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

October 19, 2023

Abstract

Data harmonization is an emerging approach to strategically combine 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, i.e., the regularization will enable 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. The GVEM algorithm is extended to a multi-group version with categorical covariates and L_1 penalty for re-MIRT. This note aims to provide empirical evidence to support feasible uses of GVEM for MIRT DIF detection, thereby providing a useful tool for integrative data analysis.

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 the simultaneous analysis of raw data pooled from different studies. It offers many advantages, including increased power due to larger sample size, enhanced external validity and generalizability due to greater heterogeneity in demographic and psycho-social characteristics, cost effectiveness due to 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, which are often required by existing methods, would likely be violated.

Procedures for evaluating and establishing measurement equivalence across samples are well developed from both factor analytic 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, a.k.a., measurement non-invariance). More recently, Bauer (2017) proposed a unified flexible model, namely, the moderated nonlinear factor analysis (MNLFA) that can handle different types of study-specific covariates simultaneously, such as gender (categorical) and age (continuous) and 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 implements 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 automatically separates true DIF by shrinking non-DIF parameters directly to 0.

The current regularized MNLFA is only restricted to unidimensional construct, and this note aims to expand the methodology to accommodate multidimensional construct. 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 counselling, 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, methodology also needs to advance on par with the demand. From statistical perspective, using a multivariate approach would also produce more accurate factor scores with reduced standard errors of measurement, due to borrowing information from correlated scales.

In this study, we will 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 adaptive Lasso, fitting re-MIRT on the integrated data will output a a commensurate scale for multidimensional constructs (e.g., depression, anxiety, alcohol use) that well accounts for study-specific idiosyncrasy resulting from diversity of study populations and use of different instruments. 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.

Wang, Zhu, and Xu (2021) first used the adaptive Lasso and Lasso penalty with the two-dimensional two-parameter logistic model and they found the two methods outperform the likelihood ratio test especially when the DIF proportion is high. However, the regularization method can be slow because it requires a full estimation at each candidate tuning parameter value. When a large grid of tuning parameters are considered, the entire algorithm may take hours to finish. In this note, we aim to extend their study by leveraging the recently developed Gaussian Variational Expectation-Maximization (GVEM) algorithm to speed up the computation. Because the GVEM algorithm relies on variational lower bound to approximate the true marginal likelihood, it is unknown whether the numerical approximation error may cause undesirable DIF detection.

The rest of the paper is organized as follows. We will first introduce the re-MIRT model for binary responses, followed by the regularized GVEM algorithm. Then we will present two small-scale simulation study to evaluate the performance and limitation of the algorithm in terms of detecting DIF. We will end the note with final discussions.

2 Method

2.1 Regularized Explanatory MIRT

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

$$P_{ij}(\boldsymbol{\theta}) \equiv P(Y_{ij} = 1 \mid \boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j, \boldsymbol{X}_i, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\left[(\boldsymbol{a}_j + \boldsymbol{\gamma}_j^\intercal \boldsymbol{X}_i)^\intercal \boldsymbol{\theta} - (b_j - \boldsymbol{\beta}_j^\intercal \boldsymbol{X}_i)\right]}}, \quad (1)$$

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). $\boldsymbol{\beta}_j \in \mathbb{R}^P$ 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 which denotes the interaction effects of $\boldsymbol{\theta}$ and grouping variables on item responses. As explained in Wang et al. (2021), in a confirmatory MIRT model, if $a_{jk} = 0$, then the kth column of $\boldsymbol{\gamma}_j$ will be zero by default. By way of this parameterization, $\boldsymbol{\gamma}_j = \mathbf{0}$ and $\boldsymbol{\beta}_j = \mathbf{0}$ if item j does not have DIF, while $\boldsymbol{\gamma}_j = \mathbf{0}$ if item j has uniform DIF. Similar to the multiple-group IRT approach, the distribution of $\boldsymbol{\theta}$ is allowed to differ across groups, i.e., $\boldsymbol{\theta}_i \sim N(\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)$ for all $i \in I_g$ where I_g is the set of all persons in group g.

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

$$\begin{aligned} L(\boldsymbol{\Delta}) &\equiv \int P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{X}) p(\boldsymbol{\theta} \mid \boldsymbol{\Delta}, \boldsymbol{X}) \mathrm{d}\boldsymbol{\theta} \qquad (2)\\ &= \prod_{g=1}^G \prod_{i \in I_g} \int_{\mathbb{R}^K} \left[\prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}, \boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j, \boldsymbol{X}_i) \right] \phi(\boldsymbol{\theta} \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g) \mathrm{d}\boldsymbol{\theta}, \qquad (3) \end{aligned}$$

where

$$P(Y_{ij} = y_{ij} \mid \boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j, \boldsymbol{X}_i, \boldsymbol{\theta}) = [P_{ij}(\boldsymbol{\theta})]^{y_{ij}} [1 - P_{ij}(\boldsymbol{\theta})]^{1-y_{ij}} \qquad (4)$$

is the conditional likelihood, $\phi(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal density, and $\bar{\boldsymbol{\mu}}_g$ and $\bar{\boldsymbol{\Sigma}}_g$ are the corresponding population mean and covariance matrix respectively.

In this study we focus on a simplified case where all persons within the same group share exactly the same covariates, that is, $\boldsymbol{X}_i \equiv \bar{\boldsymbol{X}}_g$ for all $i \in I_g$. Then, we can further restrict to 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,

$$\begin{bmatrix}\bar{\boldsymbol{X}}_1 & \bar{\boldsymbol{X}}_2 & \cdots & \bar{\boldsymbol{X}}_G\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \boldsymbol{I}_{G-1}\end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{5}$$

For covariate groups $g = 2, \ldots, G$ let $\bar{\boldsymbol{\gamma}}_{gj} = \boldsymbol{\gamma}_j^\intercal \bar{\boldsymbol{X}}_g$ and $\bar{\beta}_{gj} = \boldsymbol{\beta}_j^\intercal \bar{\boldsymbol{X}}_g$ denote the DIF slope and intercept parameters of group g against group 1 on item j respectively. As the reference group, fix $\bar{\boldsymbol{\gamma}}_{1j} = \mathbf{0}$ and $\bar{\beta}_{1j} = 0$ for $j = 1, \ldots, J$. We estimate $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ rather than $\boldsymbol{\gamma}_j$ and $\boldsymbol{\beta}_j$ in order to avoid dealing with $\boldsymbol{X}$ and $\bar{\boldsymbol{X}}$. To simplify notations, we let $\boldsymbol{\gamma}_{ij} \equiv \boldsymbol{\gamma}_j^\intercal \boldsymbol{X}_i = \bar{\boldsymbol{\gamma}}_{gj}$ and $\beta_{ij} \equiv \boldsymbol{\beta}_j^\intercal \boldsymbol{X}_i = \bar{\beta}_{gj}$ for all $g = 1, \ldots, G$ and $i \in I_g$.

