DIF Detection via Regularization

Ruoyi Zhu

9/9/2020

Contents

1 Conventional DIF Detection Methods

1.1 Likelihood Ratio Test

In likelihood ratio test, DIF will be tested one item at a time. The null hypothesis is that all but the anchor items are DIF. To test whether the jth item is DIF, the alternative hypothesis is that the anchor items and

the jth item is DIF free, and all other items are DIF. If the null hypothesis is rejected (adjust p-value <0.05), the tested item will be flagged as DIF. Since we are testing several hypotheses, multiple comparisons need to be performed. The adjust p-values are the family-wise error rate. We assume the latent traits in reference and focal groups follow normal distributions. The mean and variance of reference group are fixed to be 0 and 1, respectively. The means and variances of focal groups can be freely estimated.

2 Multidimensional 2PL Model with DIF

Assume a total sample size $N = N_1 + N_2 + N_3$, where N_1, N_2 and N_3 are sample sizes of the reference group and two focal groups, respectively. Also assume a test length m and trait dimension q. For a dichotomously scored item j, the probability that examinee i with ability vector $\boldsymbol{\theta}_i$ giving a correct response to item j is

$$P_j(\boldsymbol{\theta}_i) = \frac{1}{1 + e^{-(\mathbf{a}_j^T\boldsymbol{\theta}_i + d_j + (\boldsymbol{y}_i\boldsymbol{\gamma}_j)\boldsymbol{\theta}_i + \boldsymbol{y}_i\boldsymbol{\beta}_j)}} (i = 1, ..., N; j = 1, 2, ..., m). \tag{1}$$

$\boldsymbol{y}_i$ is a group indicator including all the grouping information related to DIF. $\boldsymbol{y}_i = (0, 0)$ if examinee i is in the reference group, $\boldsymbol{y}_i = (1, 0)$ if examinee i is in the first focal group and $\boldsymbol{y}_i = (0, 1)$ if examinee i is in the second focal group.

The ability vector of the ith examinee is $\boldsymbol{\theta}_i = (\theta_{i1},\theta_{i2},\ldots,\theta_{ir},\ldots,\theta_{iq})^T$ (i=1,...,N; r=1,2,...,q). $\mathbf{a}_j = (a_{j1}, a_{j2})^T$ is the discrimination parameter and d_j is the boundary parameter.

$$\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{j1\cdot}, \boldsymbol{\gamma}_{j2\cdot})^T = \begin{pmatrix} \gamma_{j11} & \gamma_{j12} & \cdots & \gamma_{j1r} & \cdots & \gamma_{j1q} \\ \gamma_{j21} & \gamma_{j22} & \cdots & \gamma_{j2r} & \cdots & \gamma_{j2q} \end{pmatrix} (r = 1, ..., q)$$

is the non-uniform DIF parameter, where $\boldsymbol{\gamma}_{j1\cdot}$ is the non-uniform DIF parameter for the first focal group and $\boldsymbol{\gamma}_{j2\cdot}$ is the non-uniform DIF parameter for the second focal group. $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2})^T$ is the uniform DIF parameter, where β_{j1} is the uniform DIF parameter for the first focal group and β_{j2} is the uniform DIF parameter for the second focal group. 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 $\boldsymbol{\gamma}_j = \mathbf{0}$.

Suppose the prior distribution of $\boldsymbol{\theta}_i$ in group y is multivariate normal distribution with mean vector of $\boldsymbol{\mu}_y$ and covariance matrix of $\boldsymbol{\Sigma}_y$. The prior density of $\boldsymbol{\theta}_i$ is

$$f(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_y|^{-\frac{1}{2}} e^{-0.5(\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)^T |\boldsymbol{\Sigma}_y|^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)}.$$

If i is in 1,...,N_1, then $y_i = (0, 0)$. $\boldsymbol{\mu}_y = \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_1$, $\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.

If i is in $N_1 + 1$,...,$N_1 + N_2$, then $y_i = (1, 0)$. $\boldsymbol{\mu}_y = \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_2$, $\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

If i is in $N_1 + N_2 + 1$,...,$N_1 + N_2 + N_3$, then $y_i = (0, 1)$. $\boldsymbol{\mu}_y = \boldsymbol{\mu}_3$ and $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_3$, $\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$.

We set the reference group to have zero means and unit variances, that is, $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\text{diag}(\boldsymbol{\Sigma}_1) = \mathbf{1}$. Then with some anchor items, the trait parameters for focal groups, i.e. $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$, can be freely estimated.

3 Graded Response Model with DIF

Assume a total sample size N, test length m, number of response categories p and trait dimension q. For a polytomously scored item j, the probability that examinee i with ability vector $\boldsymbol{\theta}_i$ reaching level k or higher on item j is

$$P^*_{ijk} = \frac{1}{1 + e^{-(\mathbf{a}_j\boldsymbol{\theta}_i + d_{jk} + (\boldsymbol{y}_i\boldsymbol{\gamma}_j)\boldsymbol{\theta}_i + \boldsymbol{y}_i\boldsymbol{\beta}_{jk})}} (i = 1, ..., N; j = 1, 2, ..., m; k = 1, 2, ..., p - 1). \tag{2}$$

$\boldsymbol{y}_i$ is a group indicator including all the grouping information related to DIF. $\boldsymbol{y}_i = (0, 0)$ if examinee i is in the reference group, $\boldsymbol{y}_i = (1, 0)$ if examinee i is in the first focal group and $\boldsymbol{y}_i = (0, 1)$ if examinee i is in the second focal group.

$$P_{ijk} = P^*_{ij,k-1} - P^*_{ijk} \tag{3}$$

is the probability of an examinee i with ability vector $\boldsymbol{\theta}_i$ reaching response level k on item j.

The trait variable has q dimensions. The ability vector of the ith examinee is $\boldsymbol{\theta}_i = (\theta_{i1},\theta_{i2},\ldots,\theta_{ir},\ldots,\theta_{iq})^T$ (i=1,...,N; r=1,2,...,q), and the item parameter matrices are

discrimination parameter

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_j, ..., \mathbf{a}_m)^T = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1r} & ... & a_{1q} \\ a_{21} & a_{22} & ... & a_{2r} & ... & a_{2q} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{j1} & a_{j2} & ... & a_{jr} & ... & a_{jq} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & ... & a_{mr} & ... & a_{mq} \end{pmatrix} (j = 1, 2, ..., m; r = 1, ..., q),$$

boundary parameter

$$\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_j, ..., \mathbf{d}_m)^T = \begin{pmatrix} d_{11} & d_{12} & ... & d_{1k} & ... & d_{1,p-1} \\ d_{21} & d_{22} & ... & d_{2k} & ... & d_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{j1} & d_{j2} & ... & d_{jk} & ... & d_{j,p-1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & ... & d_{mk} & ... & d_{m,p-1} \end{pmatrix} (j = 1, 2, ..., m; k = 1, ..., p-1),$$

non-uniform DIF parameter

$$\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, ..., \boldsymbol{\gamma}_j, ..., \boldsymbol{\gamma}_m)(j = 1, ..., m)$$

$$\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{j1\cdot}, \boldsymbol{\gamma}_{j2\cdot})^T = \begin{pmatrix} \gamma_{j11} & \gamma_{j12} & \cdots & \gamma_{j1r} & \cdots & \gamma_{j1q} \\ \gamma_{j21} & \gamma_{j22} & \cdots & \gamma_{j2r} & \cdots & \gamma_{j2q} \end{pmatrix} (r = 1, ..., q),$$

where q is the dimension of $\boldsymbol{\theta}$, and each row of $\boldsymbol{\gamma}_j$ is (non-uniform) DIF parameter for a focal group, i.e. $\boldsymbol{\gamma}_{j1\cdot}$ is (non-uniform) DIF parameter for the first focal group, and $\boldsymbol{\gamma}_{j2\cdot}$ is (non-uniform) DIF parameter for the second focal group,

and uniform DIF parameter

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_j, ..., \boldsymbol{\beta}_m)(j = 1, ..., m)$$

$$\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j1\cdot}, \boldsymbol{\beta}_{j2\cdot})^T = \begin{pmatrix} \beta_{j11} & \beta_{j12} & ... & \beta_{j1k} & ... & \beta_{j1,p-1} \\ \beta_{j21} & \beta_{j22} & ... & \beta_{j2k} & ... & \beta_{j2,p-1} \end{pmatrix} (k = 1, ..., p-1),$$

where p is the number of categories in GRM, and each row of $\boldsymbol{\beta}_j$ is (uniform) DIF parameter for a focal group, i.e. $\boldsymbol{\beta}_{j1\cdot}$ is (uniform) DIF parameter for the first focal group, and $\boldsymbol{\beta}_{j2\cdot}$ is (uniform) DIF parameter for the second focal group.

If an item does not have DIF, then $\boldsymbol{\Gamma} = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{0}$. If an item has uniform DIF, then $\boldsymbol{\Gamma} = \mathbf{0}$.

The $N * m$ response matrix is

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & ... & u_{1j} & ... & u_{1m} \\ u_{21} & u_{22} & ... & u_{2j} & ... & u_{2m} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{i1} & u_{i2} & ... & u_{ij} & ... & u_{im} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{N1} & u_{N2} & ... & u_{Nj} & ... & u_{Nm} \end{pmatrix} (i = 1, ..., N; j = 1, 2, ..., m).$$

A dummy variable to indicate whether examinee i gets score k for the item j

$$x_{ijk} = \begin{cases} 1, & \text{if } u_{ij} = k \\ 0, & \text{if } u_{ij} \neq k \end{cases}.$$

$\boldsymbol{y}_i$ is the group indicator. $\boldsymbol{y}_i = (0, 0)$ stands for the reference group, $\boldsymbol{y}_i = (1, 0)$ stands for the rfirst focal group and $\boldsymbol{y}_i = (0, 1)$ stands for the second focal group. The sample size of the reference group, the first focal group and the second focal group are denoted by N_1, N_2, N_3, respectively. We have the total sample size $N = N_1 + N_2 + N_3$. We have

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_{N_1} \\ \boldsymbol{y}_{N_1+1} \\ \boldsymbol{y}_{N_1+2} \\ \vdots \\ \boldsymbol{y}_{N_1+N_2} \\ \boldsymbol{y}_{N_1+N_2+1} \\ \boldsymbol{y}_{N_1+N_2+2} \\ \vdots \\ \boldsymbol{y}_{N_1+N_2+N_3} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}.$$

Suppose the prior distribution of $\boldsymbol{\theta}_i$ in group y is multivariate normal distribution with mean vector of $\boldsymbol{\mu}_y$ and covariance matrix of $\boldsymbol{\Sigma}_y$. The prior density of $\boldsymbol{\theta}_i$ is

$$f(\boldsymbol{\theta}_i \mid \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_y|^{-\frac{1}{2}} e^{-0.5(\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)^T |\boldsymbol{\Sigma}_y|^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)}.$$

If i is in 1,...,N_1, then $y_i = (0, 0)$. $\boldsymbol{\mu}_y = \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_1$, $\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.

If i is in $N_1 + 1$,...,$N_1 + N_2$, then $y_i = (1, 0)$. $\boldsymbol{\mu}_y = \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_2$, $\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

If i is in $N_1 + N_2 + 1$,...,$N_1 + N_2 + N_3$, then $y_i = (0, 1)$. $\boldsymbol{\mu}_y = \boldsymbol{\mu}_3$ and $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_3$, $\boldsymbol{\theta}_i \sim N(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$.

We have $\boldsymbol{\mu}_1 = \mathbf{0}$ and all elements on the diagonal of $\boldsymbol{\Sigma}_1$ are 1 for the reference group. Then with some anchor items, the trait parameters for focal groups, i.e. $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$, can be freely estimated.

Denote G_0 as the number of points we evenly take from each coordinate dimension. Then $G = G_0^q$ quadrature samples (same for all examinees) are denoted by $\boldsymbol{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_g, ..., \boldsymbol{q}_G)^T (g = 1, ..., G)$, and $\boldsymbol{q}_g$ = (q_{g1},q_{g2},...,q_{gr},...,q_{gq}) (r=1,2,...,q). At iteration t, we calculate $f(\boldsymbol{q}_g \mid \boldsymbol{\mu}_y^{(t-1)}, \boldsymbol{\Sigma}_y^{(t-1)})$ for each group y, where $\boldsymbol{\mu}_y^{(t-1)}$ and $\boldsymbol{\Sigma}_y^{(t-1)}$ are the estimated trait parameters from last iteration.

