DIF Detection via Regularization

Ruoyi Zhu

1/6/2020

Contents

1 Graded Response Model with DIF

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

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

$\boldsymbol{y}_i$ 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.

$$P_{ijk} = P^*_{ij,k-1} - P^*_{ijk} \tag{2}$$

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

Assume a total sample size N and test length m. 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

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

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

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 examinee i is in the first focal group, then y=2,

$$\boldsymbol{\gamma}_{jky} = \boldsymbol{\gamma}_{j1\cdot},$$

and

$$\boldsymbol{\beta}_{jky} = \boldsymbol{\beta}_{j1\cdot}.$$

If an examinee i is in the second focal group, then y=3,

$$\boldsymbol{\gamma}_{jky} = \boldsymbol{\gamma}_{j2\cdot},$$

and

$$\boldsymbol{\beta}_{jky} = \boldsymbol{\beta}_{j2\cdot}.$$

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

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 $1,\ldots,N_1$, 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 $N_1+1,\ldots,N_1+N_2$, 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 $N_1+N_2+1,\ldots,N_1+N_2+N_3$, 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.

G quadrature samples (same for all examinees) are denoted by $\boldsymbol{q}_g (g = 1, ..., G)$, and $\boldsymbol{q}_g = (q_{g1}, q_{g2}, \ldots, q_{gr}, \ldots, q_{gq})^T$ (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

$$P^*_{ijk|q_g} = P^*_{jky|q_g} = P^*_{jk1|q_g} = \frac{1}{1 + e^{-(\mathbf{a}_j \boldsymbol{q}_g + d_k)}}$$
$$(i = 1, ..., N_1; j = 1, 2, ..., m; k = 1, 2, ..., p-1; g = 1, ..., G).$$

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

$$P^*_{ijk|q_g} = P^*_{jky|q_g} = P^*_{jk2|q_g} = \frac{1}{1 + e^{-(\mathbf{a}_j \boldsymbol{q}_g + d_k + \boldsymbol{\gamma}_{j1.} \boldsymbol{q}_g + \boldsymbol{\beta}_{j1k})}}$$

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

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

$$P^*_{ijk|q_g} = P^*_{jky|q_g} = P^*_{jk3|q_g} = \frac{1}{1 + e^{-(\mathbf{a}_j \boldsymbol{q}_g + d_k + \boldsymbol{\gamma}_{j2\cdot} \boldsymbol{q}_g + \boldsymbol{\beta}_{j2k})}}.$$

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

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

2 Model Identifiability Constraint

Yet the model is not identified. Some constraints on the item parameters are required. 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.

3 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 $\boldsymbol{\beta} = (\mathbf{0}, \mathbf{0}, \boldsymbol{\beta}_{q+1}, ..., \boldsymbol{\beta}_m)$.

3.1 E step

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

$$L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\theta}_i \mid \mathbf{y}, \boldsymbol{u}_i) = \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}}. \tag{3}$$

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

$$\begin{aligned} &L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) = L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\theta}_i \mid \mathbf{y}, \mathbf{u}_i) f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i) \\ &= \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}} (2\pi)^{-p/2} |\boldsymbol{\Sigma}_y|^{-1/2} \exp(-0.5(\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)' \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)). \end{aligned} \tag{4}$$

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

$$m(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i) = \int L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i) \partial \boldsymbol{\theta}_i \tag{5}$$

Then

$$h(\boldsymbol{\theta}_i \mid \boldsymbol{u}_i, \boldsymbol{y}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \Sigma_y^{(t-1)}) = \frac{L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i)}{m(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i)} \tag{6}$$

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 $\boldsymbol{\theta}$

$$\begin{aligned} &E[log\{L(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{\Theta})\} | \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{Y}, \mathbf{U}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}] \\ &= \sum_{i}^{N} \{ \int \log L(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{u}_i, \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}) \partial \boldsymbol{\theta}_i \\ &+ \int \log f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \boldsymbol{\Sigma}_y^{(t-1)}) \partial \boldsymbol{\theta}_i \} \end{aligned} \tag{7}$$

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

$$\begin{aligned} &E[log L(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\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{8}$$

Then we can define two artificial terms.

