A Note on DIF Detection in Multidimensional IRT via Regularized Gaussian Variational Estimation

March 3, 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 while considering study-specific variation in the definition and operationalization of theoretical constructs. Because the MIRT model is computationally demanding, we leverage the recently developed Gaussian Variational Expectation-Maximization (GVEM) algorithm to speed up the computation. This note aims to provide empirical evidence to support feasible uses of GVEM for MIRT DIF detection, followed by fast estimation of item parameters ac-

counting for DIF, 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 et al., 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 simultaneously10, 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 (2019) proposed a regularized MNLFA by using a penalized likelihood function that implements a lasso 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, et al., 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, et al., 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 assess-

ment 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 (i.e., least absolute shrinkage and selection operator), 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 (2022) 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 three small-scale simulation study to evaluate the performance of the algorithm in terms of detecting DIF. We will end the paper with final discussions.

2 Method

2.1 Regularized Explanatory MIRT

For a dichotomously scored item j, the probability that person i with a latent trait vector $\boldsymbol{\theta}_i$ giving a correct response to item j is

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

Here J denote test length and K denote the total number of dimensions. In Equation 1, $\boldsymbol{a}_j$ is a K-by-1 vector of discriminations for item j, b_j is a difficulty parameter of item j, and $\boldsymbol{\theta}_i$ is a K-by-1 vector of latent trait for person i. The explanatory feature of the model is reflected by the inclusion of person level covariates, $\boldsymbol{X}_i$, which is a 1-by-P vector including all the grouping information related to DIF (Wilson, De Boeck,

& Carstensen, 2008). $\boldsymbol{\beta}_j$ is also a P-by-1 vector of regression coefficients implying the effect of grouping variables on correct item response probability. Similarly, $\boldsymbol{\gamma}_j$ is a P-by-K matrix of regression coefficients that denote the interaction effects of $\boldsymbol{\theta}$ and grouping variable on item responses. As explained in Wang et al. (2022), 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 parameterizations, if item j does not have DIF, then $\boldsymbol{\gamma}_j = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{0}$. If item j has uniform DIF, then $\boldsymbol{\gamma}_j = \mathbf{0}$. Similar to the multiple-group IRT approach, $\boldsymbol{\theta}_i$ in Equation 1 can be written as $\boldsymbol{\theta}_{i(g)} g = 1, ..., G$ to reflect that the distribution of $\boldsymbol{\theta}$ is allowed to differ across different groups, assuming there are G groups in total.

Denote all model parameters by $\boldsymbol{\Delta} = (\boldsymbol{a}_j, b_j, \boldsymbol{\gamma})_j, \boldsymbol{\beta}_j, j = 1, ..., J, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, g = 1, ..., G$, then the marginal likelihood given covariates $\boldsymbol{X}$ and response $\boldsymbol{y}$ is

$$L(\boldsymbol{\Delta}) \equiv \int L(\boldsymbol{\Delta}|\boldsymbol{X}, \boldsymbol{u}, \boldsymbol{\theta}) \partial \boldsymbol{\theta} = \prod_{g=1}^{G} \prod_{i=1}^{N_g} \int \prod_{j=1}^{J} L(\boldsymbol{a}_j, b_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{X}_i, y_{ij}, \boldsymbol{\theta}) f(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \mid \boldsymbol{\theta}) \partial \boldsymbol{\theta}, \tag{2}$$

where

$$L(\boldsymbol{a}_j, b_j, \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j | \boldsymbol{X}_i, y_{ij}, \boldsymbol{\theta}) = P_j(\boldsymbol{\theta})^{y_{ij}} (1 - P_j(\boldsymbol{\theta}))^{1 - y_{ij}}$$

is the likelihood of item parameters and $f(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \mid \boldsymbol{\theta})$ is a multivariate normal density for group g. N_g is the sample size for group g. $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are the population mean and covariance matrix respectively.

The "regularized " feature of the model is reflected by the penalized likelihood,

with L_1 penalty, defined as follows

$$l_{pen}(\boldsymbol{\Delta}) = \log L(\boldsymbol{\Delta}) - \eta||\boldsymbol{\beta}||_1 - \eta||\boldsymbol{\gamma}||_1, \tag{3}$$

where

$$||\boldsymbol{\beta}||_1 = \sum_j^J \sum_p^P |\beta_{jp}|, \ ||\boldsymbol{\gamma}||_1 = \sum_j^J \sum_p^P \sum_k^K |\gamma_{jpk}| 1_{a_{jk} \neq 0},$$

and $\eta > 0$ is regularization parameters that controls sparsity (Wang et al., 2022).