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

$$\ell^*(\bar{\boldsymbol{\Delta}}) = \log \bar{L}(\bar{\boldsymbol{\Delta}}) - \lambda \left(\|\bar{\boldsymbol{\gamma}}\|_1 + \|\bar{\boldsymbol{\beta}}\|_1 \right), \tag{6}$$

where $\bar{\boldsymbol{\Delta}} = \{\bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g\}_{g=1}^G \bigcup \{\boldsymbol{a}_j, b_j\}_{j=1}^J \bigcup \{\bar{\boldsymbol{\gamma}}_{gj}, \bar{\beta}_{gj}\}_{g=1}^G{}_{j=1}^J$,

$$\begin{aligned}\bar{L}(\bar{\boldsymbol{\Delta}}) &\equiv \int_{\mathbb{R}^{NK}} P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\theta}, \bar{\boldsymbol{\Delta}}, \boldsymbol{X}) p(\boldsymbol{\theta} \mid \bar{\boldsymbol{\Delta}}, \boldsymbol{X}) \mathrm{d}\boldsymbol{\theta} \\ &= \prod_{g=1}^G \prod_{i \in I_g} \int_{\mathbb{R}^K} \left[\prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \boldsymbol{\theta}, \boldsymbol{a}_j, b_j, \bar{\boldsymbol{\gamma}}_{gj}, \bar{\beta}_{gj}) \right] \phi(\boldsymbol{\theta} \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g) \mathrm{d}\boldsymbol{\theta},\end{aligned} \tag{7}$$

$$\|\bar{\boldsymbol{\gamma}}\|_1 = \sum_{j=1}^{J}\sum_{g=1}^{G}\sum_{k=1}^{K} |\bar{\gamma}_{jgk}|, \quad \|\bar{\boldsymbol{\beta}}\|_1 = \sum_{j=1}^{J}\sum_{g=1}^{G} |\bar{\beta}_{gj}|, \tag{8}$$

and $\lambda > 0$ is the regularization parameter that controls sparsity (Wang et al., 2021).

2.2 Regularized Multi-Group GVEM

The Gaussian variational EM algorithm for the re-MIRT model in (1) differs from the original GVEM algorithm in Cho, Wang, Zhang, and Xu (2021) in that it generalizes to the multiple-group scenario by including explanatory variables $\boldsymbol{X}$. The variational lower bound of the log-likelihood $\log \bar{L}(\bar{\Delta})$ in the E-step (i.e., Section 3.1.2 in Cho et al. (2021)) is

$$\begin{aligned} Q(\bar{\boldsymbol{\Delta}}) = \sum_{i=1}^{N}\sum_{j=1}^{J}\Bigg\{ &\log \frac{e^{\xi_{ij}}}{1+e^{\xi_{ij}}} + \left(\frac{1}{2} - Y_{ij}\right)\left[(b_j - \beta_{ij}) - (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^{\mathsf{T}}\boldsymbol{\mu}_i\right] - \frac{1}{2}\xi_{ij} \\ &- \eta(\xi_{ij})\Big[(b_j - \beta_{ij})^2 - 2(b_j - \beta_{ij})(\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^{\mathsf{T}}\boldsymbol{\mu}_i \\ &+ (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^{\mathsf{T}}\left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^{\mathsf{T}}\right)(\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij}) - \xi_{ij}^2\Big]\Bigg\} \\ &- \frac{1}{2}\sum_{g=1}^{G}\left[N_g \log\left|\bar{\boldsymbol{\Sigma}}_g\right| + \sum_{i\in I_g} \operatorname{tr}\left\{\bar{\boldsymbol{\Sigma}}_g^{-1}\left[\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)^{\mathsf{T}}\right]\right\}\right], \end{aligned} \tag{9}$$

where $N_g = |I_g|$ is the size of group g, ξ_{ij} is the variational parameter,

$$\eta(\xi) = \begin{cases} \dfrac{1}{2\xi}\left(\dfrac{1}{1+e^{-\xi}} - \dfrac{1}{2}\right), & \xi \neq 0, \\ 0.125, & \xi = 0, \end{cases} \tag{10}$$

and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and the covariance matrix of the Gaussian variational density for person i, i.e., $q_i(\boldsymbol{\theta}_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

In the M-step, each model parameter is updated separately by a closed form formula. The details are given below.

- The update formulas for variational parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and ξ_{ij} are natural extensions of those in Cho et al. (2021):

$$\xi_{ij} \leftarrow \left[(b_j - \beta_{ij})^2 - 2(b_j - \beta_{ij})(\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\intercal \boldsymbol{\mu}_i + (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\intercal \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\intercal\right)(\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})\right]^{\frac{1}{2}}, \tag{11}$$

$$\boldsymbol{\Sigma}_i^{-1} \leftarrow \bar{\boldsymbol{\Sigma}}_g^{-1} + 2\sum_{j=1}^{J} \eta(\xi_{ij})(\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})(\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\intercal, \tag{12}$$

$$\boldsymbol{\mu}_i \leftarrow \boldsymbol{\Sigma}_i \left\{ \sum_{j=1}^{J} \left[\left(Y_{ij} - \frac{1}{2}\right) + 2\eta(\xi_{ij})(b_j - \beta_{ij})\right](\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij}) + \bar{\boldsymbol{\Sigma}}_g^{-1} \bar{\boldsymbol{\mu}}_g \right\}, \tag{13}$$

  where $i \in I_g$.

