2023). Teresi, Ocepek-Welikson, Kleinman, Ramirez, and Kim (2016a, 2016b) used this dataset to study DIF on race, a categorical variable with four levels, and we also focus on detecting race DIF here. The reference group is "Non-Hispanic White" (sample size $N_1 = 2239$), and the three focal groups are "Non-Hispanic Black" ($N_2 = 1077$), "Hispanic" ($N_3 = 1012$) and "Non-Hispanic Asians/Pacific Islanders" ($N_4 = 891$). All the 21 items have ordered categorical responses: "1 = Never", "2 = Rarely", "3 = Sometimes", "4 = Often" and "5 = Always", and the proportions that "Never" is chosen fall between 50% to 65% in most items. Therefore, similar to Bauer et al. (2020), we create dichotomous item responses by collapsing all categories except "Never" (i.e., "Rarely", "Sometimes", "Often", and "Always") to "Yes", indicating that the patient exhibits this symptom.

Teresi et al. (2016a, 2016b) applied two approaches to detect DIF items. Their first method is the Wald test, which is an iterative method using backward elimination. Initially all the items are assumed to have no DIF and hence work as anchor items. For each anchor item, an IRT model is fit with the constraint that all the current anchor items but this one have the same item parameters across all groups. Then a Wald test is conducted to determine whether the item parameters of this item have significant differences across groups. If so, this item is marked as having DIF and eliminated from the set of anchor items. This procedure is run repeatedly until the set of anchor items stabilizes. Their second method is the ordinal logistic regression, where for each item they regress the response on the group, the latent trait, and their interaction term. An item is marked as having DIF if the group effect or the interaction effect is significantly different from zero. For both methods, they did not seem

28

to model impact but instead assumed a common latent trait distribution for all groups.

Before discussing our empirical findings, we propose two possible ways for finding the best constant c in GIC for model selection. Figure 1 shows the relationship between ℓ_0, the number of non-zero DIF parameters of the model with the lowest GIC, against c. Figure 1 looks similar to scree plots in principal component analysis, and it suggests that the models chosen by BIC, GIC with $c = 0.7$, and GIC with $c = 0.9$ correspond to "elbows" of the plot. Or we may choose c by focusing on predictive accuracy as a model fit index. For group g and item j, we compute the predicted proportion of respondents that choose "Yes" according to the estimated impact and item parameters:

$$\mathbb{E}_{\boldsymbol{\theta} \sim \mathcal{N}_K(\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)} P(Y_j = 1 \mid \boldsymbol{\theta}) \approx \frac{1}{\lceil N/G \rceil} \sum_{i=1}^{\lceil N/G \rceil} P\left(Y_{ij} = 1 \mid \boldsymbol{\theta}_i\right) \triangleq \hat{p}_{gj},$$

where the expectation is approximated using Monte Carlo integration to accommodate high-dimensional settings, $\boldsymbol{\theta}_i$'s are independently sampled from $\mathcal{N}_K(\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)$ and $P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i)$ is defined in (1). Then, we define RMSE as the root mean square error between predicted and observed proportions:

$$\text{RMSE} = \sqrt{\frac{1}{GJ} \sum_{g=1}^{G} \sum_{j=1}^{J} \left[\hat{p}_{gj} - \frac{\sum_{i \in I_g} Y_{ij}}{N_g}\right]^2}.$$

The RMSEs corresponding to GIC with $c = 0.7$ and $c = 0.9$ are 0.021 and 0.022 respectively, suggesting that they have similar model fit. Note that the RMSEs reported here do not serve the purpose of cross-validation because they tend to be smaller for smaller c (i.e., more complex models). This also provides support for our default choice of $c = 1$ in the simulation study, but still we recommend trying different values of c and comparing their results.