2.2 Regularized GVEM

The Gaussian variational EM algorithm for the re-MIRT model in Equation 1 differs from the original GVEM algorithm in Cho et al. (2021) in two aspects: it generalizes to the multiple-group scenario and it includes explanatory covariates $\boldsymbol{X}$. As a result, the variational lower bound in the E-step (i.e., section 3.1.2 in Cho et al. (2021)) is updated as follows, for item j

$$\begin{aligned} Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j) = \sum_{g=1}^G \sum_{i=1}^{N_g} \Bigg[& \log \frac{e^{\xi_{i,j}}}{(1+e^{\xi_{i,j}})} + (\frac{1}{2} - Y_{ij})(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j) + (Y_{ij} - \frac{1}{2})(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i \\ & - \frac{1}{2}\xi_{i,j} - \eta(\xi_{i,j})\{(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)^2 - 2(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i \\ & + (\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T[\Sigma_i + (\mu_i)(\mu_i)^T](\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j) - \xi_{i,j}^2\}\Bigg] \\ & + \sum_{g=1}^G \frac{N_g}{2} \log|\Sigma_{\theta g}^{-1}| - \sum_{g=1}^G \sum_{i=1}^{N_g} \frac{1}{2} Tr(\Sigma_{\theta g}^{-1}[\Sigma_i + (\mu_i - \mu_{\theta g})(\mu_i - \mu_{\theta g})^T]), \end{aligned} \tag{4}$$

where $\xi_{i,j}$ is the variational parameter, and $\eta(\xi_{i,j}) = (2\xi_{i,j})^{-1}[e^{\xi_{i,j}}/(1+e^{\xi_{i,j}}) - 1/2]$. μ_i and Σ_i are the mean vector and covariance matrix of the variational density for person i, i.e., $q_i(\boldsymbol{\theta}_i) \sim N(\mu_i, \Sigma_i)$, and they are updated iteratively during the EM cycles. $\boldsymbol{\theta} \sim N(\mu_{\theta g}, \Sigma_{\theta g})$ is the distribution of the latent trait $\boldsymbol{\theta}$ in group g, and N_g is the sample size of group g.

In the M-step, each model parameter is updated separately, and the details are given below.

- Variational parameters μ_i, Σ_i, and $\xi_{i,j}$ are updated using the following closed forms, which are natural extension of the update formula in Cho et al. (2021).

$$\Sigma_i^{-1} = \Sigma_{\theta g}^{-1} + 2\sum_{j=1}^{J} \eta(\xi_{i,j})(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T \tag{5}$$

$$\mu_i = \Sigma_i \times \{\sum_{j=1}^{J}[2\eta(\xi_{i,j})(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j) + Y_{ij} - \frac{1}{2}](\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T + \Sigma_{\theta g}^{-1}\mu_{\theta g}\} \tag{6}$$

$$\begin{aligned} \xi_{i,j}^2 = {} & (b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)^2 - 2(b_j - \boldsymbol{X}_i\boldsymbol{\beta}_j)(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i \\ & + (\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T[\Sigma_i + \mu_i\mu_i^T](\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j) \end{aligned} \tag{7}$$

- The latent trait distribution parameters μ_θ and Σ_θ are updated as follows. First, for model identification, we fix $\mu_{\theta 1} = \mathbf{0}$. For g=2,...,G we have

$$\mu_{\theta g} = \frac{1}{N_g}\sum_{i=1}^{N_g} \mu_i. \tag{8}$$

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

$$\Sigma_{\theta g} = \frac{1}{N_g} \sum_{i=1}^{N_g} [\Sigma_i + (\mu_i - \mu_\theta)(\mu_i - \mu_\theta)^T] \tag{9}$$

Note that for the reference group, we fix the diagonal elements of $\Sigma_{\theta 1}$ to 1, whereas re-scale $\Sigma_{\theta g}$ by

$$\Sigma^*_{\theta g} = ((\sqrt{diag(\Sigma_{\theta 1})})^{-1})^T \Sigma_{\theta g} (\sqrt{diag(\Sigma_{\theta 1})})^{-1} \tag{10}$$

- Item discrimination parameters $\boldsymbol{a}_j$ is updated by solving the following first derivative,

$$\begin{aligned} \frac{\partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)}{\partial \boldsymbol{a}_j} = \sum_{i=1}^{N} \Big[(Y_{ij} - \frac{1}{2})\mu_i - \eta(\xi_{i,j}) \big(- 2(b_j - \boldsymbol{X}_i \boldsymbol{\beta}_j)\mu_i \\ + 2(\boldsymbol{a}_j + \boldsymbol{X}_i \boldsymbol{\gamma}_j)^T [\Sigma_i + (\mu_i)(\mu_i)^T] \big) \Big] = 0. \end{aligned} \tag{11}$$