For an examinee i in the reference group (group 1), we have $y = 1$ and

$$P^*_{ijk|q_g} = P^*_{jyk|q_g} = P^*_{j1k|q_g} = \frac{1}{1+e^{-(\mathbf{a}_j \boldsymbol{q}_g + d_k)}}$$

$$(i = 1, ..., N_1; j = 1, 2, ..., m; k = 1, 2, ..., p-1; g = 1, ..., G).$$

For an examinee i in the first focal group (group 2), $y = 2$ and

$$P^*_{ijk|q_g} = P^*_{jyk|q_g} = P^*_{j2k|q_g} = \frac{1}{1+e^{-(\mathbf{a}_j \boldsymbol{q}_g + d_k + \boldsymbol{\gamma}_{j1\cdot} \boldsymbol{q}_g + \boldsymbol{\beta}_{j1k})}}$$

$$(i = N_1 + 1, ..., N_1 + N_2; j = 1, 2, ..., m; k = 1, 2, ..., p-1; g = 1, ..., G).$$

For the second focal group (group 3), $y = 3$ and

$$P^*_{ijk|q_g} = P^*_{jyk|q_g} = P^*_{j3k|q_g} = \frac{1}{1+e^{-(\mathbf{a}_j \boldsymbol{q}_g + d_k + \boldsymbol{\gamma}_{j2\cdot} \boldsymbol{q}_g + \boldsymbol{\beta}_{j2k})}}.$$

$$(i = N_1 + N_2 + 1, ..., N_1 + N_2 + N3; j = 1, 2, ..., m; k = 1, 2, ..., p-1; g = 1, ..., G).$$

$$P_{jyk|q_g} = P^*_{j,y,k-1|q_g} - P^*_{jyk|q_g}$$

4 Model Identifiability Constraint

Some constraints on the item parameters are required for model identification. Here, for each dimension, we set one anchor item which we know its DIF parameters (Γ and β) are zero for all groups.

For instance, if we have two ability dimensions (q=2), test length $m = 20$, and the each factor is loaded on 10 items, then the simple structure discrimination parameter matrix will take the form

$$\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_m)^T = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \\ a_{31} & 0 \\ a_{41} & 0 \\ a_{51} & 0 \\ . & . \\ . & . \\ . & . \\ a_{10,1} & 0 \\ a_{11,1} & 0 \\ 0 & a_{12,2} \\ 0 & a_{13,2} \\ . & . \\ . & . \\ . & . \\ 0 & a_{19,2} \\ 0 & a_{20,2} \end{pmatrix},$$

and the DIF parameters are

$$\boldsymbol{\Gamma} = (\mathbf{0}, \mathbf{0}, \boldsymbol{\Gamma}_3, ..., \boldsymbol{\Gamma}_m)$$

and

$$\boldsymbol{\beta} = (\mathbf{0}, \mathbf{0}, \boldsymbol{\beta}_3, ..., \boldsymbol{\beta}_m).$$

We further assume the reference group has mean zero and variance one and only estimate its correlation, and the means and all the elements in covariance matrices of two focal groups can be freely estimated.

5 Uniform DIF Detection via LASSO

As mentioned before, if an item has uniform DIF, then $\boldsymbol{\Gamma} = \mathbf{0}$. The DIF parameter we are estimating is only $\boldsymbol{\beta} = (\mathbf{0}, ..., \mathbf{0}, \boldsymbol{\beta}_{q+1}, ..., \boldsymbol{\beta}_m)$.

5.1 E step

For an examinee with ability $\boldsymbol{\theta}_i$ the conditional likelihood of observing $\boldsymbol{u}_i$ is

$$L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\theta}_i \mid \mathbf{y}, \boldsymbol{u}_i) = \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}}. \tag{4}$$

With the assumption of prior distribution of latent trait, the joint likelihood of $\boldsymbol{u}_i$ and $\boldsymbol{\theta}_i$ is

$$\begin{aligned} &L(\mathbf{A}, \mathbf{D}, \beta, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) = L(\mathbf{A}, \mathbf{D}, \beta, \boldsymbol{\theta}_i \mid \mathbf{y}, \mathbf{u}_i) f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i) \\ &= \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}} (2\pi)^{-p/2} |\boldsymbol{\Sigma}_y|^{-1/2} \exp(-0.5(\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)' \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)). \end{aligned} \tag{5}$$

Therefore, the marginal likelihood of $\boldsymbol{u}_i$ is

$$m(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i) = \int L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i) \partial \boldsymbol{\theta}_i \tag{6}$$

Then

$$h(\boldsymbol{\theta}_i \mid \boldsymbol{u}_i, \boldsymbol{y}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \Sigma_y^{(t-1)}) = \frac{L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i)}{m(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i)} \tag{7}$$

is the posterior density of $\boldsymbol{\theta}_i$ given the estimation of $\mathbf{A}$, $\mathbf{D}$, $\boldsymbol{\beta}$ and Σ at the iteration t.

The expected complete data log-likelihood with respect to the posterior distribution of $\boldsymbol{\theta}$

$$\begin{aligned} &E[log\{L(\boldsymbol{A}, \boldsymbol{D}, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{\Theta})\} | \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{Y}, \mathbf{U}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}] \\ &= \sum_{i}^{N} \{ \int \log L(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{u}_i, \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}) \partial \boldsymbol{\theta}_i \\ &+ \int \log f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \boldsymbol{\Sigma}_y^{(t-1)}) \partial \boldsymbol{\theta}_i \} \end{aligned} \tag{8}$$

At iteration t, applying Gauss-Hermite quadrature nodes and the integration above can be updated as

$$
\begin{aligned}
&E[logL(\boldsymbol{A},\boldsymbol{D},\beta,\boldsymbol{\mu},\boldsymbol{\Sigma}\mid \boldsymbol{Y},\boldsymbol{U})]\\
&=\sum_{i}^{N}\sum_{g}^{G}\log L(\boldsymbol{A},\boldsymbol{D},\boldsymbol{\beta}\mid \boldsymbol{u}_i,\boldsymbol{q}_g)\frac{L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}\\
&+\sum_{i}^{N}\sum_{g}^{G}\log f(\boldsymbol{\mu},\boldsymbol{\Sigma}\mid \boldsymbol{q}_g)\frac{L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}\\
&=\sum_{i}^{N}\sum_{g}^{G}\sum_{j}^{m}\sum_{k}^{p}x_{ijk}\log P_{ijk|q_g}\frac{L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}\\
&+\sum_{i}^{N}\sum_{g}^{G}\log f(\boldsymbol{\mu},\boldsymbol{\Sigma}\mid \boldsymbol{q}_g)\frac{L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{Y},\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}
\end{aligned}
\tag{9}
$$

Then we can define two artificial terms.

For the reference group, $y=1$. We have

$$
n_{gy}=n_{g1}=\sum_{i=1}^{N_1}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_1^{(t-1)},\boldsymbol{\Sigma}_1^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_1^{(t-1)},\boldsymbol{\Sigma}_1^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)},
$$

and

$$
r_{gjyk}=r_{gj1k}=\sum_{i=1}^{N_1}x_{ijk}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_1^{(t-1)},\boldsymbol{\Sigma}_1^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_1^{(t-1)},\boldsymbol{\Sigma}_1^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}.
$$

For the first focal group, $y=2$. We have

$$
n_{gy}=n_{g2}=\sum_{i=N_1+1}^{N_1+N_2}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_2^{(t-1)},\boldsymbol{\Sigma}_2^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_2^{(t-1)},\boldsymbol{\Sigma}_2^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)},
$$

and

$$
r_{gjyk}=r_{gj2k}=\sum_{i=N_1+1}^{N_1+N_2}x_{ijk}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_2^{(t-1)},\boldsymbol{\Sigma}_2^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_2^{(t-1)},\boldsymbol{\Sigma}_2^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}.
$$

For the second focal group, $y=3$. We have

$$
n_{gy}=n_{g3}=\sum_{i=N_1+N_2+1}^{N_1+N_2+N3}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_3^{(t-1)},\boldsymbol{\Sigma}_3^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_3^{(t-1)},\boldsymbol{\Sigma}_3^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}
$$

and

$$
r_{gjyk}=r_{gj3k}=\sum_{i=N_1+N_2+1}^{N_1+N_2+N3}x_{ijk}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_3^{(t-1)},\boldsymbol{\Sigma}_3^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}{\sum_{g}^{G}L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_3^{(t-1)},\boldsymbol{\Sigma}_3^{(t-1)})f(\boldsymbol{\mu}^{(t-1)},\boldsymbol{\Sigma}^{(t-1)}\mid \boldsymbol{q}_g)}.
$$

$n_g = n_{g1} + n_{g2} + n_{g3}$ represents the expected number of examinees with the ability $\boldsymbol{q}_g$, and $r_{jgk} = r_{jgk1} + r_{jgk2} + r_{jgk3}$ is the expected number of examinees who get the score level k on the item j with the ability $\boldsymbol{q}_g$.

$$E[log\{L(\mathbf{A}, \mathbf{D}, \beta, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{U}, \boldsymbol{\Theta})] = \sum_{g}^{G}\sum_{j}^{m}\sum_{y}^{3}\sum_{k}^{p}(r_{gjyk}\log P_{jyk|q_g}) + \sum_{g}^{G}\sum_{y}^{3} n_{gy}\log f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{10}$$

In the EM problem, we want to maximize the above expectation at the iteration t. Denote this unpenalized expectation as $\log M$.

For each item j, we define

$$\log M_j = \sum_{g}^{G}\sum_{y}^{3}\sum_{k}^{p}(r_{jgky}\log P_{jky|q_g}) + \sum_{g}^{G}\sum_{y}^{3} n_{gy}\log f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{11}$$

In our uniform DIF detection problem, the maximum likelihood method does not serve the purpose of DIF variable selection. We apply lasso and minimize the following objective function

$$-\log M + \eta \sum_{j}^{m} ||\boldsymbol{\beta}_j||_1 \tag{12}$$

For each item, we minimize

$$-\log M_j + \eta||\boldsymbol{\beta}_j||_1 \tag{13}$$

where η is the lasso tuning parameter.

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\beta}}) = \text{argmin}\{-\log M + \eta||\boldsymbol{\beta}||_1\} \tag{14}$$

5.2 M step

In our DIF detection problem, we assume the reference group has mean zero and variance one and only estimate the correlation, and the means and all the elements in covariance matrices of two focal groups can be freely estimated.

In quadrature method, at the iteration t, the first partial derivative with respect to μ is

$$\begin{aligned}
\frac{\partial \log M}{\partial \boldsymbol{\mu}_y} &= \sum_{g}^{G} n_{gy}\frac{\partial \log f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g)}{\partial \boldsymbol{\mu}_y} \\
&= \sum_{g}^{G} n_{gy}\frac{\partial -\frac{1}{2}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)}{\partial \boldsymbol{\mu}_y} \\
&= \sum_{g}^{G} n_{gy}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)\boldsymbol{\Sigma}_y^{-1}
\end{aligned} \tag{15}$$

Set $\frac{\partial \log M}{\partial \boldsymbol{\mu}_y} = 0$, and we know that $\sum_g^G n_{gy} = N_y$.

$\hat{\boldsymbol{\mu}}_y$ can be updated as

$$\hat{\boldsymbol{\mu}}_2 = \frac{\sum_{g=1}^{G} n_{g2}\boldsymbol{q}_g}{N_2}, \tag{16}$$

and

$$\hat{\boldsymbol{\mu}}_3 = \frac{\sum_{g=1}^{G} n_{g3}\boldsymbol{q}_g}{N_3}. \tag{17}$$

The first partial derivative with respect to $\boldsymbol{\Sigma}$ is

$$\begin{aligned}
\frac{\partial \log M}{\partial \boldsymbol{\Sigma}_y} &= \sum_{g}^{G} n_{gy} \frac{\partial \log f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g)}{\partial \boldsymbol{\Sigma}_y} \\
&= \sum_{g}^{G} n_{gy} \frac{\partial(-\frac{g}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_y| - \frac{1}{2}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{q}_g - \boldsymbol{\mu}_y))}{\partial \boldsymbol{\Sigma}_y} \\
&= \sum_{g}^{G} n_{gy}[-\frac{1}{2}\Sigma_y^{-1} + \frac{1}{2}\Sigma_y^{-1}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)(\boldsymbol{q}_g - \boldsymbol{\mu}_y)^T\Sigma_y^{-1}]
\end{aligned} \tag{18}$$

Set $\frac{\partial \log M}{\partial \boldsymbol{\mu}_y} = 0$, and use the fact that $\sum_g^G n_{gy} = N_y$.