- Latent trait distribution parameters $\bar{\boldsymbol{\mu}}_g$ and $\bar{\boldsymbol{\Sigma}}_g$ are updated by

$$\bar{\boldsymbol{\mu}}_g \leftarrow \frac{1}{N_g} \sum_{i \in I_g} \boldsymbol{\mu}_i \tag{14}$$

  and

$$\bar{\boldsymbol{\Sigma}}_g \leftarrow \frac{1}{N_g} \left[\sum_{i \in I_g} \boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)^\intercal \right]. \tag{15}$$

- Item discrimination and difficulty parameters $\boldsymbol{a}_j$ and b_j are updated by setting their corresponding first derivatives

$$\frac{\partial Q}{\partial \boldsymbol{a}_j} = \sum_{i=1}^{N} \left(Y_{ij} - \frac{1}{2} \right) \boldsymbol{\mu}_i + 2\eta(\xi_{ij}) \left[(b_j - \beta_{ij}) \boldsymbol{\mu}_i - \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\mathsf{T} \right) (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij}) \right] \quad (16)$$

and

$$\frac{\partial Q}{\partial b_j} = \sum_{i=1}^{N} \left(\frac{1}{2} - Y_{ij} \right) - 2\eta(\xi_{ij}) \left[(b_j - \beta_{ij}) - (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\mathsf{T} \boldsymbol{\mu}_i \right] \quad (17)$$

to zero, which results in closed form solutions

$$\begin{aligned} \boldsymbol{a}_j \leftarrow \left[\sum_{i=1}^{N} 2\eta(\xi_{ij}) \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\mathsf{T} \right) \right]^{-1} \Bigg\{ & \sum_{i=1}^{N} \left(Y_{ij} - \frac{1}{2} \right) \boldsymbol{\mu}_i \\ & + 2\eta(\xi_{ij}) \Big[(b_j - \beta_{ij}) \boldsymbol{\mu}_i - \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\mathsf{T} \right) \boldsymbol{\gamma}_{ij} \Big] \Bigg\} \end{aligned} \quad (18)$$

and

$$b_j \leftarrow \left[\sum_{i=1}^{N} 2\eta(\xi_{ij}) \right]^{-1} \left\{ \sum_{i=1}^{N} \left(\frac{1}{2} - Y_{ij} \right) + 2\eta(\xi_{ij}) \left[\beta_{ij} + (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^\mathsf{T} \boldsymbol{\mu}_i \right] \right\}. \quad (19)$$

- Since DIF parameters $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ are L^1 penalized in Q_j, we adopt a quadratic approximation approach similar to Wang et al. (2021): the closed form update rule of parameter φ with respect to objective function f is

$$\varphi \leftarrow -\frac{S_\lambda \left(\dfrac{\partial Q}{\partial \varphi} - \varphi \dfrac{\partial^2 Q}{\partial \varphi^2} \right)}{\dfrac{\partial^2 Q}{\partial \varphi^2}}, \quad (20)$$

where $S_\lambda(z) = \mathrm{sign}(z) \max(|z| - \lambda, 0)$ is a soft thresholding operator (Donoho

& Johnstone, 1995). Since

$$\frac{\partial Q}{\partial \bar{\boldsymbol{\gamma}}_{gj}} = \sum_{i \in I_g} \left(Y_{ij} - \frac{1}{2}\right) \boldsymbol{\mu}_i + 2\eta(\xi_{ij}) \left[(b_j - \bar{\beta}_{gj})\boldsymbol{\mu}_i - \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathsf{T}}\right)(\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj})\right], \tag{21}$$

$$\frac{\partial^2 Q}{\partial \bar{\boldsymbol{\gamma}}_{gj} \partial \bar{\boldsymbol{\gamma}}_{gj}^{\mathsf{T}}} = -2 \sum_{i \in I_g} \eta(\xi_{ij}) \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathsf{T}}\right), \tag{22}$$

$$\frac{\partial Q}{\partial \bar{\beta}_{gj}} = \sum_{i \in I_g} \left(Y_{ij} - \frac{1}{2}\right) + 2\eta(\xi_{ij}) \left[(b_j - \bar{\beta}_{gj}) - (\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj})^{\mathsf{T}} \boldsymbol{\mu}_i\right], \tag{23}$$

$$\frac{\partial^2 Q}{\partial \bar{\beta}_{gj}^2} = -2 \sum_{i \in I_g} \eta(\xi_{ij}), \tag{24}$$

we update $\bar{\boldsymbol{\gamma}}_{gj}$ and $\bar{\beta}_{gj}$ by

$$\bar{\gamma}_{gjk} \leftarrow \frac{\left[S_\lambda\left(\sum_{i \in I_g} \left(Y_{ij} - \frac{1}{2}\right) \boldsymbol{\mu}_i + 2\eta(\xi_{ij}) \left[(b_j - \bar{\beta}_{gj})\boldsymbol{\mu}_i - \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathsf{T}}\right) \boldsymbol{a}_j\right]\right)\right]_k}{\left[\sum_{i \in I_g} 2\eta(\xi_{ij}) \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathsf{T}}\right)\right]_{kk}} \tag{25}$$

and

$$\bar{\beta}_{gj} \leftarrow \frac{S_\lambda \left(\sum_{i \in I_g} \left(Y_{ij} - \frac{1}{2}\right) + 2\eta(\xi_{ij}) \left[b_j - (\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj})^{\mathsf{T}} \boldsymbol{\mu}_i\right]\right)}{\sum_{i \in I_g} 2\eta(\xi_{ij})} \tag{26}$$

for $g = 2, \ldots, G$ while fixing $\bar{\boldsymbol{\gamma}}_{1j} = \mathbf{0}$ and $\bar{\beta}_{1j} = 0$.