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

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

and

$$r_{jgky} = r_{jgk1} = \sum_{i=1}^{N_1} x_{ijk} \frac{L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_1^{(t-1)}, \boldsymbol{\Sigma}_1^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_1^{(t-1)}, \boldsymbol{\Sigma}_1^{(t-1)})}.$$

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

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

and

$$r_{jgky} = r_{jgk2} = \sum_{i=N_1+1}^{N_1+N_2} x_{ijk} \frac{L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_2^{(t-1)}, \boldsymbol{\Sigma}_2^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_2^{(t-1)}, \boldsymbol{\Sigma}_2^{(t-1)})}.$$

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

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

and

$$r_{jgky} = r_{jgk3} = \sum_{i=N_1+N_2+1}^{N_1+N_2+N3} x_{ijk} \frac{L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_3^{(t-1)}, \boldsymbol{\Sigma}_3^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_3^{(t-1)}, \boldsymbol{\Sigma}_3^{(t-1)})}.$$

$n_g = n_{g1} + n_{g2} + n_{g3}$ 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 $\boldsymbol{q}_g$.

$$E[log\{L(\mathbf{A}, \mathbf{D}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{U}, \boldsymbol{\Theta})] = \sum_y^3 \sum_g^G \sum_j^m \sum_k^p (r_{jgky} \mathrm{log} P_{jky|q_g}) + \sum_y^3 \sum_g^G n_g \mathrm{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{9}$$

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

$$\log M_j = \sum_y^3 \sum_g^G \sum_k^p (r_{jgky} \mathrm{log} P_{jky|q_g}) + \sum_y^3 \sum_g^G n_g \mathrm{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{10}$$

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

$$-\log M + \eta \sum_{j}^{m} ||\boldsymbol{\beta}_j||_1 \tag{11}$$

For each item, we minimize

$$-\log M_j + \eta ||\boldsymbol{\beta}_j||_1 \tag{12}$$

where η is the lasso tuning parameter.

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\beta}}) = \mathrm{argmin}\{-\log M + \eta ||\boldsymbol{\beta}||_1\} \tag{13}$$

3.2 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

$$\begin{aligned} \frac{\partial \log M}{\partial \boldsymbol{\mu}_y} &= \sum_{g}^{G} n_{gy} \frac{\partial \mathrm{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g)}{\partial \boldsymbol{\mu}_y} \\ &= \sum_{g}^{G} n_{gy} \frac{\partial - \frac{1}{2}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{q}_g - \boldsymbol{\mu}_y)}{\partial \boldsymbol{\mu}_y} \\ &= \sum_{g}^{G} n_{gy} (\boldsymbol{q}_g - \boldsymbol{\mu}_y) \boldsymbol{\Sigma}_y^{-1} \end{aligned} \tag{14}$$

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

$$\hat{\boldsymbol{\mu}}_2 = \frac{\sum_{g=1}^{G} n_{g2} \boldsymbol{q}_g}{N_2}, \tag{15}$$

and

$$\hat{\boldsymbol{\mu}}_3 = \frac{\sum_{g=1}^{G} n_{g3} \boldsymbol{q}_g}{N_3}. \tag{16}$$

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

$$
\begin{aligned}
\frac{\partial \log M}{\partial \boldsymbol{\Sigma}_y} &= \sum_g^G n_{gy} \frac{\partial \mathrm{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g)}{\partial \boldsymbol{\Sigma}_y} \\
&= \sum_g^G n_{gy} \frac{\partial(-\frac{q}{2}\mathrm{log}(2\pi) - \frac{1}{2}\mathrm{log}|\boldsymbol{\Sigma}_y| - \frac{1}{2}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)^T \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{q}_g - \boldsymbol{\mu}_y))}{\partial \boldsymbol{\Sigma}_y} \\
&= \sum_g^G n_{gy}[-\frac{1}{2}\Sigma_y^{-1} + \frac{1}{2}\Sigma_y^{-1}(\boldsymbol{q}_g - \boldsymbol{\mu}_y)(\boldsymbol{q}_g - \boldsymbol{\mu}_y)^T \Sigma_y^{-1}]
\end{aligned}
\tag{17}
$$