$\hat{\boldsymbol{\Sigma}}_y$ can be updated as

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\sum_{g=1}^{G} n_{g1}\boldsymbol{q}_g\boldsymbol{q}_g'}{N_1}, \tag{19}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \frac{\sum_{g=1}^{G} n_{g2}(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{20}$$

and

$$\hat{\boldsymbol{\Sigma}}_3 = \frac{\sum_{g=1}^{G} n_{g3}(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{21}$$

To standardize the covariance matrix, we calculate standardized quadrature points for the later steps.

$$\boldsymbol{q}_g^* = \frac{q_g}{\sqrt{\text{diag}\hat{\boldsymbol{\Sigma}}_1}}. \tag{22}$$

Then we do the following transformation on mean vector and covariance matrices for three groups.

$$\hat{\boldsymbol{\Sigma}}_1^* = \frac{\sum_{g=1}^{G} n_{g1}\boldsymbol{q}_g^*\boldsymbol{q}_g^{*'}}{N_1}, \tag{23}$$

$$\hat{\boldsymbol{\Sigma}}_2^* = \frac{\sum_{g=1}^{G} n_{g2}(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{24}$$

and

$$\hat{\boldsymbol{\Sigma}}_3^* = \frac{\sum_{g=1}^{G} n_{g3}(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{25}$$

the first partial derivative with respect to a_{jr} is

$$\frac{\partial \log M}{\partial a_{jr}} = \sum_{g=1}^{G}\sum_{y}^{3}\sum_{k=1}^{p}(\frac{r_{gjyk}q_{gr}}{P_{jky|q_g}}(\omega_{j,y,k-1} - \omega_{jyk})) \tag{26}$$

where $\omega_{jyk} = P^*_{jyk|q_g} - (P^*_{jky|q_g})^2$.

Similarly, we have the first partial derivative with respect to d_{jk}

$$\frac{\partial \log M}{\partial d_{jk}} = \sum_{g}^{G}\sum_{y}^{3}\omega_{jky}(\frac{r_{gj,(k+1),y}}{P_{jy,(k+1)|q_g}} - \frac{r_{gjyk}}{P_{jyk|q_g}}) \tag{27}$$

where $\omega_{jyk} = P^*_{jyk|q_g} - (P^*_{jyk|q_g})^2$,

and the first partial derivative with respect to β_{jyk}, where y=(2,3), is

$$\frac{\partial \log M}{\partial \beta_{jyk}} = \sum_{g}^{G}\omega_{jyk}(\frac{r_{gjy,(k+1)}}{P_{jy,(k+1)|q_g}} - \frac{r_{gjyk}}{P_{jyk|q_g}}) \tag{28}$$

where $\omega_{jyk} = P^*_{jyk|q_g} - (P^*_{jyk|q_g})^2$.

The second partial derivatives in the Hessian matrix are given by

$$\begin{aligned}\frac{\partial^2 \log M}{\partial a_{jr}^2} &= \sum_{g=1}^{G}\sum_{y}^{3}\sum_{k=1}^{p} -\frac{r_{gjyk}q_{gr}^2(P^*_{jy,(k-1)|q_g}Q^*_{jy,(k-1)|q_g} - P^*_{jyk|q_g}Q^*_{jyk|q_g})^2}{P^2_{jyk|q_g}} \\ &= \sum_{g=1}^{G}\sum_{y}^{3}\sum_{k=1}^{p} -\frac{r_{gjyk}q_{gr}^2(\omega_{jy(k-1)} - \omega_{jyk})^2}{P^2_{jyk|q_g}}\end{aligned} \tag{29}$$

$$\begin{aligned}\frac{\partial^2 \log M}{\partial d_{jk}^2} &= \sum_{y}^{3}\sum_{g=1}^{G} -(\frac{r_{gjyk}}{P^2_{jyk|q_g}} + \frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}})P^{*2}_{jyk|q_g}(1 - P^*_{jyk|q_g})^2 \\ &= \sum_{y}^{3}\sum_{g=1}^{G} -(\frac{r_{gjyk}}{P^2_{jyk|q_g}} + \frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}})\omega^2_{jyk}\end{aligned} \tag{30}$$

$$\begin{aligned}\frac{\partial^2 \log M}{\partial d_{jk}\partial d_{j,k+1}} &= \sum_{g=1}^{G}\sum_{y}^{3}\frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}}(P^{*2}_{jyk|q_g}(1 - P^*_{jyk|q_g})^2)(P^{*2}_{jy(k+1)|q_g}(1 - P^*_{jy(k+1)|q_g})^2) \\ &= \sum_{g=1}^{G}\sum_{y}^{3}\frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}}\omega^2_{jyk}\omega^2_{jy(k+1)}\end{aligned} \tag{31}$$

and

$$\frac{\partial^2 \log M}{\partial a_{jr}\partial d_{jk}} = \sum_{g=1}^{G}\sum_{y}^{3} P^*_{jyk}Q^*_{jyk}q_{gr}[\frac{r_{gjyk}}{P^2_{jyk|q_g}}(P^*_{jy(k-1)|q_g}Q^*_{jy(k-1)|q_g} - P^*_{jyk|q_g}Q^*_{jyk|q_g}) + \frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}}(P^*_{jyk|q_g}Q^*_{jyk|q_g} - P^*_{jy(k+1)|q_g}Q^*_{jy(k+1)|q_g})] \quad (32)$$
$$= \sum_{g=1}^{G}\sum_{y}^{3} \omega_{jyk}q_{gr}[\frac{r_{gjyk}}{P^2_{jyk|q_g}}(\omega_{jy(k-1)} - \omega_{jyk}) + \frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}}(\omega_{jyk} - \omega_{jy(k+1)})]$$

where

$$Q^*_{jyk|q_g} = 1 - P^*_{jyk|q_g}.$$
$$\omega_{jyk} = P^*_{jyk|q_g} * Q^*_{jky|q_g}$$

$$\frac{\partial^2 \log M}{\partial \beta^2_{jyk}} = \frac{\partial^2 \log M}{\partial \beta_{jyk}\partial d_{jk}} = \sum_{g=1}^{G} -(\frac{r_{gjyk}}{P^2_{jyk|q_g}} + \frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}})P^{*2}_{jyk|q_g}(1 - P^*_{jyk|q_g})^2 \quad (33)$$

$$\frac{\partial^2 \log M}{\partial a_{jr}\partial \beta_{jyk}} = \sum_{g=1}^{G} \omega_{jyk}q_{gr}[\frac{r_{gjyk}}{P^2_{jyk|q_g}}(\omega_{jy(k-1)} - \omega_{jyk}) + \frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}}(\omega_{jyk} - \omega_{jy(k+1)})] \quad (34)$$

The expectation of the second partial derivatives in the Fisher scoring method are given by

$$E(\frac{\partial^2 \log M}{\partial a^2_{jr}}) = \sum_{g=1}^{G}\sum_{y}^{3}\sum_{k=1}^{p} -\frac{n_{gy}q^2_{gr}(\omega_{jy(k-1)} - \omega_{jyk})^2}{P_{jyk|q_g}}, \quad (35)$$

$$E(\frac{\partial^2 \log M}{\partial d^2_{jk}}) = \sum_{g=1}^{G}\sum_{y}^{3} -n_{gy}(\frac{1}{P_{jyk|q_g}} + \frac{1}{P_{jy(k+1)|q_g}})\omega^2_{jyk}, \quad (36)$$

$$E(\frac{\partial^2 \log M}{\partial d_{jk}\partial d_{j,k+1}}) = \sum_{g=1}^{G}\sum_{y}^{3} \frac{n_{gy}}{P_{jy(k+1)|q_g}}\omega^2_{jyk}\omega^2_{jy(k+1)}, \quad (37)$$

and

$$E(\frac{\partial^2 \log M}{\partial a_{jr}\partial d_{jk}}) = \sum_{g=1}^{G}\sum_{y}^{3} n_{gy}\omega_{jyk}q_{gr}[\frac{1}{P_{jyk|q_g}}(\omega_{jy(k-1)} - \omega_{jyk}) + \frac{1}{P_{jy(k+1)|q_g}}(\omega_{jyk} - \omega_{jy(k+1)})]. \quad (38)$$

$$E(\frac{\partial^2 \log M}{\partial \beta^2_{jyk}}) = E(\frac{\partial^2 \log M}{\partial \beta_{jyk}\partial d_{jk}}) = \sum_{g=1}^{G} -n_{gy}(\frac{1}{P_{jyk|q_g}} + \frac{1}{P_{jy(k+1)|q_g}})\omega^2_{jyk}. \quad (39)$$

$$E(\frac{\partial^2 \log M}{\partial a_{jr}\partial \beta_{jyk}}) = \sum_{g=1}^{G} n_{gy}\omega_{jyk}q_{gr}[\frac{1}{P_{jyk|q_g}}(\omega_{jy(k-1)} - \omega_{jyk}) + \frac{1}{P_{jy(k+1)|q_g}}(\omega_{jyk} - \omega_{jy(k+1)})]. \quad (40)$$

5.2.1 Cyclical coordinate descent

By Bazaraa, Sherali, and Shetty (2006), for a convex function f, a point $\bar{\theta}$ is a global minimizer of f if and only if $\partial f(\bar{\theta})$, the subgradient of f at $\bar{\theta}$, contains 0. Hence $\hat{\theta}_\tau$ is the global minimizer only when $\hat{\theta}_\tau = \text{sign}(s)(|s| - \tau)_+$, where $(u)_+ = u1(u > 0)$. This is called the soft-threshold of s and τ, and can be denoted by

$$\begin{aligned}\hat{\theta}_\tau &= \text{soft}(s, \tau) \equiv \text{sign}(s)(|s| - \tau)_+ \\ &= \arg\min_{\theta \in \mathbb{R}} \{0.5\theta^2 - s\theta + \tau|\theta|\}.\end{aligned} \tag{41}$$

Then, to minimize our objective function with respect to $\boldsymbol{\beta}$, we calculate a second-order Tylor approximation of the marginal log-likelihood at $\boldsymbol{\beta}^{(t-1)}$, and our lasso estimator in (13) can be updated by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \text{argmin}\{-\log M(\boldsymbol{\beta}) + \eta||\boldsymbol{\beta}||_1\} \\ &= \text{argmin}\{-\log M(\boldsymbol{\beta}^{(t-1)}) - \partial_\beta \log M(\boldsymbol{\beta}^{(t-1)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t-1)}) - \frac{\partial^2_\beta \log M(\boldsymbol{\beta}^{(t-1)})}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t-1)})^2 + \eta||\boldsymbol{\beta}||_1\} \\ &= -\frac{\text{soft}(\partial_{\boldsymbol{\beta}} \log M - \boldsymbol{\beta}_j^{(t-1)} * \partial^2_{\boldsymbol{\beta}} \log M, \eta)}{\partial^2_{\boldsymbol{\beta}} \log M}\end{aligned} \tag{42}$$

We run an EM and cyclical coordinate descent algorithm given by following.

Algorithm 1: Uniform DIF Detection via LASSO

Input : $\boldsymbol{A}_0$,$\boldsymbol{D}_0$,$\boldsymbol{\beta}_0$,$\boldsymbol{\mu}_0$,$\boldsymbol{\Sigma}_0$,U,η,ε_1,ε_2
Output: $\hat{\boldsymbol{A}}$,$\hat{\boldsymbol{D}}$,$\hat{\boldsymbol{\beta}}$,$\hat{\boldsymbol{\mu}}$,$\hat{\boldsymbol{\Sigma}}$

set $t_1 = 1$, $\delta^{(t_1-1)}$ = any value greater than ε_1
while $\delta_1^{(t_1-1)} > \varepsilon_1$ **do**
  Calculate n_{gy} and r_{gjyk}
  Update $\boldsymbol{\mu}^{(t_1)}$ and $\boldsymbol{\Sigma}^{(t_1)}$
  for *j=1,...m* **do**
    set $t_2 = 1$, $\delta^{(t_2-1)}$ = any value greater than ε_2
    while $\delta_2^{(t_2-1)} > \varepsilon_2$ **do**
      Calculate $P^*_{jyk|q_g}, Q^*_{jyk|q_g}$
      $a_{jr}^{(t_2)} = a_{jr}^{(t_2-1)} - \frac{\partial_{a_{jr}} \log M}{\partial^2_{a_{jr}} \log M}$
      $d_{jk}^{(t_2)} = d_{jk}^{(t_2-1)} - \frac{\partial_{d_{jk}} \log M}{\partial^2_{d_{jk}} \log M}$
      $\beta_{jyk}^{(t_2)} = -\frac{\text{soft}(\partial_{\beta_{jyk}} \log M - \beta_{jyk}^{(t_2-1)} * \partial^2_{\beta_{jyk}} \log M, \eta)}{\partial^2_{\beta_{jyk}} \log M}$
      $\delta_2^{(t_2)} = ||\boldsymbol{A}_j^{(t_2)} - \boldsymbol{A}_j^{(t_2-1)}|| + ||\boldsymbol{D}_j^{(t_2)} - \boldsymbol{D}_j^{(t_2-1)}|| + ||\boldsymbol{\beta}_j^{(t_2)} - \boldsymbol{\beta}_j^{(t_2-1)}||$
      $t_2 = t_2 + 1$
    end
    $a_{jr}^{(t_1)*} = a_{jr}^{(t_1)} * \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})}$
  end
  $\delta_1^{(t_1)} = ||\boldsymbol{A}^{(t_1)} - \boldsymbol{A}^{(t_1-1)}|| + ||\boldsymbol{D}^{(t_1)} - \boldsymbol{D}^{(t_1-1)}|| + ||\boldsymbol{\beta}^{(t_1)} - \boldsymbol{\beta}^{(t_1-1)}||$
  $t_1 = t_1 + 1$
end

Note that $\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})$ is the rth element on the diagonal of the estimated covariance matrix of the reference group $\hat{\boldsymbol{\Sigma}}_1$.