- Finally, for model identification, we fix $\bar{\boldsymbol{\mu}}_1 = \mathbf{0}$ and $\operatorname{diag}(\bar{\boldsymbol{\Sigma}}_1) = \boldsymbol{I}_K$ by

$$\begin{aligned}
b_j &\leftarrow b_j - \boldsymbol{a}_j^\mathsf{T} \bar{\boldsymbol{\mu}}_1, && (27)\\
\bar{\beta}_{gj} &\leftarrow \bar{\beta}_{gj} + \bar{\boldsymbol{\gamma}}_{gj}^\mathsf{T} \bar{\boldsymbol{\mu}}_1, && (28)\\
\boldsymbol{\mu}_i &\leftarrow \boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_1, && (29)\\
\bar{\boldsymbol{\mu}}_g &\leftarrow \bar{\boldsymbol{\mu}}_g - \bar{\boldsymbol{\mu}}_1, && (30)\\
\boldsymbol{a}_j &\leftarrow \bar{\boldsymbol{\Lambda}}_1 \boldsymbol{a}_j, && (31)\\
\bar{\boldsymbol{\gamma}}_{gj} &\leftarrow \bar{\boldsymbol{\Lambda}}_1 \bar{\boldsymbol{\gamma}}_{gj}, && (32)\\
\boldsymbol{\Sigma}_i &\leftarrow \bar{\boldsymbol{\Lambda}}_1^{-\mathsf{T}} \boldsymbol{\Sigma}_i \bar{\boldsymbol{\Lambda}}_1^{-1}, && (33)\\
\bar{\boldsymbol{\Sigma}}_g &\leftarrow \bar{\boldsymbol{\Lambda}}_1^{-\mathsf{T}} \bar{\boldsymbol{\Sigma}}_g \bar{\boldsymbol{\Lambda}}_1^{-1}, && (34)
\end{aligned}$$

  for $j = 1, \ldots, J$ and $g = 1, \ldots, G$, where $\bar{\boldsymbol{\Lambda}}_1 = \sqrt{\operatorname{diag}(\bar{\boldsymbol{\Sigma}}_1)}$ is a diagonal matrix whose diagonal elements are the square roots of those of $\bar{\boldsymbol{\Sigma}}_1$.

Note that even with the transformations from (27) to (34), the model is still not identified under (5) because any group $g' \in \{2, \ldots, G\}$ can be rescaled without affecting other groups. Let $\odot$ denote the elementwise multiplication operator. Then for any $\boldsymbol{u} \in \mathbb{R}_+^K$ and $\boldsymbol{v} \in \mathbb{R}^K$ we have

$$(\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj})^\mathsf{T} \boldsymbol{\theta}_i - (b_j - \bar{\boldsymbol{\beta}}_{gj}) = (\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}'_{gj})^\mathsf{T} \boldsymbol{\theta}'_i - (b_j - \bar{\boldsymbol{\beta}}'_{gj}) \qquad (35)$$

for all $g = 1, \dots, G$ and $i \in I_g$, where

$$\boldsymbol{\theta}_i' = \begin{cases} \boldsymbol{\theta}_i, & g \neq g', \\ \boldsymbol{u} \odot \boldsymbol{\theta}_i + \boldsymbol{v}, & g = g', \end{cases} \tag{36}$$

$$\bar{\boldsymbol{\gamma}}_{gj}' = \begin{cases} \bar{\boldsymbol{\gamma}}_{gj}, & g \neq g', \\ (\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj}) \odot \dfrac{1}{\boldsymbol{u}} - \boldsymbol{a}_j, & g = g', \end{cases} \tag{37}$$

and

$$\bar{\boldsymbol{\beta}}_{gj}' = \begin{cases} \bar{\boldsymbol{\beta}}_{gj}, & g \neq g', \\ \bar{\boldsymbol{\beta}}_{gj} + (\boldsymbol{a}_j + \bar{\boldsymbol{\gamma}}_{gj}')^\intercal \boldsymbol{v}, & g = g'. \end{cases} \tag{38}$$

For example, under uniform DIF, we cannot statistically distinguish between the case where items are easier to group 2 than to group 1 (i.e., $\beta_{2j} > 0$) and the case where the mean latent traits of group 2 are higher than those of group 1 (i.e., $\boldsymbol{\mu}_2 \in \mathbb{R}_+^K$ while $\boldsymbol{\mu}_1 = \mathbf{0}$). We detect DIF in the former but not the latter case. Therefore, in the simulation study below we require that each group's DIF items are not in the same direction so that differences in item parameters will not be absorbed into differences in distributions of latent traits.

After each M-step, it is necessary to re-estimate all non-zero model parameters and use them as initial values for the next E-step. One additional M-step can be carried out without a penalty (i.e., $\lambda = 0$) while fixing current zero elements in $\bar{\boldsymbol{\gamma}}$'s and $\bar{\boldsymbol{\beta}}$'s at zero. That is, each E-step is followed by two M-steps, the first one with a penalty, and the second one without a penalty. This algorithm is therefore termed as the GVEMM algorithm.

Compared to the Lasso EMM method proposed by (Wang et al., 2021), one advantage of this regularized GVEMM method is that it better handles higher dimensionality, such as multiple categorical covariates. In our simulation study, we only consider three groups based on one categorical variable in order to be consistent with the prior study. However, consider the case where we have multiple categorical variables, say sex and race, resulting in, for instance, 8 different groups by interactions. Even if the latent trait is unidimensional and only uniform DIF is considered, each item j will have a 7-by-1 vector $\boldsymbol{\beta}_j$, a discrimination parameter a_j, and an intercept b_j. Then the M-step of the Lasso EMM method will involve inverting a 9-by-9 Hessian matrix, which can be numerically unstable and lead to unreliable estimates when the Hessian matrix is nearly singular. In contrast, all model parameters have closed-form update formulas in the regularized GVEMM algorithm, so their estimates are more viable.

2.3 Bias Reduction via Importance Sampling

The proposed GVEMM algorithm is know to have large bias in discrimination parameters when latent traits of different dimensions are highly correlated and the sample size is not large (Cho et al., 2021), so it is not good at detecting non-uniform DIF. To reduce bias in model parameter estimates, we apply a similar approach to IW-GVEM citation by performing additional importance sampling steps after GVEMM converges. The main idea is to sample from the estimated posterior distributions of person parameters and approximate a tighter variational lower bound of the marginal log-likelihood function $\log \bar{L}(\bar{\boldsymbol{\Delta}})$ in (7):