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

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\sum_{g=1}^G n_{g1} \boldsymbol{q}_g \boldsymbol{q}_g'}{N_1}, \tag{18}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \frac{\sum_{g=1}^G n_{g2} (\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{19}$$

and

$$\hat{\boldsymbol{\Sigma}}_3 = \frac{\sum_{g=1}^G n_{g3} (\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{20}$$

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

$$\boldsymbol{q}_g^* = \frac{q_g}{\sqrt{\mathrm{diag}\hat{\boldsymbol{\Sigma}}_1}}. \tag{21}$$

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

$$\hat{\boldsymbol{\Sigma}}_1^* = \frac{\sum_{g=1}^G n_{g1} \boldsymbol{q}_g^* \boldsymbol{q}_g^{*\prime}}{N_1}, \tag{22}$$

$$\hat{\boldsymbol{\Sigma}}_2^* = \frac{\sum_{g=1}^G n_{g2} (\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{23}$$

and

$$\hat{\boldsymbol{\Sigma}}_3^* = \frac{\sum_{g=1}^G n_{g3} (\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{24}$$

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

$$\frac{\partial \log M}{\partial a_{jr}} = \sum_y^3 \sum_{k=1}^p \sum_{g=1}^G (\frac{r_{jgky} q_{gr}}{P_{jky|q_g}} (\omega_{j,(k-1),y} - \omega_{jky})) \tag{25}$$

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

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

$$\frac{\partial \log M}{\partial d_{jk}} = \sum_{y}^{3}\sum_{g}^{G} \omega_{jky}(\frac{r_{jg,(k+1),y}}{P_{j,(k+1),y|q_g}} - \frac{r_{jgky}}{P_{jky|q_g}}) \tag{26}$$

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

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

$$\frac{\partial \log M}{\partial \beta_{jky}} = \sum_{g}^{G} \omega_{jky}(\frac{r_{jg(k+1)y}}{P_{j(k+1)y|q_g}} - \frac{r_{jgky}}{P_{jky|q_g}}) \tag{27}$$

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

The second partial derivatives in the Hessian matrix are given by

$$\begin{aligned}\frac{\partial^2 \log M}{\partial a_{jr}^2} &= \sum_{y}^{3}\sum_{k=1}^{p}\sum_{g=1}^{G} -\frac{r_{jgky} q_{gr}^2 (P^*_{j(k-1)y|q_g} Q^*_{j(k-1)y|q_g} - P^*_{jky|q_g} Q^*_{jky|q_g})^2}{P^2_{jky|q_g}} \\ &= \sum_{y}^{3}\sum_{k=1}^{p}\sum_{g=1}^{G} -\frac{r_{jgky} q_{gr}^2 (\omega_{j(k-1)y} - \omega_{jky})}{P^2_{jky|q_g}}\end{aligned} \tag{28}$$

$$\begin{aligned}\frac{\partial^2 \log M}{\partial d_{jk}^2} &= \sum_{y}^{3}\sum_{g=1}^{G} -(\frac{r_{jgky}}{P^2_{jky|q_g}} + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}}) P^{*2}_{jky|q_g} (1 - P^*_{jky|q_g})^2 \\ &= \sum_{y}^{3}\sum_{g=1}^{G} -(\frac{r_{jgky}}{P^2_{jky|q_g}} + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}}) \omega^2_{jky}\end{aligned} \tag{29}$$