$\varepsilon_1 = 10^{-3}$ and $\varepsilon_2 = 10^{-7}$.

5.3 Tuning Parameter Selectiion

The Bayesian information criterion (BIC) is applied for tuning parameter selection. The BIC can be calculated by

$$BIC_{\hat{\mathbf{A}},\hat{\mathbf{D}},\hat{\boldsymbol{\beta}}} = -2 \max_{\hat{\mathbf{A}},\hat{\mathbf{D}},\hat{\boldsymbol{\beta}}} \log M + \|\hat{\boldsymbol{\beta}}\|_0 \log N, \tag{11}$$

where N is the sample size. The first term control the bias of the estimator, and the second term penalize the complexity of the model. We want to choose a tuning parameter η which gives us the smallest BIC value.

5.4 Simulation

5.4.1 Simulation 1

Sample Size. The total sample size is $N = 1500$, and the group sample sizes are $N_1 = N_2 = N_3 = 500$.

Test Length. $m = 20$. Simple structure. 10 items per dimension.

Proportion of DIF. 4 items with DIF. 2 DIF items per dimension.

Magnitude of DIF. The first focal group with larger difficulty parameters (+0.5) on the 4 DIF items.

The second focal group with much larger difficulty parameters (+1) on the 4 DIF items.

Generated parameters.

$a_{j1} \sim U(1.5, 2.5), j = 1, ..., 10$

$a_{j2} \sim U(1.5, 2.5), j = 11, ..., 20$

$d_1 \sim N(0, 1)$

$$\boldsymbol{A} = \begin{pmatrix} 2.17 & 0 \\ 0 & 2.46 \\ 2.41 & 0 \\ 2.45 & 0 \\ 2.34 & 0 \\ 1.84 & 0 \\ 1.85 & 0 \\ 1.92 & 0 \\ 1.94 & 0 \\ 1.90 & 0 \\ 1.92 & 0 \\ 0 & 2.43 \\ 0 & 1.82 \\ 0 & 2.22 \\ 0 & 1.93 \\ 0 & 1.88 \\ 0 & 1.84 \\ 0 & 2.12 \\ 0 & 2.42 \\ 0 & 2.15 \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} 0.03 \\ -1.28 \\ 0.58 \\ -2.06 \\ 0.12 \\ 3.25 \\ -0.41 \\ -0.51 \\ 0.89 \\ 1.33 \\ 0.85 \\ 0.82 \\ -0.37 \\ -0.99 \\ -0.27 \\ 0.19 \\ 1.73 \\ 0.05 \\ -1.86 \\ -0.63 \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

wABC

item	4	5	12	13
Focal 1	0.886	1.052	0.998	1.275
Focal 2	1.857	2.068	1.924	2.537

The first two items are used as anchor items in estimation.

No impact. $\theta_i \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.85 \\ 0.85 & 1 \end{pmatrix})$

Tuning parameters. Tuning parameters η are chosen from a sequence starting from 21 to 51 with increment 3. The range [21,51] was chosen because in all simulation studies I've run so far, the true model will be selected by the tuning parameter in this range for 100%.

Results of 50 Replications

Table 1. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.985	0.34	0.985
Type I	0.047	0.024	0.024

Omnibus DIF in regularization is defined as if at least one focal group showd DIF on an item, then that item is flagged as DIF. But in mirt package, omnibus DIf is testing H_0: the item has no DIF v.s. H_α : all focal groups have DIF.

Table 2. *Type I error and Power of mirt LRT* (0.05 *significance level*)

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.86	0.23	0.91
Type I	0.00714	0.0085	0.002857

mirt wald

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.2031	0.2031	0.9531
Type I	0.0045	0.0045	0.0134

Table 3. *Item parameter estimates by regularization*

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	-0.011145	-0.00684	-0.04356
RMSE	0.1676937	0.1614306	0.1576307

Table 4. *Item parameter estimates by mirt LRT* (0.05 *significance level*)

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	0.01262	0.00636	-0.00532
RMSE	0.17259	0.16686	0.15026

Our regularization method has slightly better non-DIF item parameter estimates.

Table 5. *Absolute bias for DIF magnitude recoveries for true DIF items*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Regularization (include false negative)	0.1529	0.3779	0.2284
mirt LRT (include false negative)	0.2222	0.4181	0.2181

Table 6. *Trait Estimation*

Item Parameters	θ_1	θ_2
Bias	0.0365	0.0374
RMSE	0.3823682	0.3806336
SE	0.3644813	0.3590251

The results in Table 6 are the average of estimated DIF for false positive items. LRT by mirt performs better when type I error happens. The type I error is low, so the probability to have these bias is low.

5.4.2 Simulation 2

Increase DIF proportion to 60% and keep everything else the same in simulation 1.

Sample Size. The total sample size is $N = 1500$, and the group sample sizes are $N_1 = N_2 = N_3 = 500$.

Proportion of DIF. 12 items with DIF. 6 DIF items per dimension.

$$\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

wABC

item	4	5	6	7	8	9	12	13	14	15	16	17
Focal 1	0.886	1.052	0.636	1.265	1.255	1.116	0.998	1.275	1.091	1.234	1.233	0.918
Focal 2	1.857	2.067	1.385	2.521	2.513	2.117	1.923	2.537	2.218	2.448	2.399	1.684

Table 7. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.872	0.136	0.872
Type I	0.135	0.085	0.053

Omnibus DIF is defined as if at least one focal group showd DIF on an item, then that item is flagged as DIF.

Table 8. *Type I error and Power of mirt LRT* (0.05 *significance level*)

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.963	0.395	0.987
Type I	0.0175	0.0125	0.0175

5.4.3 Simulation 3

Increase total sample size to 3000 and keep everything else the same in simulation 1.

Sample Size. The total sample size is $N = 3000$, and the group sample sizes are $N_1 = N_2 = N_3 = 1000$.

Proportion of DIF. 4 items with DIF. 2 DIF items per dimension.

Table 9. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	1	0.775	1
Type I	0.051	0.027	0.028

Omnibus DIF is defined as if at least one focal group showd DIF on an item, then that item is flagged as DIF.

Table 10. *Type I error and Power of mirt LRT* (0.05 *significance level*)

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.995	0.63	0.995
Type I	0.008	0.01	0.012

Table 11. *Item parameter estimates by regularization*

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	-0.01364	-0.00353	-0.0316
RMSE	0.11486	0.11917	0.111221

Table 12. *Item parameter estimates by mirt LRT* (0.05 *significance level*)

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	0.00612	0.01542	0.009
RMSE	0.11613	0.12292	0.0953

Table 13. *Absolute bias for DIF magnitude recoveries that were true DIF*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Regularization (include false negative)	0.1111	0.1884	0.1451
mirt LRT (include false negative)	0.1145	0.2557	0.1155

5.4.4 Simulation 4

Increase total sample size to 3000 and keep everything else the same in simulation 2.

Sample Size. The total sample size is $N = 3000$, and the group sample sizes are $N_1 = N_2 = N_3 = 1000$.

Proportion of DIF. 12 items with DIF. 6 DIF items per dimension.

Table 14. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.998	0.58	0.998
Type I	0.05	0.04	0.01

Table 15. *Type I error and Power of mirt LRT* (0.05 *significance level*)

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.996	0.772	1
Type I	0.02	0.02	0.0225

Table 16. *Item parameter estimates by regularization*

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	-0.0316	0.0244	0.0192
RMSE	0.118	0.127	0.131

Table 17. *Item parameter estimates by mirt LRT* (0.05 *significance level*)

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	0.0075	0.0197	0.0093
RMSE	0.1188	0.1269	0.1062

Table 18. *Absolute bias for DIF magnitude recoveries that were true DIF*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Regularization (include false negative)	0.216	0.2663	0.167
mirt LRT (include false negative)	0.1843	0.3556	0.1155

6 Non-uniform DIF Detection via LASSO

When the items have non-uniform DIF on slope only, i.e., there is no DIF on the intercepts, the DIF parameter we are estimating is $\boldsymbol{\Gamma} = (\boldsymbol{0}, ..., \boldsymbol{0}, \boldsymbol{\Gamma}_{q+1}, ..., \boldsymbol{\Gamma}_m)$.

6.1 E step

For the expectation step, we can use the result in Section 3.1 only replacing the DIF parameter $\boldsymbol{\beta}$ by $\boldsymbol{\Gamma}$, and minimize the following objective function

$$-\log M + \eta \sum_{j}^{m} ||\boldsymbol{\Gamma}_j||_1 \tag{43}$$

For each item, we minimize

$$-\log M_j + \eta ||\boldsymbol{\Gamma}_j||_1 \tag{44}$$

where η is the lasso tuning parameter.

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\Gamma}}) = \text{argmin}\{-\log M + \eta ||\boldsymbol{\Gamma}||_1\} \tag{45}$$

6.2 M step

Again, we assume the reference group has mean zero and variance one and only estimate its correlations. The means and all elements in the covariance matrices of two focal groups can be freely estimated.

$\hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\mu}}_3, \hat{\boldsymbol{\Sigma}}_1^*, \hat{\boldsymbol{\Sigma}}_2^*, \hat{\boldsymbol{\Sigma}}_3^*$ are same in (15), (16), (22), (23) and (24).

The first partial derivative with respect to a_{jr} and d_{jk} are same as in (25) and (26).

the first partial derivative with respect to γ_{jry}, where y=(2,3), is

$$\begin{aligned} \frac{\partial \log M}{\partial \gamma_{jyr}} &= \sum_{g}^{G} \sum_{k}^{p} \frac{r_{gjyk} q_{gr} [P^*_{jy(k-1)|q_g}(1 - P^*_{jy(k-1)|q_g}) - P^*_{jyk|q_g}(1 - P^*_{jyk|q_g})]}{P_{jyk|q_g}} \\ &= \sum_{g}^{G} \sum_{k}^{p} (\frac{r_{gjyk} q_{gr}}{P_{jyk|q_g}} (\omega_{jy(k-1)} - \omega_{jyk})) \end{aligned} \tag{46}$$

where $\omega_{jyk} = P^*_{jky|q_g} - (P^*_{jyk|q_g})^2$.

The second partial derivatives in the Hessian matrix $\frac{\partial^2 \log M}{\partial a_{jr}^2}$, $\frac{\partial^2 \log M}{\partial d_{jk}^2}$, $\frac{\partial^2 \log M}{\partial d_{jk} \partial d_{j,k+1}}$ and $\frac{\partial^2 \log M}{\partial a_{jr} \partial d_{jk}}$ are given by (28)-(31) and their expectations are (34)-(37).

$$\frac{\partial^2 \log M}{\partial \gamma_{jyr}^2} = \frac{\partial^2 \log M}{\partial \gamma_{jyr} \partial a_{jr}} = \sum_{g=1}^{G} \sum_{k=1}^{p} -\frac{r_{gjyk} q_{gr}^2 (\omega_{jy(k-1)} - \omega_{jyk})^2}{P^2_{jyk|q_g}} \tag{47}$$

$$\frac{\partial^2 \log M}{\partial \gamma_{jyr} \partial d_{jk}} = \sum_{g=1}^{G} \omega_{jky} q_{gr} [\frac{r_{gjyk}}{P^2_{jyk|q_g}} (\omega_{jy(k-1)} - \omega_{jyk}) + \frac{r_{gjy(k+1)}}{P^2_{jy(k+1)|q_g}} (\omega_{jyk} - \omega_{jy(k+1)})] \tag{48}$$

The expectation of the second partial derivatives in the Fisher scoring method are given by

$$E(\frac{\partial^2 \log M}{\partial \gamma_{jyr}^2}) = E(\frac{\partial^2 \log M}{\partial \gamma_{jyr} \partial a_{jr}}) = \sum_{g=1}^{G} \sum_{k=1}^{p} -\frac{n_{gy} q_{gr}^2 (\omega_{jy(k-1)} - \omega_{jyk})^2}{P_{jyk|q_g}}. \tag{49}$$

$$E(\frac{\partial^2 \log M}{\partial \gamma_{jyr} \partial d_{jk}}) = \sum_{g=1}^{G} n_{gy} \omega_{jyk} q_{gr} [\frac{1}{P_{jyk|q_g}} (\omega_{jy(k-1)} - \omega_{jyk}) + \frac{1}{P_{jy(k+1)|q_g}} (\omega_{jyk} - \omega_{jy(k+1)})]. \tag{50}$$

6.2.1 Cyclical coordinate descent

Same as in 3.2, to minimize our objective function with respect to $\mathbf{\Gamma}$, our lasso estimator in (13) can be written by

$$\begin{aligned}
\hat{\mathbf{\Gamma}} &= \text{argmin}\{-\log M(\mathbf{\Gamma}) + \eta ||\mathbf{\Gamma}||_1\} \\
&= \text{argmin}\{-\log M(\mathbf{\Gamma}_0) - \partial_\Gamma \log M(\mathbf{\Gamma}_0)(\mathbf{\Gamma} - \mathbf{\Gamma}_0) - \frac{\partial_\Gamma^2 \log M(\mathbf{\Gamma}_0)}{2} (\mathbf{\Gamma} - \mathbf{\Gamma}_0)^2 + \eta ||\mathbf{\Gamma}||_1\} \\
&= -\frac{\text{soft}(\partial_{\mathbf{\Gamma}} \log M - \mathbf{\Gamma}_j^{(t-1)} * \partial_{\mathbf{\Gamma}}^2 \log M, \eta)}{\partial_{\mathbf{\Gamma}}^2 \log M}
\end{aligned} \tag{51}$$

We run a cyclical coordinate descent algorithm for each group (item) with all other groups fixed. For item j, our algorithm is given by following.