$$\begin{aligned}
\log \bar{L}(\bar{\boldsymbol{\Delta}}) &= \sum_{g=1}^{G}\sum_{i\in I_g} \log \int_{\mathbb{R}^K} P(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\theta}_i, \bar{\boldsymbol{\Delta}})\phi(\boldsymbol{\theta}_i \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)\mathrm{d}\boldsymbol{\theta}_i \\
&= \sum_{g=1}^{G}\sum_{i\in I_g} \log \int_{\mathbb{R}^K} \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\theta}_i, \bar{\boldsymbol{\Delta}})\phi(\boldsymbol{\theta}_i \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)}{\phi(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}\phi(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\mathrm{d}\boldsymbol{\theta}_i \\
&\geq \sum_{g=1}^{G}\sum_{i\in I_g} \int_{\mathbb{R}^{MK}} \left[\log \frac{1}{M}\sum_{m=1}^{M} \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\theta}_i^{(m)}, \bar{\boldsymbol{\Delta}})\phi(\boldsymbol{\theta}_i^{(m)} \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)}{\phi(\boldsymbol{\theta}_i^{(m)} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}\phi(\boldsymbol{\theta}_i^{(m)} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right] \mathrm{d}\boldsymbol{\theta}_i^{(1)} \cdots \mathrm{d}\boldsymbol{\theta}_i^{(} \\
&\approx \sum_{g=1}^{G}\sum_{i\in I_g} \frac{1}{S}\sum_{s=1}^{S} \left[\log \frac{1}{M}\sum_{m=1}^{M} \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\theta}_i^{(s,m)}, \bar{\boldsymbol{\Delta}})\phi(\boldsymbol{\theta}_i^{(s,m)} \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)}{\phi(\boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}\right],
\end{aligned} \tag{39}$$

where $\boldsymbol{\theta}_i^{(s,m)} \sim q_i(\boldsymbol{\theta}_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is sampled from the posterior distribution of $\boldsymbol{\theta}_i$. Then, the objective function with L_1 penalty becomes

$$Q^*(\bar{\boldsymbol{\Delta}}) = \sum_{g=1}^{G}\sum_{i\in I_g} \frac{1}{S}\sum_{s=1}^{S} \left[\log \frac{1}{M}\sum_{m=1}^{M} \frac{P(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\theta}_i^{(s,m)}, \bar{\boldsymbol{\Delta}})\phi(\boldsymbol{\theta}_i^{(s,m)} \mid \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)}{\phi(\boldsymbol{\theta}_i^{(s,m)} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}\right] - \lambda\left(\|\bar{\boldsymbol{\gamma}}\|_1 + \|\bar{\boldsymbol{\beta}}\|_1\right), \tag{40}$$

which is minimized using gradient-based algorithms due to its complexity. To ensure the positive definiteness of $\bar{\boldsymbol{\Sigma}}_g$, we consider the Cholesky decomposition $\bar{\boldsymbol{\Sigma}}_g = \bar{\boldsymbol{L}}_g\bar{\boldsymbol{L}}_g^{\mathsf{T}}$ and minimize (40) with respect to $\bar{\boldsymbol{L}}_g$ instead of $\bar{\boldsymbol{\Sigma}}_g$. Further, we fix $\bar{\boldsymbol{\mu}}_1 = \mathbf{0}$ and

$\operatorname{diag}(\bar{\boldsymbol{\Sigma}}_1) = \operatorname{diag}(\bar{\boldsymbol{L}}_1\bar{\boldsymbol{L}}_1^\intercal) = \boldsymbol{I}_K$ by applying the transformation

$$\bar{\boldsymbol{L}}_1 = \begin{bmatrix} 1 & & & & \\ z_{21} & \sqrt{1-z_{21}^2} & & & \\ z_{31} & z_{32}\sqrt{1-z_{31}^2} & \sqrt{(1-z_{31})^2(1-z_{32})^2} & & \\ \vdots & \vdots & \vdots & \ddots & \\ z_{K1} & z_{K2}\sqrt{1-z_{K1}^2} & z_{K3}\sqrt{(1-z_{K1})^2(1-z_{K2})^2} & \cdots & \sqrt{\prod_{k=1}^{K}(1-z_{Kk})^2} \end{bmatrix}, \tag{41}$$

where $z_{ij} = \tanh y_{ij} \in (-1,1)$ and $y_{ij} \in \mathbb{R}$ for $i = 2,\ldots,K$ and $j = 1,\ldots,i-1$ (Lewandowski, Kurowicka, & Joe, 2009).

We apply the Adam optimization algorithm by Kingma and Ba (2014), a popular optimizer in deep learning, to minimize $Q^{(S,M)}(\bar{\boldsymbol{\Delta}})$ with respect to $\boldsymbol{a}, \boldsymbol{b}, \bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$.

2.4 Information Criteria for Tuning Parameter Selection

Information criteria are used in this study to find the best tuning parameter λ. Since the log-likelihood $\log \bar{L}(\bar{\boldsymbol{\Delta}})$ in (7) is difficult to compute due to numerical integration while its variational lower bound $Q(\bar{\boldsymbol{\Delta}})$ in (9) is a by-product of our proposed algorithm that is easier to compute, we modify the generalized information criterion (GIC; Zhang, Li, & Tsai, 2010)

$$\begin{aligned} \text{GIC} &= -2\log \bar{L}(\bar{\boldsymbol{\Delta}}) + k_N\left(\|\bar{\boldsymbol{\gamma}}\|_0 + \|\bar{\boldsymbol{\beta}}\|_0\right) \\ &= -2\sum_{g=1}^{G}\sum_{i\in I_g}\log\int\prod_{j=1}^{J}P(Y_{ij}=y_{ij}\mid \boldsymbol{a}_j, b_j, \bar{\boldsymbol{\gamma}}_{gj}, \bar{\beta}_{gj}, \boldsymbol{X}_i, \boldsymbol{\theta})\phi(\boldsymbol{\theta}\mid\bar{\boldsymbol{\mu}}_g,\bar{\boldsymbol{\Sigma}}_g)\mathrm{d}\boldsymbol{\theta} + k_N\left(\|\bar{\boldsymbol{\gamma}}\|_0 + \|\bar{\boldsymbol{\beta}}\|_0\right) \end{aligned} \tag{42}$$