Table 8 shows the DIF detection results. DIF is marked by "$\star$" for the Wald test and the logistic regression, and these results were obtained directly from Teresi et al. (2016a, 2016b). Since IW-GVEMM found no non-uniform DIF, non-zero estimates of DIF intercept
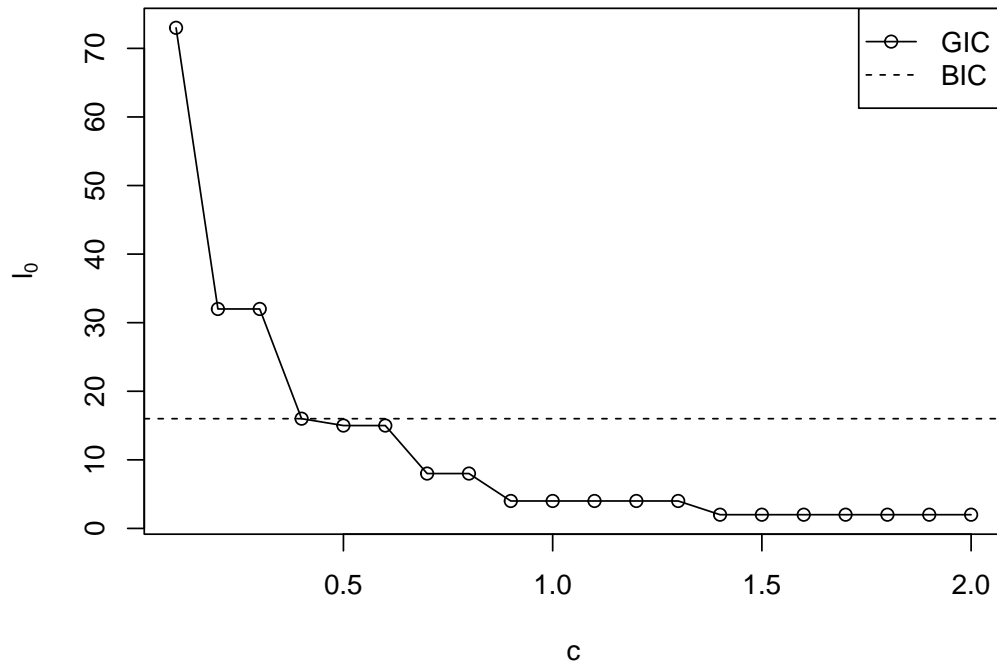
Figure 1: Relationship between Number of Non-Zero DIF Parameters and c of GIC in PROMIS Data

parameters ($\boldsymbol{\beta}$) are shown instead of "$\star$" in Table 8. DIF items detected by the three approaches do not agree with each other. In particular, regardless of the information criteria, IW-GVEMM finds much fewer DIF items for the "Non-Hispanic Asians/Pacific Islanders" group. This striking difference may be attributed to the fact that Teresi et al. (2016a, 2016b) did not consider impact, i.e., the differences among groups' population distributions were not considered. As shown in Table 9, the three focal groups all have higher mean anxiety and depression levels than the reference group, so ignoring this difference will inevitably bias DIF detection. Another possible reason is that ordinal responses are collapsed into binary to use our proposed method, whereas Teresi et al. (2016a, 2016b) used the original ordinal responses for DIF analysis. If DIF is absent between "Never" and "Yes" but is present among the four positive responses that are collapsed into "Yes", then only their approaches are able to detect DIF. As a result, extending our proposed methods to ordinal responses would be an important and useful future direction.