1. Calculate $P^*_{jky|q_g}$ and $Q^*_{jky|q_g}$.
2. The parameter a_{jr} and d_{jk} can be updated by

$$a_{jr}^{(t)} = a_{jr}^{(t-1)} - \frac{\partial_{a_{jr}} \log M}{\partial_{a_{jr}}^2 \log M},$$

$$d_{jk}^{(t)} = d_{jk}^{(t-1)} - \frac{\partial_{d_{jk}} \log M}{\partial_{d_{jk}}^2 \log M}$$

and

$$\hat{\mathbf{\Gamma}}_{jyr} = -\frac{\text{soft}(\partial_{\mathbf{\Gamma}_{jyr}} \log M - \mathbf{\Gamma}_{jyr}^{(t-1)} * \partial_{\mathbf{\Gamma}_{jyr}}^2 \log M, \eta)}{\partial_{\mathbf{\Gamma}_{jyr}}^2 \log M}$$

Then we update $P^*_{jyk|q_g}$ and $Q^*_{jyk|q_g}$ by plugging in $\hat{\boldsymbol{A}}$,$\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{\beta}}$ from last coordinate descent cycle and repeat above steps until a convergence criterion is met.

After we get optimizers for item j, we do transforamtions on all estimates as following

$$a_{jr}^{(t)*} = a_{jr}^{(t)} * \sqrt{\text{diag}(\hat{\mathbf{\Sigma}}_{1r})},$$

$$\gamma_{jr}^{(t)*} = \gamma_{jr}^{(t)} * \sqrt{\text{diag}(\hat{\mathbf{\Sigma}}_{1r})},$$

We run an EM and cyclical coordinate descent algorithm given by following.

Algorithm 2: Non-uniform DIF Detection via LASSO

Input : $\boldsymbol{A}_0$,$\boldsymbol{D}_0$,$\boldsymbol{\Gamma}_0$,$\boldsymbol{\mu}_0$,$\boldsymbol{\Sigma}_0$,U,η,ε_1,ε_2

Output: $\hat{\boldsymbol{A}}$,$\hat{\boldsymbol{D}}$,$\hat{\boldsymbol{\Gamma}}$,$\hat{\boldsymbol{\mu}}$,$\hat{\boldsymbol{\Sigma}}$

set $t_1 = 1$, $\delta^{(t_1-1)} =$ any value greater than ε_1
while $\delta_1^{(t_1-1)} > \varepsilon_1$ **do**
  Calculate n_{gy} and r_{gjyk}
  Update $\boldsymbol{\mu}^{(t_1)}$ and $\boldsymbol{\Sigma}^{(t_1)}$
  for *j=1,...m* **do**
    set $t_2 = 1$, $\delta_2^{(t_2-1)} =$ any value greater than ε_2
    while $\delta_2^{(t_2-1)} > \varepsilon_2$ **do**
      Calculate $P^*_{jyk|q_g}, Q^*_{jyk|q_g}$
      $a_{jr}^{(t_2)} = a_{jr}^{(t_2-1)} - \frac{\partial_{a_{jr}} \log M}{\partial^2_{a_{jr}} \log M}$
      $d_{jk}^{(t_2)} = d_{jk}^{(t_2-1)} - \frac{\partial_{d_{jk}} \log M}{\partial^2_{d_{jk}} \log M}$
      $\Gamma_{jyr}^{(t_2)} = -\frac{\text{soft}(\partial_{\Gamma_{jyr}} \log M - \Gamma_{jyr}^{(t_2-1)} * \partial^2_{\boldsymbol{\Gamma}_{jyr}} \log M, \eta)}{\partial^2_{\Gamma_{jyr}} \log M}$
      $\delta_2^{(t_2)} = ||\boldsymbol{A}_j^{(t_2)} - \boldsymbol{A}_j^{(t_2-1)}|| + ||\boldsymbol{D}_j^{(t_2)} - \boldsymbol{D}_j^{(t_2-1)}|| + ||\boldsymbol{\Gamma}_j^{(t_2)} - \boldsymbol{\Gamma}_j^{(t_2-1)}||$
      $t_2 = t_2 + 1$
    end
    $a_{jr}^{(t_1)*} = a_{jr}^{(t_1)} * \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})}$
    $\Gamma_{jr}^{(t_1)*} = \Gamma_{jr}^{(t_1)} * \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})}$
  end
  $\delta_1^{(t_1)} = ||\boldsymbol{A}^{(t_1)} - \boldsymbol{A}^{(t_1-1)}|| + ||\boldsymbol{D}^{(t_1)} - \boldsymbol{D}^{(t_1-1)}|| + ||\boldsymbol{\Gamma}^{(t_1)} - \boldsymbol{\Gamma}^{(t_1-1)}||$
  $t_1 = t_1 + 1$
end

Note that $\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})$ is the rth element on the diagonal of the estimated covariance matrix of the reference group $\hat{\boldsymbol{\Sigma}}_1$.

$\varepsilon_1 = 10^{-3}$ and $\varepsilon_2 = 10^{-7}$.

6.3 Tuning Parameter Selectiion

The BIC can be calculated by

$$BIC_{\hat{\mathbf{A}},\hat{\mathbf{D}},\hat{\mathbf{\Gamma}}} = -2 \max_{\hat{\mathbf{A}},\hat{\mathbf{D}},\hat{\mathbf{\Gamma}}} \log M + \|\hat{\mathbf{\Gamma}}\|_0 \log N, \tag{11}$$

where N is the sample size.

6.4 Simulation

6.4.1 Simulation 5

Sample Size. The total sample size is $N = 1500$, and the group sample sizes are $N_1 = N_2 = N_3 = 500$.

Test Length. $m = 20$. Simple structure. 10 items per dimension.

Proportion of DIF. 4 items with DIF. 2 DIF items per dimension.

Magnitude of DIF. The first focal group with smaller discrimination parameter (-0.5) on the 4 DIF items. The second focal group with much smaller discrimination parameter (-1) on the 4 DIF items.

Generated parameters.

$a_{j1} \sim U(1.5, 2.5), j = 1, ..., 10$

$a_{j2} \sim U(1.5, 2.5), j = 11, ..., 20$

$d_1 \sim N(0, 1)$

$$\boldsymbol{A} = \begin{pmatrix} 2.17 & 0 \\ 0 & 2.46 \\ 2.41 & 0 \\ 2.45 & 0 \\ 2.34 & 0 \\ 1.84 & 0 \\ 1.85 & 0 \\ 1.92 & 0 \\ 1.94 & 0 \\ 1.90 & 0 \\ 1.92 & 0 \\ 0 & 2.43 \\ 0 & 1.82 \\ 0 & 2.22 \\ 0 & 1.93 \\ 0 & 1.88 \\ 0 & 1.84 \\ 0 & 2.12 \\ 0 & 2.42 \\ 0 & 2.15 \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} 0.03 \\ -1.28 \\ 0.58 \\ -2.06 \\ 0.12 \\ 3.25 \\ -0.41 \\ -0.51 \\ 0.89 \\ 1.33 \\ 0.85 \\ 0.82 \\ -0.37 \\ -0.99 \\ -0.27 \\ 0.19 \\ 1.73 \\ 0.05 \\ -1.86 \\ -0.63 \end{pmatrix},$$

$$\boldsymbol{\Gamma}_j = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, for\ j = 1, 2, 3, 6, 7, 8, 9, 10, 11, 14, 15, 16, 17, 18, 19, 20$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} -0.5 & 0 \\ -1 & 0 \end{pmatrix}, for\ j = 4, 5$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} 0 & -0.5 \\ 0 & -1 \end{pmatrix}, for\ j = 12, 13$$

wABC

item	4	5	12	13
Focal 1	0.620	0.530	0.560	0.774
Focal 2	1.331	1.242	1.292	1.813

The first two items are used as anchor items in estimation.

No impact. $\theta_i \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.85 \\ 0.85 & 1 \end{pmatrix})$

Tuning parameters. Tuning parameters η are chosen from a sequence starting from 21 to 51 with increment 3.

Results of 50 Replications

Table 19. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.625	0.14	0.625
Type I	0.0314	0.0171	0.0185

Table 20. *Type I error and Power of mirt LRT*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.515	0.375	0.646
Type I	0.0185	0.0238	0.0149

6.4.2 Simulation 6

Increase DIF proportion to 60%, and keep everything else the same as simulation 5.

Sample Size. The total sample size is $N = 1500$, and the group sample sizes are $N_1 = N_2 = N_3 = 500$.

Proportion of DIF. 12 items with DIF. 6 DIF items per dimension.

$$\mathbf{\Gamma}_j = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, for\ j = 1, 2, 3, 10, 11, 18, 19, 20$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} -0.5 & 0 \\ -1 & 0 \end{pmatrix}, for\ j = 4, 5, 6, 7, 8, 9$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} 0 & -0.5 \\ 0 & -1 \end{pmatrix}, for\ j = 12, 13, 14, 15, 16, 17$$

wABC

item	4	5	6	7	8	9	12	13	14	15	16	17
Focal 1	0.620	0.530	0.556	0.764	0.760	0.737	0.560	0.774	0.635	0.724	0.744	0.750
Focal 2	1.331	1.242	1.076	1.788	1.771	1.686	1.292	1.813	1.452	1.702	1.751	1.589

Table 21. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.45	0.06	0.443
Type I	0.0067	0.0033	0.0033

Adaptive LASSO (selected by GIC)

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.383	0.0086	0.3836
Type I	0	0	0

Table 22. *Type I error and Power of mirt LRT*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.656	0.185	0.746
Type I	0.0067	0.01	0.0233

6.4.3 Simulation 7

Increase total sample size to 3000, and keep everything else the same as simulation 5.

Sample Size. The total sample size is $N = 3000$, and the group sample sizes are $N_1 = N_2 = N_3 = 1000$.

Proportion of DIF. 4 items with DIF. 2 DIF items per dimension.

Table 23. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.96	0.345	0.96
Type I	0.0371	0.0271	0.0214

Table 24. *Type I error and Power of mirt LRT*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.815	0.555	0.885
Type I	0.005	0.031	0.0119

6.4.4 Simulation 8

Increase DIF proportion to 60%, and keep everything else the same as simulation 7.

Sample Size. The total sample size is $N = 3000$, and the group sample sizes are $N_1 = N_2 = N_3 = 1000$.

Proportion of DIF. 12 items with DIF. 6 DIF items per dimension.

$$\mathbf{\Gamma}_j = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, for\ j = 1, 2, 3, 10, 11, 18, 19, 20$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} -0.5 & 0 \\ -1 & 0 \end{pmatrix}, for\ j = 4, 5, 6, 7, 8, 9$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} 0 & -0.5 \\ 0 & -1 \end{pmatrix}, for\ j = 12, 13, 14, 15, 16, 17$$

Table 25. *Type I error and Power of regularization method*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.9933	0.481	0.9933
Type I	0.0333	0.02	0.02

Adaptive LASSO (selected by GIC)

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.866	0.112	0.866
Type I	0.0083	0.0083	0

Table 26. *Type I error and Power of mirt LRT*

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	0.9367	0.3	0.9683
Type I	0.0233	0.0133	0.0333

7 Non-uniform DIF Detection via Group LASSO

When the items have non-uniform DIF on both slope and intercept, the DIF parameter we are estimating are $\mathbf{\Gamma} = (\mathbf{0}, ..., \mathbf{0}, \mathbf{\Gamma}_{q+1}, ..., \mathbf{\Gamma}_m)$ and $\boldsymbol{\beta} = (\mathbf{0}, ..., \mathbf{0}, \boldsymbol{\beta}_{q+1}, ..., \boldsymbol{\beta}_m)$.