This results in a closed-form solution for $\boldsymbol{a}_j$ as follows,

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

When the unit variance constraint is imposed on the reference group, $\boldsymbol{a}_j$'s also need to be rescaled using the following simple transformation

$$\hat{\boldsymbol{a}}_j = \hat{\boldsymbol{a}}_j \sqrt{diag(\Sigma_{\theta 1})} \tag{12}$$

- The item difficulty parameter b_j is updated by solving the first derivative of Q_j with respect to b_j as follows,

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

It results in a closed-form solution of $\hat{b}_j$

$$\hat{b}_j = \frac{\sum_{i=1}^{N} \left((\frac{1}{2} - Y_{ij}) + 2\eta(\xi_{i,j})\boldsymbol{X}_i\boldsymbol{\beta}_j + 2\eta(\xi_{i,j})(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i \right)}{\sum_{i=1}^{N} 2\eta(\xi_{i,j})}. \tag{13}$$

- Let $\boldsymbol{\gamma}_{jg} = \boldsymbol{X}_i\boldsymbol{\gamma}_j$ denote the DIF parameter on slope for the covariate group g (g=1,...,G), and similar to Wang et al. (2022), we used a quadratic approximation of Q_j, and derive a closed-form update of $\boldsymbol{\gamma}_{jg}$ as follows. First, obtain the first and second derivatives of Q_j with respect to $\boldsymbol{\gamma}_{jg}$ respectively, i.e.,

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

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

where N_g is the sample size of group g.

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

Rescale $\hat{\boldsymbol{\gamma}}_{jg} = \hat{\boldsymbol{\gamma}}_{jg}\sqrt{diag(\Sigma_{\theta 1})}$. Note that writing $\boldsymbol{\gamma}_{jg} = \boldsymbol{X}_i\boldsymbol{\gamma}_j$ implies that only categorical covariates are considered. $S(f, \lambda) = sign(f)(|f| - \lambda)_+$ is a soft thresholding operator (Donoho & Johnstone, 1995).

- Let $\beta_{jg} = \boldsymbol{X}_i\boldsymbol{\beta}_j$ denote the uniform DIF on item j caused by covariate group g, then similarly, we can first derive the first and second derivatives of Q_j with respect to β_{jg}

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

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

$$\begin{aligned}\hat{\beta}_{jg} &= -\frac{S\left(-\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j) * \beta^*_{jg} + \partial Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j), \lambda\right)}{\partial^2 Q_j(\boldsymbol{a}_j, b_j, \boldsymbol{\gamma}_j, \boldsymbol{\beta}_j)} \\ &= \frac{S\left(\sum_{i=1}^{N_g} -(\frac{1}{2} - Y_{ij}) - \eta(\xi_{i,j})\{-2b_j + 2(\boldsymbol{a}_j + \boldsymbol{X}_i\boldsymbol{\gamma}_j)^T\mu_i\}, \lambda\right)}{\sum_{i=1}^{N_g} 2\eta(\xi_{i,j})}\end{aligned} \tag{15}$$

During each EM cycle, it is necessary to re-estimate the non-zero model parameters and use the updated coefficients to perform the next E-step. One additional M-step can be carried out without a penalty, using the DIF detection results from

the previous M-step within the same EM cycle to constrain certain elements in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ at 0. As a result, each E-step is followed by two M-steps, the first one with a penalty, and the second one without a penalty. The algorithm is therefore termed as GVEMM algorithm.

To find the best value of tuning parameter η, two information criteria are calculated. Zhang, Li, and Tsai (2010) proposed the generalized information criterion (GIC) for selecting tuning parameters, and we modify it for GVEM, as follows

$$GIC_\eta = -2\sum_{j=1}^{J} Q_j(\hat{\boldsymbol{a}}_j, \hat{b}_j, \hat{\boldsymbol{\gamma}}_j, \hat{\boldsymbol{\beta}}_j) + k_n(\|\hat{\boldsymbol{\beta}}\|_0 + \|\hat{\boldsymbol{\gamma}}\|_0). \tag{16}$$

In Equation 16, k_n can take on different values. For instance, when $k_n = \log(N)$, GIC becomes BIC. Otherwise, $k_n = c\log(\log(N)) \times \log(N)$, where c is a constant. For each value of η, we can get a set of L_1 regularized Marginal Maximum Likelihood Estimator $\boldsymbol{\Delta}_\lambda$.