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

and

$$\begin{aligned}\frac{\partial^2 \log M}{\partial a_{jr} \partial d_{jk}} &= \sum_{y}^{3}\sum_{g=1}^{G} P^*_{jky} Q^*_{jky} q_{gr} [\frac{r_{jgky}}{P^2_{jky|q_g}} (P^*_{j(k-1)y|q_g} Q^*_{j(k-1)y|q_g} - P^*_{jky|q_g} Q^*_{jky|q_g}) \\ &+ \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} (P^*_{jky|q_g} Q^*_{jky|q_g} - P^*_{j(k+1)y|q_g} Q^*_{j(k+1)y|q_g})] \\ &= \sum_{y}^{3}\sum_{g=1}^{G} \omega_{jky} q_{gr} [\frac{r_{jgky}}{P^2_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})]\end{aligned} \tag{31}$$

where

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

$$\omega_{jky} = P^*_{jky|q_g} * Q^*_{jky|q_g}$$

$$\frac{\partial^2 \log M}{\partial \beta_{jky}^2} = \frac{\partial^2 \log M}{\partial \beta_{jky} \partial d_{jk}} = \sum_{g=1}^{G} -(\frac{r_{jgky}}{P_{jky|q_g}^2} + \frac{r_{jg(k+1)y}}{P_{j(k+1)y|q_g}^2}) P_{jky|q_g}^{*2} (1 - P_{jky|q_g}^{*})^2 \tag{32}$$

$$\frac{\partial^2 \log M}{\partial a_{jr} \partial \beta_{jky}} = \sum_{g=1}^{G} \omega_{jky} q_{gr} [\frac{r_{jgky}}{P_{jky|q_g}^2} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{r_{jg(k+1)y}}{P_{j(k+1)y|q_g}^2} (\omega_{jky} - \omega_{j(k+1)y})] \tag{33}$$

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

$$E(\frac{\partial^2 \log M}{\partial a_{jr}^2}) = \sum_{y}^{3} \sum_{k=1}^{p} \sum_{g=1}^{G} - \frac{n_{gy} q_{gr}^2 (\omega_{j(k-1)y} - \omega_{jky})}{P_{jky|q_g}}, \tag{34}$$

$$E(\frac{\partial^2 \log M}{\partial d_{jk}^2}) = \sum_{y}^{3} \sum_{g=1}^{G} -n_{gy} (\frac{1}{P_{jky|q_g}} + \frac{1}{P_{j(k+1)y|q_g}}) \omega_{jky}^2, \tag{35}$$

$$E(\frac{\partial^2 \log M}{\partial d_{jk} \partial d_{j,k+1}}) = \sum_{y}^{3} \sum_{g=1}^{G} \frac{n_{gy}}{P_{j(k+1)y|q_g}} \omega_{jky}^2 \omega_{j(k+1)y}^2, \tag{36}$$

and

$$E(\frac{\partial^2 \log M}{\partial a_{jr} \partial d_{jk}}) = \sum_{y}^{3} \sum_{g=1}^{G} n_{gy} \omega_{jky} q_{gr} [\frac{1}{P_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{1}{P_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})]. \tag{37}$$

$$E(\frac{\partial^2 \log M}{\partial \beta_{jky}^2}) = E(\frac{\partial^2 \log M}{\partial \beta_{jky} \partial d_{jk}}) = \sum_{g=1}^{G} -n_{gy} (\frac{1}{P_{jky|q_g}} + \frac{1}{P_{j(k+1)y|q_g}}) \omega_{jky}^2. \tag{38}$$

$$E(\frac{\partial^2 \log M}{\partial a_{jr} \partial \beta_{jky}}) = \sum_{g=1}^{G} n_{gy} \omega_{jky} q_{gr} [\frac{1}{P_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{1}{P_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})]. \tag{39}$$

3.2.1 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

$$\begin{aligned} \hat{\theta}_\tau &= \text{soft}(s, \tau) \equiv \text{sign}(s)(|s| - \tau)_+ \\ &= \arg\min_{\theta \in \mathbb{R}} \{0.5\theta^2 - s\theta + \tau|\theta|\}. \end{aligned} \tag{40}$$