7.1 E step

The equations for expectations are same as before.

In our DIF detection problem, we minimize the following objective function

$$-\log M + \eta \sum_j^m ||(\mathbf{\Gamma}_j, \boldsymbol{\beta}_j)||_2 \tag{52}$$

For each item, we minimize

$$-\log M_j + \eta||(\boldsymbol{\Gamma}_j, \boldsymbol{\beta}_j)||_2 \tag{53}$$

where η is the group lasso tuning parameter.

We denote by $\boldsymbol{\tau} \in \mathbb{R}^{(y-1)*r+(y-1)*(m-1)}$ the whole DIF parameter vector, i.e. $\boldsymbol{\tau} = (\boldsymbol{\Gamma}, \boldsymbol{\beta})^T$.

Then, our objective function is

$$S_\eta(\boldsymbol{\tau}) = -\log M + \eta \sum_j^m ||\boldsymbol{\tau}||_2. \tag{54}$$

For each item j,

$$S_\eta(\boldsymbol{\tau}_j) = -\log M_j + \eta||\boldsymbol{\tau}_j||_2. \tag{55}$$

7.2 M step

The equations are same as in section 3.2. The Block co-ordinate gradient descent for solving the group lasso problem as follow.

7.2.1 Block co-ordinate gradient descent

Using a second-order Taylor series expansion at $\hat{\boldsymbol{\tau}}^{(t-1)}$ we define

$$M_\eta^{(t-1)}(\boldsymbol{\epsilon}^{(t)}) = -\{\log M + \boldsymbol{\epsilon}^{(t)T}\nabla \log M + \frac{1}{2}\boldsymbol{\epsilon}^{(t)T}H^{(t-1)}\boldsymbol{\epsilon}^{(t)}\} + \eta \sum_j^m ||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2,$$

where $\boldsymbol{\tau}^{(t)} = \boldsymbol{\tau}^{(t-1)} + \boldsymbol{\epsilon}^{(t)}$, and

$$\nabla \log M = (\frac{\partial \log M}{\partial \boldsymbol{\Gamma}}, \frac{\partial \log M}{\partial \boldsymbol{\beta}})$$

and

$$H^{(t-1)} = \begin{pmatrix} \frac{\partial^2 \log M}{\partial \boldsymbol{\Gamma}^2} & \frac{\partial^2 \log M}{\partial \boldsymbol{\Gamma} \partial \boldsymbol{\beta}} \\ \frac{\partial^2 \log M}{\partial \boldsymbol{\Gamma} \partial \boldsymbol{\beta}} & \frac{\partial^2 \log M}{\partial \boldsymbol{\beta}^2} \end{pmatrix}.$$

We have $M_\eta^{(t-1)}(\boldsymbol{\epsilon}) \approx S_\eta(\hat{\boldsymbol{\tau}}^{(t-1)} + \boldsymbol{\epsilon}^{(t)})$.

We run a block co-ordinate gradient descent algorithm for each group (item) with all other groups fixed.

For item j, denote u to be the subgradient of $||\boldsymbol{\tau}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2$. We have

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

The subgradient equation $\partial_{\epsilon_j} M_\eta^{(t-1)}(\boldsymbol{\epsilon}^{(t)}) = -\nabla \log M_j - \boldsymbol{\epsilon}_j^{(t)T} H_j^{(t-1)} + \eta u = 0$ is satisfied with $\boldsymbol{\tau}_j^{(t-1)} + \boldsymbol{\epsilon}_j = 0$ if

$$||u||_2 = ||\frac{\nabla \log M_j + \boldsymbol{\epsilon}_j^{(t)T} H_j^{(t-1)}}{\eta}||_2 \leq 1$$

$$||\nabla \log M_j + \boldsymbol{\epsilon}_j^{(t)T} H_j^{(t-1)}||_2 \leq \eta$$

$$||\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)}||_2 \leq \eta,$$

the minimizer of $M_\eta^{(t-1)}(\boldsymbol{\epsilon})$ is

$$\hat{\boldsymbol{\epsilon}}_j^{(t)} = -\hat{\boldsymbol{\tau}}_j^{(t-1)}.$$

Otherwise,

Then the subgradient equation is

$$\partial_{\epsilon_j} M_\eta^{(t-1)}(\boldsymbol{\epsilon}^{(t)}) = -\nabla \log M_j - \boldsymbol{\epsilon}_j^{(t)T} H_j^{(t-1)} + \eta \frac{\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}}{||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2} = 0 \tag{56}$$

$$\partial_{\epsilon_j} M_\eta^{(t-1)}(\boldsymbol{\epsilon}^{(t)}) = -\nabla \log M_j - (\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}) H_j^{(t-1)} + \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)} + \eta \frac{\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}}{||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2} = 0$$

$$\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)} = -(\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}) H_j^{(t-1)} + \eta \frac{\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}}{||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2} \tag{57}$$

$$\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)} = \frac{(\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)})||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2}{\eta - H_j^{(t-1)}||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2} \tag{58}$$

Taking the norm of both sides of (57) we see that

$$||\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)}||_2 = (\frac{\eta}{||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2} - H_j^{(t-1)})||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2$$

$$||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2 = \frac{\eta - ||\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)}||_2}{H_j^{(t-1)}} \tag{59}$$

Plugging (59) into (58), we have

$$\boldsymbol{\epsilon}_j^{(t)} = -(H_j^{(t-1)})^{-1}\{\nabla \log M_j - \eta \frac{\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)}}{||\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_j^{(t-1)}||_2}\}. \tag{60}$$

$$\nabla \log M_j = (\frac{\partial \log M}{\partial \gamma_{j11}}, ..., \frac{\partial \log M}{\partial \gamma_{jyr}}, ..., \frac{\partial \log M}{\partial \gamma_{j3q}}, \frac{\partial \log M}{\partial \beta_{j11}}, ..., \frac{\partial \log M}{\partial \beta_{jyk}}, ... \frac{\partial \log M}{\partial \beta_{j3(p-1)}}), r = 1, ..., q; k = 1, ..., p-1; y = 2, 3.$$

If $\boldsymbol{\epsilon}_j^{(t)} \neq 0$, performing a Backtracking-Armijo line search: let $\alpha^{(t)}$ be the largest value in $\{\alpha_0 \delta^l\}_{l \geq 0}$ s.t.

$$S_\eta(\hat{\boldsymbol{\tau}}_j^{(t-1)} + \alpha^{(t)}\boldsymbol{\epsilon}_j^{(t)}) - S_\eta(\hat{\boldsymbol{\tau}}_j^{(t-1)}) \leq \alpha^{(t)}\sigma\Delta^{(t)}, \tag{61}$$

where $\alpha_0 = 1$, $\delta = 0.5$ and $\sigma = 0.1$, and Δ is the improvement in the objective function $S_\eta(\cdot)$ when using a linear approximation for the log-likelihood, i.e.

$$\Delta_j^{(t)} = -\boldsymbol{\epsilon}_j^{(t)T}\nabla \log M_j + \eta||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j^{(t)}||_2 - \eta||\hat{\boldsymbol{\tau}}_j^{(t-1)}||_2. \tag{62}$$

$$\hat{\boldsymbol{\tau}}_j^{(t)} = (\hat{\boldsymbol{\Gamma}}_j^{(t)}, \hat{\boldsymbol{\beta}}_j^{(t)}) = \hat{\boldsymbol{\tau}}_j^{(t-1)} + \alpha^{(t)}\boldsymbol{\epsilon}_j^{(t)}$$

Then we update $P^*_{jky|q_g}$ and $Q^*_{jky|q_g}$ by plugging in $\hat{\boldsymbol{A}}$,$\hat{\boldsymbol{D}}$, $\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\beta}}$ from last coordinate descent cycle and repeat above steps until a convergence criterion is met.

After we get optimizers for item j, we do transforamtions on all estimates as following

$$a_{jr}^{(t)*} = a_{jr}^{(t)} * \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})},$$

$$\gamma_{jr}^{(t)*} = \gamma_{jr}^{(t)} * \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})},$$

where μ_{1r} is the rth element of the estimated mean vector of the reference group $\hat{\boldsymbol{\mu}}_1$, and $\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})$ is the rth element on the diagonal of the estimated covariance matrix of the reference group $\hat{\boldsymbol{\Sigma}}_1$.

We run an EM and block coordinate gradient descent algorithm given by following.

Algorithm 3: Non-uniform DIF Detection via group LASSO

Input : $\boldsymbol{A}_0$,$\boldsymbol{D}_0$,$\boldsymbol{\Gamma}_0$,$\boldsymbol{\beta}_0$,$\boldsymbol{\mu}_0$,$\boldsymbol{\Sigma}_0$,U,η,ε_1,ε_2
Output: $\hat{\boldsymbol{A}}$,$\hat{\boldsymbol{D}}$,$\hat{\boldsymbol{\Gamma}}$,$\boldsymbol{\beta}$,$\hat{\boldsymbol{\mu}}$,$\hat{\boldsymbol{\Sigma}}$

set $t_1 = 1$, $\delta^{(t_1-1)} =$ any value greater than ε_1
while $\delta_1^{(t_1-1)} > \varepsilon_1$ **do**
  Calculate n_{gy} and r_{gjyk}
  Update $\boldsymbol{\mu}^{(t_1)}$ and $\boldsymbol{\Sigma}^{(t_1)}$
  for *j=1,...m* **do**
    set $t_2 = 1$, $\delta_2^{(t_2-1)} =$ any value greater than ε_2
    while $\delta_2^{(t_2-1)} > \varepsilon_2$ **do**
      Calculate $P^*_{jyk|q_g}, Q^*_{jyk|q_g}$
      $a_{jr}^{(t_2)} = a_{jr}^{(t_2-1)} - \frac{\partial_{a_{jr}} \log M}{\partial^2_{a_{jr}} \log M}$
      $d_{jk}^{(t_2)} = d_{jk}^{(t_2-1)} - \frac{\partial_{d_{jk}} \log M}{\partial^2_{d_{jk}} \log M}$
      if $||\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t_2-1)} H_j^{(t_2-1)}||_2 \leq \eta$ **then**
        $\boldsymbol{\tau}_j = \mathbf{0}$
      else
        $\boldsymbol{\epsilon}_j^{(t_2)} = -(H_j^{(t_2-1)})^{-1}\{\nabla \log M_j - \eta \frac{\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t_2-1)} H_j^{(t_2-1)}}{||\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t_2-1)} H_j^{(t_2-1)}||_2}\}$
        $\Delta_j^{(t_2)} = -\boldsymbol{\epsilon}_j^{(t_2)T} \nabla \log M_j + \eta||\hat{\boldsymbol{\tau}}_j^{(t_2-1)} + \boldsymbol{\epsilon}_j^{(t_2)}||_2 - \eta||\hat{\boldsymbol{\tau}}_j^{(t_2-1)}||_2$
        $\alpha^{(t_2)}$ is the max value in $\{\alpha^{(0)}\delta^l\}_{l\geq 0}$ such that
        $S_\eta(\hat{\boldsymbol{\tau}}_j^{(t_2-1)} + \alpha^{(t_2)}\boldsymbol{\epsilon}_j^{(t_2)}) - S_\eta(\hat{\boldsymbol{\tau}}_j^{(t_2-1)}) \leq \alpha^{(t_2)}\sigma\Delta^{(t_2)}$.
        $\boldsymbol{\tau}_j^{(t_2)} = \boldsymbol{\tau}_j^{(t_2-1)} + \alpha^{(t_2)}\boldsymbol{\epsilon}_j^{(t_2)}$
      end
      $\delta_2^{(t_2)} = ||\boldsymbol{A}_j^{(t_2)} - \boldsymbol{A}_j^{(t_2-1)}|| + ||\boldsymbol{D}_j^{(t_2)} - \boldsymbol{D}_j^{(t_2-1)}|| + ||\alpha^{(t_2-1)}\boldsymbol{\epsilon}_j^{(t_2-1)}||$
      $t_2 = t_2 + 1$
    end
    $a_{jr}^{(t_1)*} = a_{jr}^{(t_1)} * \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})}$
    $\Gamma_{jr}^{(t_1)*} = \Gamma_{jr}^{(t_1)} * \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})}$
  end
  $\delta_1^{(t_1)} = ||\boldsymbol{A}^{(t_1)} - \boldsymbol{A}^{(t_1-1)}|| + ||\boldsymbol{D}^{(t_1)} - \boldsymbol{D}^{(t_1-1)}|| + ||\boldsymbol{\Gamma}^{(t_1)} - \boldsymbol{\Gamma}^{(t_1-1)}|| + ||\boldsymbol{\beta}^{(t_1)} - \boldsymbol{\beta}^{(t_1-1)}||$
  $t_1 = t_1 + 1$
end

Note that $\text{diag}(\hat{\boldsymbol{\Sigma}}_{1r})$ is the rth element on the diagonal of the estimated covariance matrix of the reference group $\hat{\boldsymbol{\Sigma}}_1$.