30

Table 8: DIF Detection Results of PROMIS Anxiety and Depression Scales

Item	Wald Test			Logistic Regression			IW-GVEMM (BIC)			IW-GVEMM (GIC, $c = 0.7$)			IW-GVEMM (GIC, $c = 0.9$)		
	Black	Hisp.	NHAPI	Black	Hisp.	NHAPI	Black	Hisp.	NHAPI	Black	Hisp.	NHAPI	Black	Hisp.	NHAPI
1	★	★	★	★	★	★	−0.489								
2			★	★	★	★									
3			★	★	★	★									
4					★	★		0.520							
5	★		★	★	★	★		0.765	1.123		0.729	1.152			0.940
6					★	★									
7		★	★		★	★		−0.411			−0.485				
8		★	★		★	★		1.215			1.154			1.098	
9	★	★	★	★	★	★	−0.576		−0.510	−0.428					
10	★	★	★	★	★	★		−0.228							
11			★		★	★		0.351							
12	★	★	★	★	★	★	−0.508	−0.538			−0.454				
13		★		★		★	−0.480								
14	★	★		★		★									
15					★	★									
16						★									
17	★					★			−0.416						
18			★	★	★	★		0.647			0.601			0.631	
19			★	★	★	★									
20			★	★		★		0.675			0.634			0.663	
21						★									

Reference group: Non-Hispanic White

For the Wald test and the logistic regression, "★" indicates DIF detected.

IW-GVEMM detected no non-uniform DIF, and non-empty cells display the estimates of DIF intercept parameters β.

31

Table 9: Estimated Mean and Covariance Matrix (Impact) of PROMIS Anxiety and Depression Scales Using IW-GVEMM

Information Criterion	Parameter	White	Black	Hisp.	NHAPI
BIC	$\bar{\mu}_{g1}$	0	0.054	0.216	0.084
	$\bar{\mu}_{g2}$	0	0.121	0.265	0.161
	$[\bar{\Sigma}_g]_{11}$	1	1.002	0.962	1.033
	$[\bar{\Sigma}_g]_{12}$	0.941	0.975	0.898	1.010
	$[\bar{\Sigma}_g]_{22}$	1	1.048	0.957	1.078
GIC, $c = 0.7$	$\bar{\mu}_{g1}$	0	0.048	0.222	0.079
	$\bar{\mu}_{g2}$	0	0.111	0.271	0.159
	$[\bar{\Sigma}_g]_{11}$	1	1.000	0.956	1.037
	$[\bar{\Sigma}_g]_{12}$	0.941	0.979	0.895	1.012
	$[\bar{\Sigma}_g]_{22}$	1	1.058	0.956	1.078
GIC, $c = 0.9$	$\bar{\mu}_{g1}$	0	0.044	0.227	0.081
	$\bar{\mu}_{g2}$	0	0.112	0.267	0.159
	$[\bar{\Sigma}_g]_{11}$	1	0.999	0.966	1.036
	$[\bar{\Sigma}_g]_{12}$	0.942	0.977	0.901	1.011
	$[\bar{\Sigma}_g]_{22}$	1	1.056	0.959	1.079

5 Discussion

This study demonstrates the feasibility of applying regularized IW-GVEMM to detect DIF within the regularized explanatory MIRT framework. Because all model parameters can be updated in closed forms in the M-step of the GVEM algorithm, it is computationally more efficient than the traditional EM algorithm. However, it may have unsatisfactory performance under non-uniform DIF conditions, which is likely due to the fact that GVEM generates a relatively large bias in discrimination parameters. Such an issue is common in variational estimation for various statistical models (Bishop & Nasrabadi, 2006). As a remedy, we further adopt the importance weighted variational technique (Ma et al., 2023), which gives a tighter variational lower bound of the marginal log-likelihood function. Simulation shows that importance sampling greatly improves the accuracy of estimation, although this additional step is slower due to gradient-based numerical optimization. Information criteria help determine the best tuning parameter λ for DIF detection. BIC often works well, but it

32

can lead to inflated false positive rates under some conditions. GIC provides more flexible control over the degree of penalization, but it involves another parameter c that may be hard to determine in practice.