Then, to minimize our objective function with respect to $\boldsymbol{\beta}$, our lasso estimator in (13) can be written by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \text{argmin}\{-\log M(\boldsymbol{\beta}) + \eta||\boldsymbol{\beta}||_1\} \\ &= \text{argmin}\{-\log M(\boldsymbol{\beta}_0) - \partial_\beta \log M(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{\partial_\beta^2 \log M(\boldsymbol{\beta}_0)}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^2 + \eta||\boldsymbol{\beta}||_1\} \\ &= -\frac{\text{soft}(\partial_{\boldsymbol{\beta}} \log M - \boldsymbol{\beta}_j^{(t-1)} * \partial_{\boldsymbol{\beta}}^2 \log M, \eta)}{\partial_{\boldsymbol{\beta}}^2 \log M} \end{aligned} \tag{41}$$

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

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

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

and

$$\hat{\boldsymbol{\beta}}_{jky} = -\frac{\text{soft}(\partial_{\boldsymbol{\beta}_{jky}} \log M - \boldsymbol{\beta}^{(t-1)}_{jky} * \partial^2_{\boldsymbol{\beta}_{jky}} \log M, \eta)}{\partial^2_{\boldsymbol{\beta}_{jky}} \log M}$$

Then we update $P^*_{jky|q_g}$ and $Q^*_{jky|q_g}$ 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

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

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

3.3 Simulation

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 extreme difficulty parameter (+0.5) on the 4 DIF items.

The second focal group with more extreme difficulty parameter (+1) on the 4 DIF items.

Generated parameters.

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

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

$d_1 \sim N(0, 1)$

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

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

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

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

3.3.1 Results of 16 Replications

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

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	1	0.5	1
Type I	0.049	0.01875	0.022

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

Table 2. *Type I error and Power of mirt LRT*

Group	Omnibus DIF
Power	1
Type I	0.053

mirt LRT can only do the omnibus DIF test.

Both regularization and mirt LRT can detect DIF maginitude 1 with power 100%. Our regularization method has slightly lower Type I error.

Table 3. *Item parameter estimates by regularization*

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	-0.0160795	-0.0099595	0.02939
RMSE	0.141	0.145	0.146

Table 4. Item parameter estimates by mirt LRT

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	0.002604	0.0118656	-0.040605
RMSE	0.14481	0.1484	0.15307

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

Table 5. 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)	0.255	0.302	0.208
Regularization (exclude false negative)	0.173	0.104	0.208
mirt LRT (include false negative)	0.1746	0.1666	0.1826

Table 6. Absolute bias for DIF magnitude recoveries that were $non-DIF$

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Regularization	0.336	0.33	0.341
mirt LRT	0.15696	0.1516	0.1622

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.

4 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 $\mathbf{\Gamma} = (\mathbf{0}, ..., \mathbf{0}, \mathbf{\Gamma}_{q+1}, ..., \mathbf{\Gamma}_m)$.

4.1 E step

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

$$L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\theta}_i \mid \mathbf{y}, \boldsymbol{u}_i) = \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}}. \tag{42}$$

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

$$\begin{aligned} L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) &= L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\theta}_i \mid \mathbf{y}, \mathbf{u}_i) f(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{\theta}_i) \\ &= \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}} (2\pi)^{-p/2} |\mathbf{\Sigma}_y|^{-1/2} \exp(-0.5(\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)' \mathbf{\Sigma}_y^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)). \end{aligned} \tag{43}$$

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

$$m(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i) = \int L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{\theta}_i) \partial \boldsymbol{\theta}_i \tag{44}$$

Then

$$h(\boldsymbol{\theta}_i \mid \boldsymbol{u}_i, \boldsymbol{y}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \Sigma_y^{(t-1)}) = \frac{L(\mathbf{A}, \mathbf{D}, \boldsymbol{\Gamma} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i)}{m(\mathbf{A}, \mathbf{D}, \boldsymbol{\Gamma}, \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i)} \tag{45}$$