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

$$
\begin{aligned}
\mathrm{GIC} &= Q(\bar{\boldsymbol{\Delta}}) + k_N \left(\|\bar{\boldsymbol{\gamma}}\|_0 + \|\bar{\boldsymbol{\beta}}\|_0 \right) \\
&= -2 \sum_{i=1}^{N} \sum_{j=1}^{J} \left\{ \log \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}} + \left(\frac{1}{2} - Y_{ij} \right) \left[(b_j - \beta_{ij}) - (\boldsymbol{a}_j + \boldsymbol{\gamma}_{ij})^{\mathsf{T}} \boldsymbol{\mu}_i \right] - \frac{1}{2} \xi_{ij} \right\} \\
&\quad + \sum_{g=1}^{G} \left[N_g \log \left| \bar{\boldsymbol{\Sigma}}_g \right| + \sum_{i \in I_g} \operatorname{tr} \left\{ \bar{\boldsymbol{\Sigma}}_g^{-1} \left[\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g) (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}_g)^{\mathsf{T}} \right] \right\} \right] \\
&\quad + k_N \left(\|\bar{\boldsymbol{\gamma}}\|_0 + \|\bar{\boldsymbol{\beta}}\|_0 \right),
\end{aligned}
\tag{43}
$$

where ξ_{ij} is defined in (11),

$$
\|\bar{\boldsymbol{\gamma}}\|_0 = \sum_{j=1}^{J} \sum_{g=1}^{G} \sum_{k=1}^{K} |\mathbb{1}\{\bar{\gamma}_{jgk} \neq 0\}, \quad \|\bar{\boldsymbol{\beta}}\|_0 = \sum_{j=1}^{J} \sum_{g=1}^{G} \mathbb{1}\{\bar{\beta}_{gj} \neq 0\}, \tag{44}
$$

and k_N is an increasing function of N. In particular, GIC becomes BIC by taking $k_N = \log N$. We also use $k_N = c \log N \log\log N$ where $c > 0$ is a constant that controls the magnitude of penalty, i.e., larger c means higher penalty and shrinks more parameters toward zero. Our simulation study shows that $Q(\bar{\boldsymbol{\Delta}})$ works as a good proxy of $\log L(\bar{\boldsymbol{\Delta}})$.

3 Simulation

Two simulation studies were performed to examine the GVEMM algorithm for DIF detection in a two-parameter re-MIRT model. Study I focuses on uniform DIF detection whereas study II focuses on non-uniform DIF detection. The results are compared to the Lasso EMM algorithm in Wang et al. (2021). Test length was fixed at

20. Discrimination parameters were generated from Uniform(1.5, 2.5) and boundary parameters were generated from N(0,1). The generated items parameters are given in table 1, which is the same as Table 1 in Wang et al. (2021). The latent variables of three groups, one reference group and two focal groups, were generated from bivariate normal distributions with mean vector $\mathbf{0}$ and variances 1. The correlation between two trait dimensions was 0.85. Similar to Wang et al. (2021), no impact was generated for the two focal groups. Two factors, sample size (1500 and 3000) and DIF proportion (20% and 60%), were manipulated.

Table 1: 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

3.1 Simulation I

Three groups were considered with one reference group and two focal groups. The first focal group had small DIF magnitude ($\boldsymbol{\beta}_1 = 0.5$) and the second focal group had large DIF ($\boldsymbol{\beta}_1 = 1$). In the 20% DIF condition, four items 4, 5, 12, 13 had DIF. In the 60% DIF condition, 12 items 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17 had DIF. We considered both BIC and GIC as the criterion for selecting the best tuning parameter for the GVEMM algorithm, and for GIC, we set $c = 5$. Table 2 and table 3 show the results of type I error and power of DIF detection. Values in the parenthesis are standard deviations computed across 25 replications. As shown, the two versions of GVEMM appears to perform similarly to the original EMM algorithm, although BIC tends to penalize more severely than GIC, especially when sample size is large and DIF proportion is high, yielding slightly inflated Type I error with power of almost 1. Besides, all three methods produce higher power with larger sample size, and larger DIF size is easier to detect, all unsurprisingly.

Table 2: Study I Type I error of detecting uniform DIF

Corr	N	DIF%	Group	EMM	GVEMM BIC	GVEMM GIC
0.85	1500	20%	Omnibus DIF	0.021 (0.005)	0.035 (0.015)	0 (0)
			Low DIF	0.013 (0.004)	0.025 (0.012)	0 (0)
			High DIF	0.011 (0.003)	0.018 (0.008)	0 (0)
		60%	Omnibus DIF	0.035 (0.011)	0.035 (0.026)	0.013 (0.005)
			Low DIF	0.025 (0.009)	0.005 (0.005)	0.013 (0.005)
			High DIF	0.013 (0.005)	0.030 (0.026)	0 (0)
	3000	20%	Omnibus DIF	0.026 (0.006)	0.078 (0.024)	0.018 (0.005)
			Low DIF	0.021 (0.005)	0.043 (0.013)	0.015 (0.005)
			High DIF	0.006 (0.003)	0.055 (0.020)	0.008 (0.004)
		60%	Omnibus DIF	0.060 (0.015)	0.175 (0.044)	0.055 (0.053)
			Low DIF	0.233 (0.015)	0.130 (0.032)	0.050 (0.040)
			High DIF	0.008 (0.004)	0.085 (0.030)	0.020 (0.026)

3.2 Simulation II

A second simulation study focused on detecting non-uniform DIF, with DIF effects on both intercept and slope are simulated. The first focal group with small DIF magnitude had $\boldsymbol{\beta}_1 = 0.25$ and $\boldsymbol{\gamma}_1 = -0.4$ on items with DIF, and the second focal group with large DIF had $\boldsymbol{\beta}_1 = 0.6$ and $\boldsymbol{\gamma}_1 = -0.6$. Again, as in Study I, in the

Table 3: Study I Power of detecting uniform DIF

Corr	N	DIF%	Group	EMM	GVEMM BIC	GVEMM GIC
0.85	1500	20%	Omnibus DIF	0.96 (0.017)	0.97 (0.022)	0.95 (0.025)
			Low DIF	0.55 (0.043)	0.33 (0.065)	0.21 (0.048)
			High DIF	0.96 (0.017)	0.97 (0.022)	0.95 (0.025)
		60%	Omnibus DIF	0.885 (0.019)	0.72 (0.074)	0.653 (0.070)
			Low DIF	0.208 (0.024)	0.233 (0.026)	0.063 (0.017)
			High DIF	0.885 (0.019)	0.72 (0.074)	0.653 (0.070)
	3000	20%	Omnibus DIF	1 (0)	1 (0)	1 (0)
			Low DIF	0.84 (0.029)	0.95 (0.025)	0.91 (0.029)
			High DIF	1 (0)	1 (0)	1 (0)
		60%	Omnibus DIF	0.998 (0.002)	1 (0)	1 (0)
			Low DIF	0.632 (0.032)	0.907 (0.038)	0.883 (0.038)
			High DIF	0.998 (0.002)	1 (0)	1 (0)

20% DIF condition, four items (i.e., 4, 5, 12, 13) were simulated to have DIF. In the 60% DIF condition, 12 items (i.e., 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17) were simulated to have DIF. All three methods were used, and they are EMM, GVEMM with BIC, as well as GVEMM with GIC. The penalty is smaller than the penalty in GIC in simulation I as we set $c = 3$. The results were presented in Tables **??** and 5. As shown, using BIC for GVEMM tended to generate extremely low Type I error and therefore lowest power across all conditions. Using GIC for GVEMM improves the power but the Type I error for large sample size and/or high DIF proportion conditions are inflated. Overall, the GVEMM method does not perform as well as the orginal EMM method, and this is partly due to the fact that GVEM algorithm generates larger bias in discrimination parameters than EM.