This study has certain limitations that suggest potential directions for future research. First, following C. Wang et al. (2023), although our proposed approach allows and estimates impact, we let all the groups have the same latent trait distribution in the simulation study. Studying the performance of the proposed and other existing methods when impact exists and is large will provide useful guidance for users. Second, we proposed ways to find good values for c for GIC, but did not extensively study their performance using simulation because it is beyond the focus of this study. In practice, we may consider cross-validation for model comparison and selection, i.e., split the data into training and test data, fit the models to training data, and then compare their prediction accuracy over test data.

Similar to C. Wang et al. (2023), we only consider the Lasso or L_1 penalty for DIF detection. Due to the inherent bias introduced by Lasso penalty, one additional M-step without penalty is needed. A future direction is to use nonconcave penalties, such as a truncated L_1 penalty (TLP; Shen, Pan, & Zhu, 2012), whose idea is to replace (3) by $\ell_{\mathrm{TLP}}^*(\boldsymbol{\Delta}) = \log L(\boldsymbol{\Delta}) - \eta \left[J_\tau(\bar{\boldsymbol{\gamma}}) + J_\tau(\bar{\boldsymbol{\beta}}) \right]$, where $J_\tau(\delta) = \min(|\delta|, \tau)$ is an elementwise function and $\tau > 0$ is a tuning parameter. TLP corrects the bias of Lasso by combining adaptive shrinkage with thresholding, so there is no need to run an additional M-step to reduce bias. The optimal tuning parameter may be determined by BIC or GIC as well.

Properly identifying DIF and adjusting for DIF is essential for data harmonization because assuming strict item invariance across groups may be too strict and lead to inaccurate findings. Regularized explanatory MIRT is a flexible modeling framework that simultaneously handles multidimensional traits and potential DIF explained by multiple covariates. It obviates the tedious process of detecting DIF on each item and each covariate one at a time, which is often the case in traditional likelihood-ratio-based DIF detection, and the reliance on modification indices in confirmatory factor analysis. The IW-GVEMM algo-

33

rithm provides a computationally efficient alternative to the classic EM algorithm, and it can naturally handle high dimensional latent traits. Hence, it has a great potential to serve as a screening tool when analyzing integrated item response data. It is worth noting that we utilize dummy coding when dealing with multiple categorical covariates or one categorical covariate with multiple levels. This requires us to choose one group as a reference and all other groups become focal groups, and the proposed regularized DIF detection method identifies DIF items by comparing each focal group to the reference group. As a result, the proposed method may find different DIF items if a different reference group is chosen. This poses no problem when the goal is to adjust for non-invariance during data harmonization. However, if the goal is to detect DIF, then the selection of a designated reference group needs careful justification because we cannot directly compare two focal groups unless we run the algorithm again where one focal group becomes the new reference group. Hence, future research is needed to develop reference group agnostic DIF detection methods that will pinpoint items behaving differently across pairs of groups without designating a specific reference group.

References

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. doi: 10.1037/met0000077

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. doi: 10.1080/10705511.2019.1642754

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

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

Burda, Y., Grosse, R., & Salakhutdinov, R. (2016). *Importance weighted autoencoders.* doi: 10.48550/arXiv.1509.00519

Carrasco, M. A., Arias, R., & Figueroa, M. E. (2017). The multidimensional nature of HIV stigma: evidence from Mozambique. *African Journal of AIDS Research*, *16*(1), 11–18. doi: 10.2989/16085906.2016.1264983

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2014). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*(1), 18–36. doi: 10.1177/0146621613488643

Chen, Y., Li, C., Ouyang, J., & Xu, G. (2023). DIF statistical inference without knowing anchoring items. *Psychometrika*, *88*(4), 1097–1122. doi: 10.1007/s11336-023-09930-9

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*(S1), 52–85. doi: 10.1111/bmsp.12219

Cho, A. E., Xiao, J., Wang, C., & Xu, G. (2024). Regularized variational estimation for exploratory item factor analysis. *Psychometrika*, *89*(1), 347–375. doi: 10.1007/s11336-022-09874-6