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

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

$$\begin{aligned} &E[log\{L(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\Gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{\Theta})\} | \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \mathbf{Y}, \mathbf{U}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}] \\ &= \sum_i^N \{ \int \log L(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\Gamma} \mid \boldsymbol{y}, \boldsymbol{u}_i, \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}) \partial \boldsymbol{\theta}_i \\ &+ \int \log f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \boldsymbol{\Sigma}_y^{(t-1)}) \partial \boldsymbol{\theta}_i \} \end{aligned} \tag{46}$$

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

$$\begin{aligned} &E[logL(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\Gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{U})] \\ &= \sum_i^N \sum_g^G \log L(\boldsymbol{A}, \boldsymbol{D}, \boldsymbol{\Gamma} \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{\Gamma}^{(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{\Gamma}^{(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{\Gamma}^{(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{\Gamma}^{(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{\Gamma}^{(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{\Gamma}^{(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{\Gamma}^{(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{\Gamma}^{(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{47}$$

Then we can define two artificial terms.

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

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

and

$$r_{jgky} = r_{jgk1} = \sum_{i=1}^{N_1} x_{ijk} \frac{L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_1^{(t-1)}, \boldsymbol{\Sigma}_1^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_1^{(t-1)}, \boldsymbol{\Sigma}_1^{(t-1)})}.$$

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

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

and

$$r_{jgky} = r_{jgk2} = \sum_{i=N_1+1}^{N_1+N_2} x_{ijk} \frac{L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_2^{(t-1)}, \boldsymbol{\Sigma}_2^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_2^{(t-1)}, \boldsymbol{\Sigma}_2^{(t-1)})}.$$

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

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

and

$$r_{jgky} = r_{jgk3} = \sum_{i=N_1+N_2+1}^{N_1+N_2+N3} x_{ijk} \frac{L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_3^{(t-1)}, \boldsymbol{\Sigma}_3^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g \mid \mathbf{y}_i, \boldsymbol{u}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \boldsymbol{\Gamma}^{(t-1)}, \boldsymbol{\mu}_3^{(t-1)}, \boldsymbol{\Sigma}_3^{(t-1)})}.$$

$n_g = n_{g1} + n_{g2} + n_{g3}$ 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 $\boldsymbol{q}_g$.

$$E[log\{L(\mathbf{A}, \mathbf{D}, \boldsymbol{\Gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{U}, \boldsymbol{\Theta})] = \sum_y^3 \sum_g^G \sum_j^m \sum_k^p (r_{jgky} \text{log} P_{jky|q_g}) + \sum_y^3 \sum_g^G n_g \text{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{48}$$

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

$$\log M_j = \sum_y^3 \sum_g^G \sum_k^p (r_{jgky} \text{log} P_{jky|q_g}) + \sum_y^3 \sum_g^G n_g \text{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{49}$$

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

$$-\log M + \eta \sum_j^m ||\boldsymbol{\Gamma}_j||_1 \tag{50}$$

For each item, we minimize

$$-\log M_j + \eta ||\boldsymbol{\Gamma}_j||_1 \tag{51}$$

where η is the lasso tuning parameter.

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\Gamma}}) = \text{argmin}\{-\log M + \eta ||\boldsymbol{\Gamma}||_1\} \tag{52}$$

4.2 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.

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

$$\hat{\boldsymbol{\mu}}_2 = \frac{\sum_{g=1}^{G} n_{g2}\boldsymbol{q}_g}{N_2}, \tag{53}$$

and

$$\hat{\boldsymbol{\mu}}_3 = \frac{\sum_{g=1}^{G} n_{g3}\boldsymbol{q}_g}{N_3}. \tag{54}$$

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

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\sum_{g=1}^{G} n_{g1}\boldsymbol{q}_g\boldsymbol{q}_g'}{N_1}, \tag{55}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \frac{\sum_{g=1}^{G} n_{g2}(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{56}$$

and

$$\hat{\boldsymbol{\Sigma}}_3 = \frac{\sum_{g=1}^{G} n_{g3}(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{57}$$

Standardized quadrature points.

$$\boldsymbol{q}_g^* = \frac{q_g}{\sqrt{\text{diag}\hat{\boldsymbol{\Sigma}}_1}}. \tag{58}$$