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. The latent variables of three groups, one reference group and

two focal groups, were generated from bi-variate 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 has small DIF magnitude ($\boldsymbol{\beta}_1 = 0.5$) and the second focal group has large DIF ($\boldsymbol{\beta}_1 = 1$). In 20% DIF condition, four items 4, 5, 12, 13 are DIF. In 60% DIF condition, 12 items 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17 are DIF. We considered both BIC and GIC as the criterion for selecting the best tuning paramter 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.

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)	0.0125 (0.005)
			Low DIF	0.025 (0.009)	0.005 (0.005)	0.0125 (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)

Table 3: Study I Power of detecting uniform DIF

Corr	N	DIF%	Group	LASSO EMM	LASSO GVEMM	LASSO GVEMM2
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)

LASSO GVEMM2 are the results with larger penalty in GIC.

Table 4: BIC and GIC for different tuning parameter values in Study I condition 1 (N=1500, 20% DIF)

λ	10	12	14	16	18	20	22	24	26
-2*LL	26185	26185	26179	26179	26204	26238	26086	26086	26173
BIC	26236	26236	26223	26223	26233	26260	26101	26101	26180
$\log N * \lvert\beta\rvert_0$	51.19	51.19	43.87	43.87	29.25	21.93	14.62	14.62	7.31
GIC	26287	26286	26266	26266	26262	26281	26115	26115	26187
$\log(\log N) * \log N * \lvert\beta\rvert_0$	101.8	101.8	87.31	87.31	58.20	43.65	29.10	29.10	14.55
Group 1 power	0.25	0.25	0.25	0.25	0.25	0	0	0	0
Group 1 Type I	0.1	0.1	0	0	0	0	0	0	0
Group 2 power	1	1	1	1	0.75	0.75	0.5	0.5	0.25
Group 2 Type I	0.1	0.1	0.1	0.1	0	0	0	0	0

LASSO GVEMM2 are the results with larger penalty in GIC.

3.2 Simulation II

Non-uniform DIF is studied in simulation II, with DIF effects on both intercept and slope are simulated. The first focal group with small DIF magnitude has $\boldsymbol{\beta}_1 = 0.25$ and $\boldsymbol{\gamma}_1 = -0.4$ on items with DIF, and the second focal group with large DIF has $\boldsymbol{\beta}_1 = 0.6$ and $\boldsymbol{\gamma}_1 = -0.6$. Again, in 20% DIF condition, four items (4, 5, 12, 13) are DIF. In 60% DIF condition, 12 items 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17) are DIF.

The LASSO GVEM and LASSO GVEMM are the same as in simulation I. In table 4 and table 5, the LASSO GVEMM2 are the results of $GIC = -2 * ll + 3 * \log(\log N) * \log N * |\beta|_0$. The penalty is smaller than the penalty in GIC in simulation I, but the power is still relatively low when sample size is small. We can further reduce the constant in GIC in the LASSO GVEMM2 method.

In addition to the above three methods, we further try to use uniform DIF GVEMM to detect non-uniform DIF. The method was only applied to one condition (correlation 0.85, N=1500, 20% DIF) and the results named LASSO GVEMM$_u$ are

shown in the small separate table below.

Table 5: Study II Type I error of detecting non-uniform DIF

Corr	N	DIF%	Group	LASSO EMM	LASSO GVEMM	LASSO GVEMM2
0.85	1500	20%	Omnibus DIF	0.021 (0.005)	0.075 (0.023)	0.010 (0.006)
			Low DIF	0.013 (0.004)	0.045 (0.016)	0.005 (0.004)
			High DIF	0.011 (0.003)	0.075 (0.017)	0.008 (0.006)
		60%	Omnibus DIF	0.035 (0.011)	0.0875 (0.028)	0.025 (0.018)
			Low DIF	0.025 (0.009)	0.05 (0.022)	0.025 (0.018)
			High DIF	0.013 (0.005)	0.0375 (0.020)	0 (0)
	3000	20%	Omnibus DIF	0.026 (0.006)	0.18 (0.048)	0.043 (0.017)
			Low DIF	0.021 (0.005)	0.13 (0.033)	0.043 (0.017)
			High DIF	0.006 (0.003)	0.1 (0.046)	0.006 (0.007)
		60%	Omnibus DIF	0.060 (0.015)	0.150 (0.051)	0.088 (0.040)
			Low DIF	0.233 (0.015)	0.113 (0.036)	0.063 (0.022)
			High DIF	0.008 (0.004)	0.075 (0.022)	0.05 (0.029)

Corr	N	DIF%	Group	LASSO GVEMM$_u$
0.85	1500	20%	Omnibus DIF	0.143 (0.053)
			Low DIF	0.106 (0.043)
			High DIF	0.081 (0.031)

LASSO GVEMM2 are the results with larger penalty in GIC.