Curran, P. J., & Hussong, A. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. doi: 10.1037/a0015914

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*(2), 121–136. doi: 10.1080/15248371003699969

Debelak, R., & Strobl, C. (2019). Investigating measurement invariance by means of parameter instability tests for 2PL and 3PL models. *Educational and Psychological Measurement*, *79*(2), 385–398. doi: 10.1177/0013164418777784

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, *90*(432), 1200–1224. doi: 10.1080/01621459.1995.10476626

Edelen, M. O., Stucky, B., & Chandra, A. (2015). Quantifying 'problematic' DIF within an IRT framework: application to a cancer stigma index. *Quality of Life Research*, *24*(1), 95–103. doi: 10.1007/s11136-013-0540-4

Falbel, D., & Luraschi, J. (2023). torch: Tensors and neural networks with 'gpu' acceleration [Computer software manual]. (https://torch.mlverse.org/docs, https://github.com/mlverse/torch)

Fayers, P. M. (2007). Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. *Quality of Life Research*, *16*(1), 187–194. doi: 10.1007/s11136-007-9197-1

Genz, A., & Keister, B. (1996). Fully symmetric interpolatory rules for multiple integrals over infinite regions with Gaussian weight. *Journal of Computational and Applied Mathematics*, *71*(2), 299–309. doi: 10.1016/0377-0427(95)00232-4

Hastie, T., Tibshirani, R., & Wainwright, M. (2015, May). Optimization methods. In *Statistical learning with sparsity: the lasso and generalizations* (pp. 111–154). CRC Press. doi: 10.1201/b18401-7

Heiss, F., & Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics*, *144*(1), 62–80. doi: 10.1016/j.jeconom.2007.12.004

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. doi: 10.48550/arXiv.1412.6980

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. doi: 10.1016/j.jmva.2009.04.008

Ma, C., Ouyang, J., Wang, C., & Xu, G. (2023). A note on improving variational estimation

for multidimensional item response theory. *Psychometrika*. doi: 10.1007/s11336-023-09939-0

Michel, P., Baumstarck, K., Lancon, C., Ghattas, B., Loundou, A., Auquier, P., & Boyer, L. (2018). Modernizing quality of life assessment: Development of a multidimensional computerized adaptive questionnaire for patients with schizophrenia. *Quality of Life Research*, *27*(4), 1041–1054. doi: 10.1007/s11136-017-1553-1

Nance, R., Delaney, J., Golin, C., Wechsberg, W., Cunningham, C., Altice, F., . . . Springer, S. (2017). Co-calibration of two self-reported measures of adherence to antiretroviral therapy. *AIDS Care*, *29*(4), 464–468. doi: 10.1080/09540121.2016.1263721

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research*, *206*(1), 647–662. doi: 10.1007/s10479-012-1181-7

Shen, X., Pan, W., & Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, *107*(497), 223–232. doi: 10.1080/01621459.2011.645783

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016a). Measurement equivalence of the patient reported outcomes measurement information system® (PROMIS®) anxiety short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(1), 183–219.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016b). Psychometric properties and performance of the patient reported outcomes measurement information system® (PROMIS®) depression short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(1), 141—181.

van de Geer, S. A. (2008, April). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, *36*(2). Retrieved from http://dx.doi.org/10.1214/009053607000000929 doi: 10.1214/009053607000000929

37

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. doi: 10.1080/00273171.2021.1985950

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116–136. doi: 10.1037/1082-989X.9.1.116

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91–120). Hogrefe & Huber Publishers.

Zhang, Y., Li, R., & Tsai, C.-L. (2012). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*(489), 312–323. doi: 10.1198/jasa.2009.tm08013

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*(90), 2541–2563. Retrieved from `http://jmlr.org/papers/v7/zhao06a.html`

Zheng, Y., Chang, C.-H., & Chang, H.-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research*, *22*(3), 491–499. doi: 10.1007/s11136-012-0179-6

38