Table 4: Study II Type I error of detecting non-uniform DIF

Corr	N	DIF%	Group	EMM	GVEMM BIC	GVEMM GIC
0.85	1500	20%	Omnibus DIF	0.021 (0.005)	0.030 (0.009)	0.145 (0.023)
			Low DIF	0.013 (0.004)	0.015 (0.007)	0.073 (0.015)
			High DIF	0.011 (0.003)	0.015 (0.008)	0.095 (0.022)
		60%	Omnibus DIF	0.035 (0.011)	0.025 (0.021)	0.17 (0.035)
			Low DIF	0.025 (0.009)	0.005 (0.005)	0.10 (0.023)
			High DIF	0.013 (0.005)	0.020 (0.020)	0.09 (0.027)
	3000	20%	Omnibus DIF	0.026 (0.006)	0.010 (0.007)	0.270 (0.030)
			Low DIF	0.021 (0.005)	0.008 (0.004)	0.173 (0.026)
			High DIF	0.006 (0.003)	0.003 (0.002)	0.160 (0.017)
		60%	Omnibus DIF	0.060 (0.015)	0.025 (0.012)	0.295 (0.047)
			Low DIF	0.233 (0.015)	0.020 (0.009)	0.180 (0.031)
			High DIF	0.008 (0.004)	0.025 (0.012)	0.200 (0.043)

LASSO GVEMM2 are the results with larger penalty in GIC (c=0.2 for N=1500, and c=0.1 for N=3000).

Table 5: Study II Power of detecting non-uniform DIF

Corr	N	DIF%	Group	EMM	GVEMM BIC	GVEMM GIC
0.85	1500	20%	Omnibus DIF	0.96 (0.017)	0.63 (0.055)	0.90 (0.029)
			Low DIF	0.55 (0.043)	0.10 (0.033)	0.27 (0.052)
			High DIF	0.96 (0.017)	0.63 (0.055)	0.90 (0.029)
		60%	Omnibus DIF	0.885 (0.019)	0.347 (0.057)	0.817 (0.046)
			Low DIF	0.208 (0.024)	0.037(0.011)	0.243 (0.038)
			High DIF	0.885 (0.019)	0.347 (0.057)	0.813 (0.046)
	3000	20%	Omnibus DIF	1 (0)	0.86 (0.029)	0.98 (0.014)
			Low DIF	0.84 (0.029)	0.20 (0.041)	0.56 (0.051)
			High DIF	1 (0)	0.86 (0.029)	0.98 (0.014)
		60%	Omnibus DIF	0.998 (0.002)	0.840 (0.032)	0.973 (0.054)
			Low DIF	0.632 (0.032)	0.097 (0.018)	0.467 (0.009)
			High DIF	0.998 (0.002)	0.840 (0.032)	0.973 (0.047)

LASSO GVEMM2 are the results with larger penalty in GIC (c=0.2 for N=1500, and c=0.1 for N=3000).

4 Real Data Analysis

To demonstrate the feasibility of the GVEMM algorithm for detecting DIF in real data, we used 5,219 cancer patients' responses to 21 items from two scales, depression (10 items) and anxiety scales (11 items). We especially focused on detecting race DIF because in the sample, race is a categorical variable with four levels and prior researches have studied race DIF on these items using the same sample (Teresi, Ocepek-Welikson, Kleinman, Ramirez, & Kim, 2016a, 2016b), and we intend to compare the results with theirs.

Among the four groups, the reference group is 'Non-Hispanic White' (sample size $n = 2,239$), and three focal groups are 'Non-Hispanic Black' ($n = 1,077$), 'Hispanic' ($n = 1,012$) and 'Non-Hispanic Asians/ Pacific Islanders ($n = 891$)' respectively. As we focused on M2PL model throughout the paper, we created a dichotomous data set by combining response categories. Given the proportion of the 'never' response falls between 50%-65% in most items, we combined the other four response categories (i.e., 'rarely', 'sometimes', 'often' and 'always') and made all 21 items dichotomous. That is, the patient response to each item is either yes or no. This treatment was similar to Bauer et al. (2020). The item content can be found in Wang et al. (2021).

Table 6 presents the DIF detection results using the Lasso GVEMM algorithm, and meanwhile, we also tabulated the flagged items reported in Teresi et al. (2016a, 2016b) as a comparison. In their studies, they used two DIF detection methods: Wald test with Bonferroni correction and ordinal logistic regression. The last three columns of Table 6 are the DIF estimates from our GVEMM algorithm. As shown, our algorithm detected far fewer items as having DIF compare to theirs, and unsurprisingly, our flagged DIF items for the specific pairwise comparison are a complete subset of the marked cases from their methods, except item 18 (i.e., "My worries overwhelmed me") between White vs. Hispanic subgroups. The positive β's indicate that items are harder for the focal group, implying patients in the focal group are less likely to say they have the stated symptoms. This striking difference between GVEMM and traditional methods could be due to the following reason: the version of the traditional methods used in Teresi et al. (2016a, b) did not consider impact. As shown in Table 7, the estimated latent trait distributions in the three focal groups were different from those in the reference group, hence treating them as the same would inevitably bias DIF detection.