LASSO GVEMM$_u$ are the results estimated by uniform DIF algorithm.

3.3 Simulation III

Non-uniform DIF only on slope is studied in simulation III. The first focal group with small DIF magnitude has $\boldsymbol{\gamma}_1 = -0.5$ and the second focal group with large DIF

Table 6: Study II Power of detecting non-uniform DIF

Corr	N	DIF%	Group	LASSO EMM	LASSO GVEMM	LASSO GVEMM2
0.85	1500	20%	Omnibus DIF	0.96 (0.017)	0.59 (0.067)	0.38 (0.044)
			Low DIF	0.55 (0.043)	0.13 (0.044)	0.05 (0.020)
			High DIF	0.96 (0.017)	0.59 (0.067)	0.38 (0.044)
		60%	Omnibus DIF	0.885 (0.019)	0.497 (0.074)	0.508 (0.097)
			Low DIF	0.208 (0.024)	0.113 (0.029)	0.108 (0.045)
			High DIF	0.885 (0.019)	0.490 (0.071)	0.508 (0.097)
	3000	20%	Omnibus DIF	1 (0)	0.95 (0.035)	0.925 (0.040)
			Low DIF	0.84 (0.029)	0.475 (0.073)	0.35 (0.070)
			High DIF	1 (0)	0.95 (0.035)	0.925 (0.040)
		60%	Omnibus DIF	0.998 (0.002)	0.983 (0.012)	0.966 (0.019)
			Low DIF	0.632 (0.032)	0.350 (0.054)	0.275 (0.064)
			High DIF	0.998 (0.002)	0.983 (0.012)	0.966 (0.019)

Corr	N	DIF%	Group	LASSO GVEMM$_u$
0.85	1500	20%	Omnibus DIF	0.825 (0.056)
			Low DIF	0.275 (0.121)
			High DIF	0.825 (0.053)

LASSO GVEMM2 are the results with larger penalty in GIC.

LASSO GVEMM$_u$ are the results estimated by uniform DIF algorithm.

has $\boldsymbol{\gamma}_1 = -1$.

According to our multiple-group studies, GVEM algorithms could not estimate discrimination parameters accurately. So DIF could not be detected when we have DIF only on slope. Here we tried a few different conditions, including reduce the correlation between two trait dimensions to 0.1, increase test length to 60, the DIF detection result was improved but the power was still low.

Table 7: Study III (first version MBR) Type I error of detecting non-uniform DIF

Corr	N	DIF%	Group	LASSO EMM	LASSO GVEMM
0.85	1500	20%	Omnibus DIF	0.036	0.038
			Low DIF	0.020	0.018
			High DIF	0.020	0.025
	3000	20%	Omnibus DIF	0.035	0
			Low DIF	0.017	0
			High DIF	0.026	0

Corr	N	Test Len	DIF%	Group	LASSO GVEMM
0.1	3000	60	20%	Omnibus DIF	0.047
				Low DIF	0.004
				High DIF	0.043

4 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 preceding tuning parameter will be used as warm starting values of GVEMM 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). For detecting non-uniform DIF, the reguarlized GVEMM generates slightly inflated Type I error rate. 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, 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 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 Equation 3 by

$$l_{TLP}(\boldsymbol{\Delta}) = \log L(\boldsymbol{\Delta}) - \sum_{j=1}^{J} \Big(\eta J_\tau(|\boldsymbol{\beta}_j|) + \eta J_\tau(|\boldsymbol{\gamma}_j|) \Big), \tag{17}$$

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 Equation 17 reduces to an adaptive Lasso problem as follows, for item j as an example,

$$l_{TLP}(\boldsymbol{\Delta}_j) = \log L_j(\boldsymbol{\Delta}) - \frac{\eta}{\tau}|\boldsymbol{\beta}_j| I(|\hat{\boldsymbol{\beta}}_j^{(t-1)}| \leq \tau) - \frac{\eta}{\tau}|\boldsymbol{\gamma}_j| I(|\hat{\boldsymbol{\gamma}}_j^{(t-1)}| \leq \tau), \tag{18}$$

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.

References