$\varepsilon_1 = 10^{-3}$ and $\varepsilon_2 = 10^{-7}$.

7.3 Tuning Parameter Selection

The Bayesian information criterion (BIC) in Tutz et al.(2015) is applied for tuning parameter selection. The degrees of freedom of penalized parameters $\boldsymbol{\tau} = (\boldsymbol{\Gamma}, \boldsymbol{\beta})^T$ are approximated by

$$\tilde{df}_{\boldsymbol{\tau}}(\eta) = \sum_{j}^{m} I(||\boldsymbol{\tau}_j(\eta)||_2 > 0) + \sum_{j}^{m} \frac{||\boldsymbol{\tau}_j(\eta)||_2}{||\boldsymbol{\tau}_j^{ML}||_2}(m' - 1)$$

where m' is the number of DIF parameters for each item. Here we have $m' = 4$.

The total degrees of freedom are

$$df(\eta) = m + N + \tilde{df}_{\boldsymbol{\tau}}(\eta) - 1.$$

The BIC can be calculated by

$$BIC(\eta) = -2 \max_{\hat{\mathbf{A}}, \hat{\mathbf{D}}, \hat{\boldsymbol{\tau}}} \log M + df(\eta) \log(m \cdot N).$$

where N is the sample size. The first term control the bias of the estimator, and the second term penalize the complexity of the model. We want to choose a tuning parameter η which gives us the smallest BIC value.

7.4 Simulation

7.4.1 Simulation 9

Sample Size. The total sample size is $N = 1500$, and the group sample sizes are $N_1 = N_2 = N_3 = 500$.

Test Length. $m = 20$. Simple structure. 10 items per dimension.

Proportion of DIF. 4 items with DIF. 2 DIF items per dimension.

Magnitude of DIF. The first focal group with smaller discrimination parameter (-0.5) and larger difficulty parameters (+0.5) on the 4 DIF items.

The second focal group with much smaller difficulty parameter (-1) and much larger difficulty parameters (+1) on the 4 DIF items.

Generated parameters.

$a_{j1} \sim U(1.5, 2.5), j = 1, ..., 10$

$a_{j2} \sim U(1.5, 2.5), j = 11, ..., 20$

$d_1 \sim N(0, 1)$

$$\boldsymbol{A} = \begin{pmatrix} 2.17 & 0 \\ 0 & 2.46 \\ 2.41 & 0 \\ 2.45 & 0 \\ 2.34 & 0 \\ 1.84 & 0 \\ 1.85 & 0 \\ 1.92 & 0 \\ 1.94 & 0 \\ 1.90 & 0 \\ 1.92 & 0 \\ 0 & 2.43 \\ 0 & 1.82 \\ 0 & 2.22 \\ 0 & 1.93 \\ 0 & 1.88 \\ 0 & 1.84 \\ 0 & 2.12 \\ 0 & 2.42 \\ 0 & 2.15 \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} 0.03 \\ -1.28 \\ 0.58 \\ -2.06 \\ 0.12 \\ 3.25 \\ -0.41 \\ -0.51 \\ 0.89 \\ 1.33 \\ 0.85 \\ 0.82 \\ -0.37 \\ -0.99 \\ -0.27 \\ 0.19 \\ 1.73 \\ 0.05 \\ -1.86 \\ -0.63 \end{pmatrix},$$

$$\boldsymbol{\Gamma}_j = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, for\ j = 1, 2, 3, 6, 7, 8, 9, 10, 11, 14, 15, 16, 17, 18, 19, 20$$

$$\boldsymbol{\Gamma}_j = \begin{pmatrix} -0.3 & 0 \\ -0.6 & 0 \end{pmatrix}, for\ j = 4, 5$$

$$\boldsymbol{\Gamma}_j = \begin{pmatrix} 0 & -0.3 \\ 0 & -0.6 \end{pmatrix}, for\ j = 12, 13$$

$$\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

The first two items are used as anchor items in estimation.

No impact. $\theta_i \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.85 \\ 0.85 & 1 \end{pmatrix})$

Tuning parameters. Tuning parameters η are chosen from a sequence starting from 21 to 51 with increment 3.

Results of 50 Replications

Table 27. *Type I error and Power of regularization method*

Group	Omnibus DIF
Power	0.93
Type I	0.0457

Table 28. *Type I error and Power of mirt LRT*

Group	Omnibus DIF
Power	0.925
Type I	0.0142

7.4.2 Simulation 10

Keep everything the same as simulation 9. Increase the DIF proportion to 60%. *Sample Size*. The total sample size is $N = 1500$, and the group sample sizes are $N_1 = N_2 = N_3 = 500$.

Proportion of DIF. 12 items with DIF. 6 DIF items per dimension.

$$\boldsymbol{\Gamma}_j = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, for\ j = 1, 2, 3, 10, 11, 18, 19, 20$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} -0.3 & 0 \\ -0.6 & 0 \end{pmatrix}, for\ j = 4, 5, 6, 7, 8, 9$$

$$\mathbf{\Gamma}_j = \begin{pmatrix} 0 & -0.3 \\ 0 & -0.6 \end{pmatrix}, for\ j = 12, 13, 14, 15, 16, 17$$

$$\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

Results of 50 Replications

Table 29. *Type I error and Power of regularization method*

Group	Omnibus DIF
Power	0.842
Type I	0.0033

Table 30. *Type I error and Power of mirt LRT*

Group	Omnibus DIF
Power	0.926
Type I	0.023

7.4.3 Simulation 11

Keep everything the same as simulation 9. Increase the sample size. *Sample Size*. The total sample size is $N = 3000$, and the group sample sizes are $N_1 = N_2 = N_3 = 1000$.

Proportion of DIF. 4 items with DIF. 2 DIF items per dimension.

Results of 50 Replications

Table 31. *Type I error and Power of regularization method*

Group	Omnibus DIF
Power	1
Type I	0.0557

Table 32. *Type I error and Power of mirt LRT*

Group	Omnibus DIF
Power	0.985
Type I	0.008

7.4.4 Simulation 12

Keep everything the same as simulation 10. Increase the sample size to 3000. *Sample Size*. The total sample size is $N = 3000$, and the group sample sizes are $N_1 = N_2 = N_3 = 1000$.

Proportion of DIF. 12 items with DIF. 6 DIF items per dimension.

Results of 50 Replications

Table 33. *Type I error and Power of regularization method*

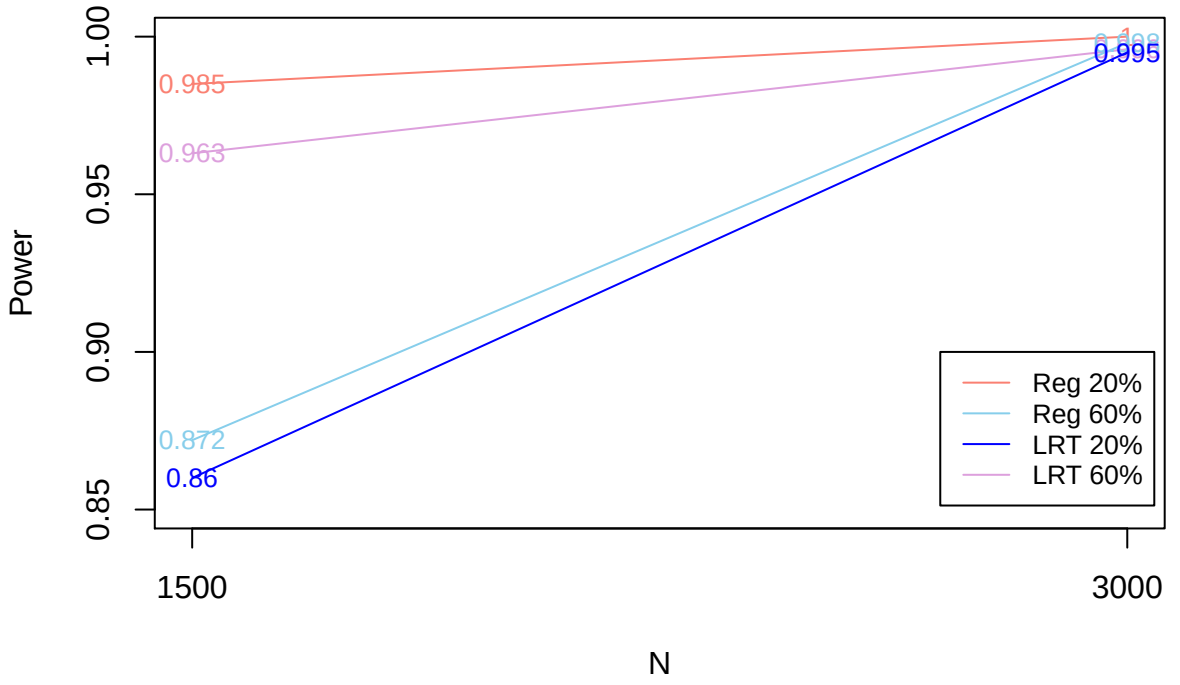
Group	Omnibus DIF
Power	0.941
Type I	0.0459

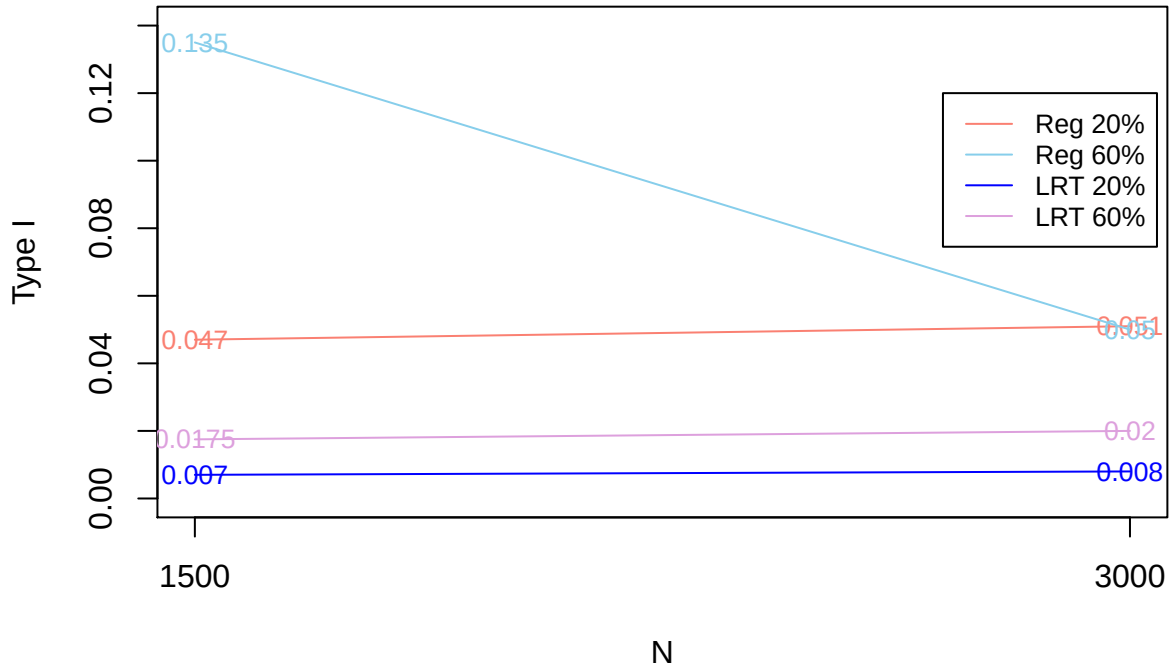
Table 34. *Type I error and Power of mirt LRT*