Table 6: DIF Detection result from PROMIS anxiety and depression scales

Item	3—c—Wald test	3—c—Logistic regression	3—c—LASSO GVEMM			
	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	Wh v NH
1	✓	✓	✓	✓	✓	✓
2			✓	✓	✓	✓
3			✓	✓	✓	✓
4					✓	✓
5	✓		✓	✓	✓	✓
6					✓	✓
7		✓	✓		✓	✓
8		✓	✓		✓	✓
9	✓	✓	✓	✓	✓	✓
10	✓	✓	✓	✓	✓	✓
11			✓		✓	✓
12	✓	✓	✓	✓	✓	✓
13		✓		✓		✓
14	✓	✓		✓		✓
15					✓	✓
16						✓
17	✓					✓
18			✓	✓	✓	✓
19			✓	✓	✓	✓
20			✓	✓		✓
21						✓

5 Discussion

This note demonstrates the feasibility of using regularized GVEMM for detecting DIF within the regularized explanatory-MIRT framework. Because all model parameters can be updated in closed-forms in the M-step of the GVEMM algorithm, it is computationally more efficient than the traditional EM algorithm. As a grid of tuning parameters are considered in sequel, the final parameter estimates from a

Table 7: Estimated mean and covariance matrix (Impact) from PROMIS analysis

	4—c—Lasso GVEMM			
	White	Black	Hisp.	NHAPI
$\boldsymbol{\mu}_1$	0	0.027	0.279	0.076
$\boldsymbol{\mu}_2$	0	0.130	0.364	0.215
$\boldsymbol{\Sigma}_1^2$	1	1.118	1.021	1.151
$\boldsymbol{\Sigma}_{12}$	.958	1.143	1.023	1.197
$\boldsymbol{\Sigma}_2^2$	1	1.251	1.122	1.322

preceding tuning parameter will be used as warm starting values for the next tuning parameter. This further speeds up the algorithm convergence. According to the simulation results, regularized GVEMM produces almost the same, and sometimes better, Type I error control and power than the EMM algorithm proposed in Wang et al. (2021) to detect uniform DIF. For detecting non-uniform DIF, the reguarlized GVEMM generates slightly either inflated Type I error when using GIC or much lower power when using BIC to select optimal tuning parameter[1]. This is likely due to the fact that the current GVEM algorithm may generate relatively large bias on discrimination parameters in confirmatory MIRT models. Such a bias issue is common to variational estimation for various statistical models (Bishop & Nasrabadi, 2006). In fact, we ran a simple simulation check by introducing DIF only on slope parameters and noted that the regularized GVEMM could barely detect such DIF, resulting in a power lower than 0.2. Therefore, a natural extension of the method is to further fine tune the GVEM algorithm to reduce the bias on discrimination parameter estimates. One possible solution is to use a importance weighted variational

[1] We checked the interim results and noted that the correct solution was actually on the solution path. But neither BIC nore GIC could consistently pick the best tuning parameter.

technique to create a tighter variational lower bound to the marginal likelihood.

Similar to Wang et al. (2021), we only considered L_1 penalty in this note such that a direct comparison between GVEMM and EMM can be established. Due to the inherent bias introduced by the L_1 penalty, one additional M-step without penalty is always needed in the DIF detection context. Another future direction would be to replace L_1 penalty with a truncated L_1 penalty (TLP) proposed by Shen, Pan, and Zhu (2012). The idea is to update (6) by

$$l_{TLP}(\boldsymbol{\Delta}) = \log L(\boldsymbol{\Delta}) - \sum_{j=1}^{J}\Bigg(\eta J_\tau(|\boldsymbol{\beta}_j|) + \eta J_\tau(|\boldsymbol{\gamma}_j|)\Bigg), \tag{45}$$

where for every element in $|\boldsymbol{\beta}_j|$, $J_\tau(|\beta_{jp}|) = \min(|\beta_{jp}|, \tau)$ and $\tau > 0$ is also a tuning parameter. Using the difference of convex method, the objective function in (45) reduces to an adaptive Lasso problem as follows, for item j as an example,

$$l_{TLP}(\boldsymbol{\Delta}_j) = \log L_j(\boldsymbol{\Delta}) - \frac{\eta}{\tau}|\boldsymbol{\beta}_j| I(|\hat{\boldsymbol{\beta}}_j^{(t-1)}| \leq \tau) - \frac{\eta}{\tau}|\boldsymbol{\gamma}_j| I(|\hat{\boldsymbol{\gamma}}_j^{(t-1)}| \leq \tau), \tag{46}$$

where $\hat{\boldsymbol{\beta}}_j^{(t-1)}$ denotes the interim parameter estiamtes from $(t-1)$th iteration. This TLP penalty corrects the Lasso bias through adaptive shrinkage combing shrinkage with thresholding, hence only small $\hat{\boldsymbol{\beta}}_j^{(t-1)}$ will be further shrinked to 0. Certainly, the optimal combination of τ and η will be determined based on information criteria such as BIC or GIC.

Properly identifying DIF is essential for data harmonization, because assuming strict item invariance across groups may sometimes be too strict. The re-MIRT is a flexible modeling framework that simultaneously handles multidimensional traits and potential DIF caused by multiple covariates. It obviates the tedious process of detecting DIF on each item and each covriate one at a time, which is often the case in traditional likelihood-ratio based DIF detection, or the reliance on modification index in confirmatory factor analysis. The regularized GVEM algorithm provides a computationally more efficient alternative to the classic EM algorithm, and it performs very well when DIF occurs on the difficulty parameters. Hence, it has a great potential to serve as a screening tool when analyzing integrated item response data. It is important to note that when dealing with multiple categorical covariates or a categorical covariate with multiple levels, such as in our simulation demonstration, we utilize dummy coding. This involves treating one group as a reference and the remaining groups as focal groups. The regularization DIF detection method then identifies items that exhibit DIF for one or more specific focal groups, relative to the reference group. It is worth noting that selecting a different reference group can yield different results. Additionally, the method does not automatically compare one focal group to another. However, it is possible to manually compare the estimated β and γ parameters for each group. If selecting a specific reference group is difficult to justify, one alternative is to use effect coding by using the population mean as the reference.

References

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

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.

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

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.

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.

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

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.

Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, *90*(432), 1200–1224.

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.

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

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.

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.

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

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.

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*.

Wang, C., Zhu, R., & Xu, G. (2021). Using Lasso and adaptive Lasso to identify DIF in multidimensional 2PL models. *Multivariate Behavioral Research*, 1–21.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. *Assessment of Competencies in Educational Contexts*, 91–120.

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

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.