Group	Omnibus DIF
Power	0.985
Type I	0.023

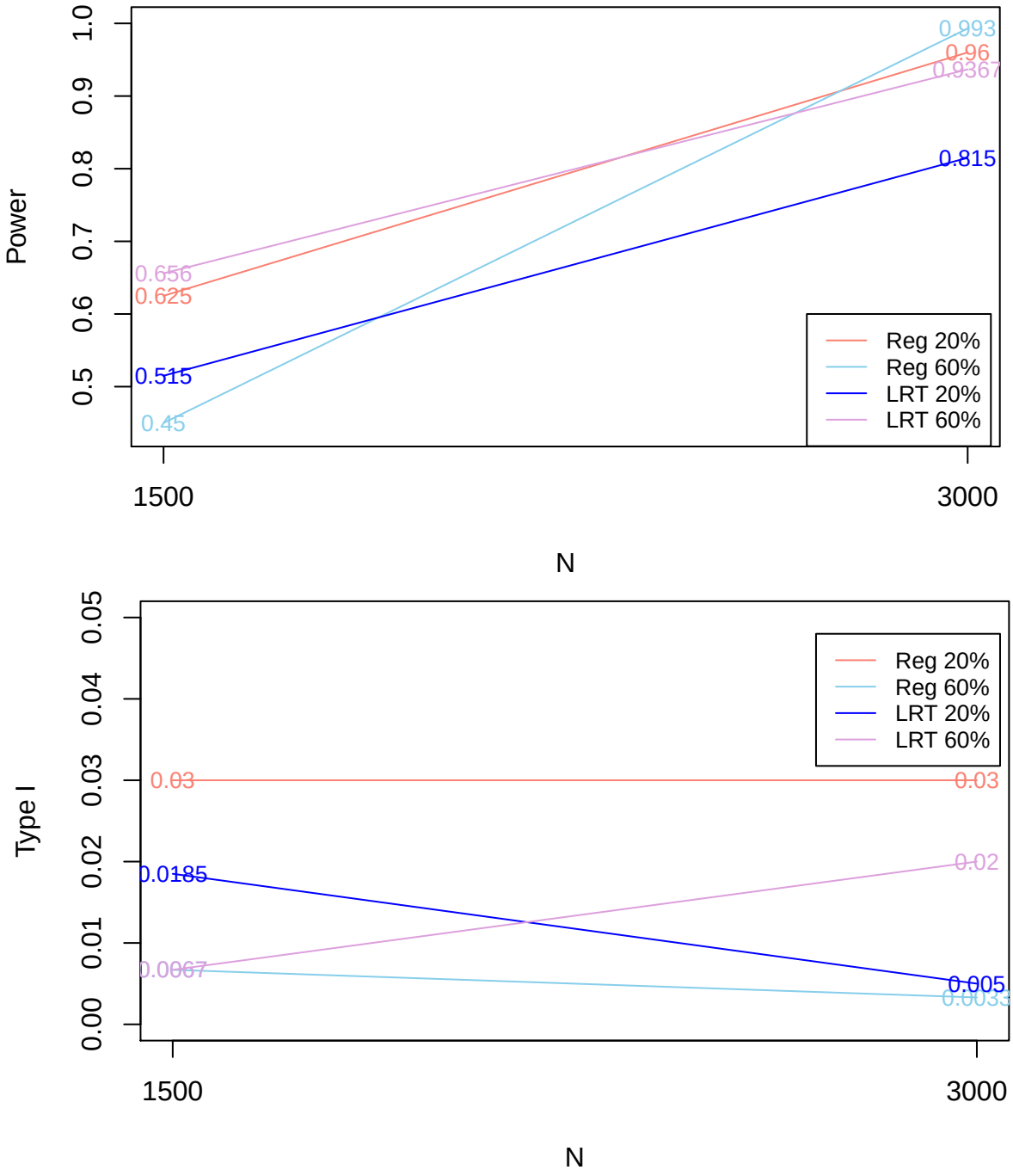
8 Plots of Results

8.1 DIF only on intercept

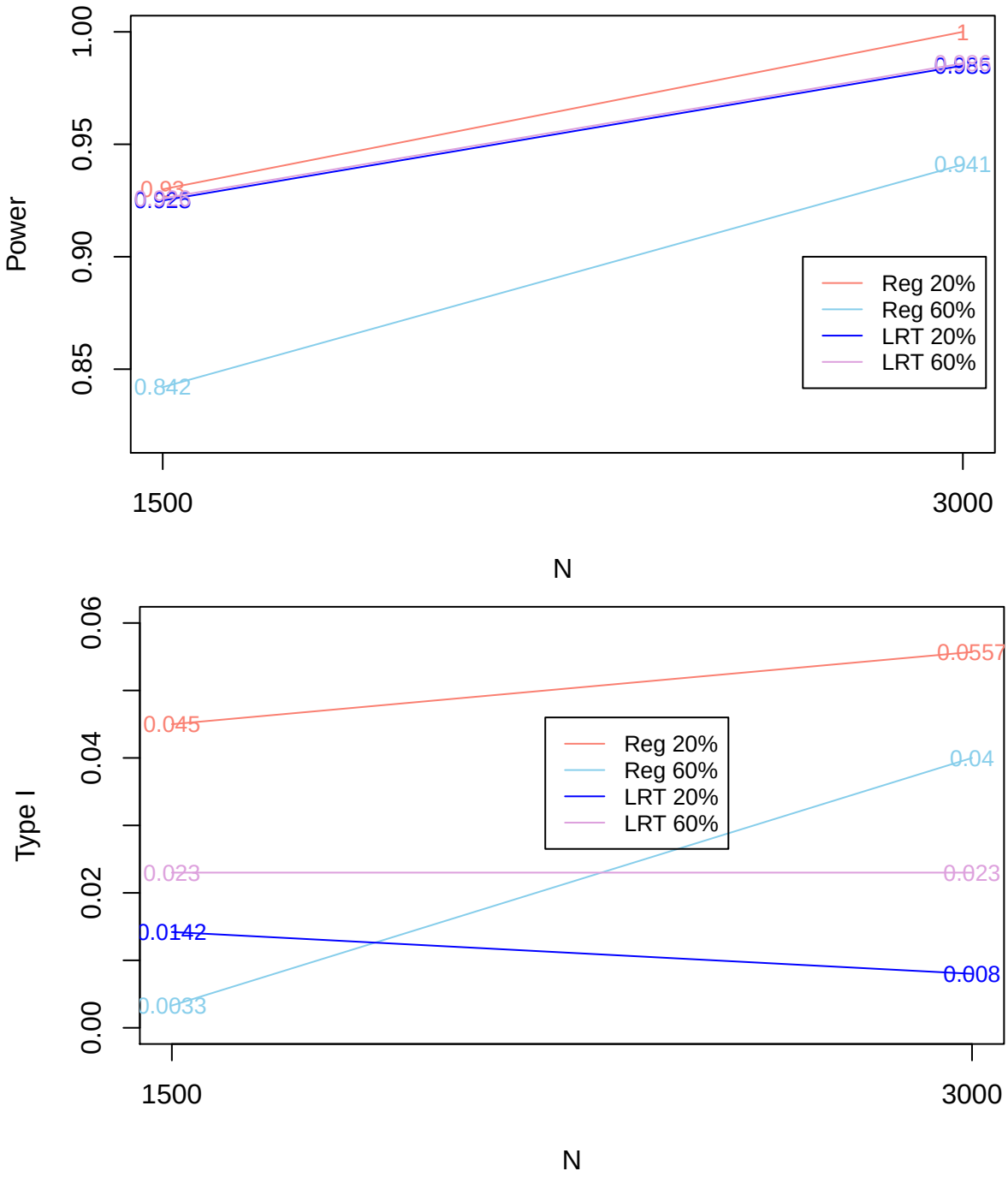




8.2 DIF only on slope



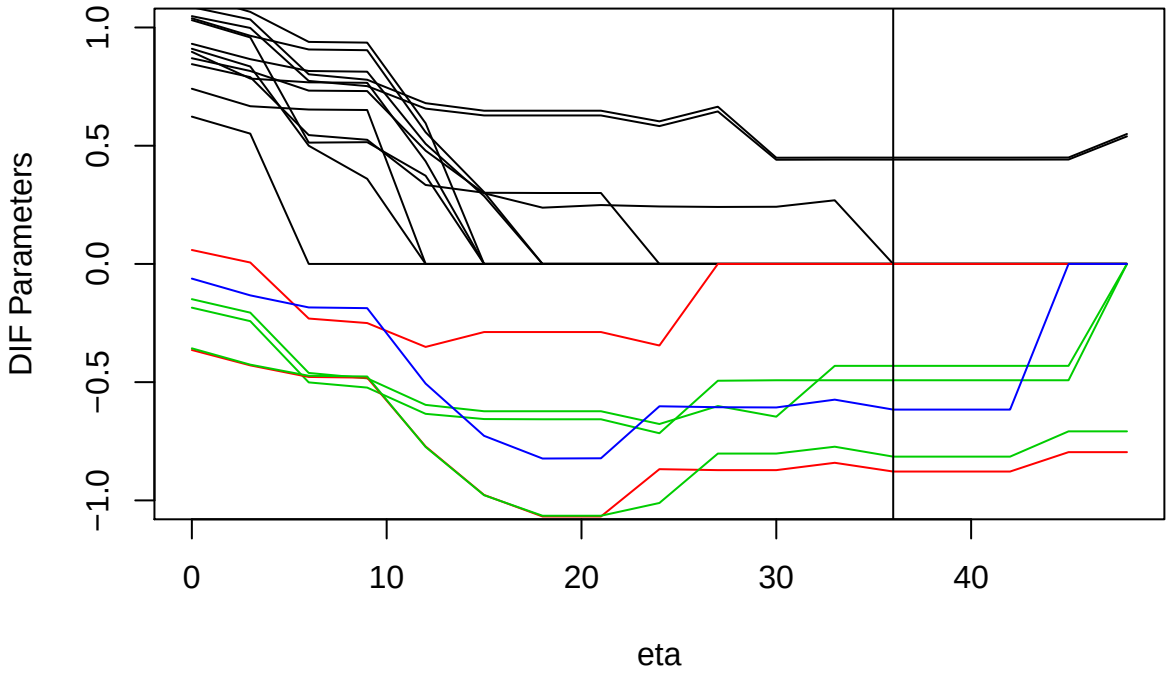
8.3 DIF on slope and intercept



9 High DIF Proportion Issue

When DIF proportion is higher than 50%, we found either our original LASSO or group LASSO algorithm work. Both of two methods would give us "reverse" type solutions– the true DIF parameters would be estimated as zero and non-DIF parameters would be non-zero. Below is a plot of the solution path of LASSO algorithm in 60% DIF on intercept condition. At the selected tuning parameter ($\eta = 36$), most DIF items

(black lines) are shrunk to zero and non-DIF items (colored lines) are non-zero.



To solve this issue, we consider following methods:

1. "EMM" algorithm

From the above solution path plot, we can see when the tuning parameter is zero, the parameter estimations are good (DIF parameters are large and non-DIF parameters are close to zero). But the bias of LASSO estimates increases as tuning parameter increasing. So we considered adding a re-estimation step, that is, perform another M-step without penalty term after each EM cycle. The results of simulation 2, 4, 6, 8, 10 and 12 in this document are estimated by EMM algorithm and they look acceptable.

2. Using a different constraint

In section 4, we mentioned using anchor items for model identification. Those anchor items have no DIF on all focal groups. Here we can use a different constraint by setting "group-specific" anchors. Suppose we have two focal groups. Two anchor items can be used and each of them has DIF on one focal group but has no DIF on the other focal group.

To be more specific, recall

$$\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j1\cdot}, \boldsymbol{\beta}_{j2\cdot})^T = \begin{pmatrix} \beta_{j11} & \beta_{j12} & \dots & \beta_{j1k} & \dots & \beta_{j1,p-1} \\ \beta_{j21} & \beta_{j22} & \dots & \beta_{j2k} & \dots & \beta_{j2,p-1} \end{pmatrix} (k = 1, ..., p-1),$$

where p is the number of categories in GRM, and each row of $\boldsymbol{\beta}_j$ is (uniform) DIF parameter for a focal group, i.e. $\boldsymbol{\beta}_{j1\cdot}$ is (uniform) DIF parameter for the first focal group, and $\boldsymbol{\beta}_{j2\cdot}$ is (uniform) DIF parameter for the second focal group.

If the first two items are anchors, we want to set $\beta_{j11} = 0$ and $\beta_{j22} = 0$, and β_{j21} and β_{j12} be non-zero and their magnitudes can be freely estimated. This constraint strongly relies on a priori knowledge of the items and the entire study but it works well when the DIF proportion is high.

3. Adaptive LASSO

Pick a $\lambda > 0$, and define the weight vector

$$\hat{w} = 1/|\hat{\boldsymbol{\beta}}|^{\lambda}$$

($\hat{\boldsymbol{\beta}}$ can be the MLE).

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\beta}}) = \text{argmin}\{-\log M + \eta_n \sum_{j}^{m} \hat{w}_j ||\boldsymbol{\beta}_j||_1\} \tag{63}$$

Works.

4. SCAD penalty (Fan & Li, 2001)