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

$$\hat{\boldsymbol{\Sigma}}_1^* = \frac{\sum_{g=1}^{G} n_{g1}\boldsymbol{q}_g^*\boldsymbol{q}_g^{*'}}{N_1}, \tag{59}$$

$$\hat{\boldsymbol{\Sigma}}_2^* = \frac{\sum_{g=1}^{G} n_{g2}(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{60}$$

and

$$\hat{\boldsymbol{\Sigma}}_3^* = \frac{\sum_{g=1}^{G} n_{g3}(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{61}$$

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

$$\frac{\partial \log M}{\partial a_{jr}} = \sum_{y}^{3}\sum_{k=1}^{p}\sum_{g=1}^{G}\left(\frac{r_{jgky}q_{gr}}{P_{jky|q_g}}(\omega_{j,(k-1),y} - \omega_{jky})\right) \tag{62}$$

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

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

$$\frac{\partial \log M}{\partial d_{jk}} = \sum_{y}^{3} \sum_{g}^{G} \omega_{jky} \left(\frac{r_{jg,(k+1),y}}{P_{j,(k+1),y|q_g}} - \frac{r_{jgky}}{P_{jky|q_g}} \right) \tag{63}$$

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

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

$$\begin{aligned} \frac{\partial \log M}{\partial \gamma_{jry}} &= \sum_{g}^{G} \sum_{k}^{p} \frac{r_{jgky} q_{gr} [P^*_{j(k-1)y|q_g}(1 - P^*_{j(k-1)y|q_g}) - P^*_{jky|q_g}(1 - P^*_{jky|q_g})]}{P_{jky|q_g}} \\ &= \sum_{k}^{p} \sum_{g}^{G} \left(\frac{r_{jgky} q_{gr}}{P_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) \right) \end{aligned} \tag{64}$$

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

The second partial derivatives in the Hessian matrix are given by

$$\begin{aligned} \frac{\partial^2 \log M}{\partial a^2_{jr}} &= \sum_{y}^{3} \sum_{k=1}^{p} \sum_{g=1}^{G} - \frac{r_{jgky} q^2_{gr} (P^*_{j(k-1)y|q_g} Q^*_{j(k-1)y|q_g} - P^*_{jky|q_g} Q^*_{jky|q_g})^2}{P^2_{jky|q_g}} \\ &= \sum_{y}^{3} \sum_{k=1}^{p} \sum_{g=1}^{G} - \frac{r_{jgky} q^2_{gr} (\omega_{j(k-1)y} - \omega_{jky})}{P^2_{jky|q_g}} \end{aligned} \tag{65}$$

$$\begin{aligned} \frac{\partial^2 \log M}{\partial d^2_{jk}} &= \sum_{y}^{3} \sum_{g=1}^{G} - \left(\frac{r_{jgky}}{P^2_{jky|q_g}} + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} \right) P^{*2}_{jky|q_g} (1 - P^*_{jky|q_g})^2 \\ &= \sum_{y}^{3} \sum_{g=1}^{G} - \left(\frac{r_{jgky}}{P^2_{jky|q_g}} + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} \right) \omega^2_{jky} \end{aligned} \tag{66}$$

$$\begin{aligned} \frac{\partial^2 \log M}{\partial d_{jk} \partial d_{j,k+1}} &= \sum_{y}^{3} \sum_{g=1}^{G} \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} (P^{*2}_{jky|q_g} (1 - P^*_{jky|q_g})^2)(P^{*2}_{j(k+1)y|q_g} (1 - P^*_{j(k+1)y|q_g})^2) \\ &= \sum_{y}^{3} \sum_{g=1}^{G} \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} \omega^2_{jky} \omega^2_{j(k+1)y} \end{aligned} \tag{67}$$

and

$$\begin{aligned} \frac{\partial^2 \log M}{\partial a_{jr} \partial d_{jk}} &= \sum_{y}^{3} \sum_{g=1}^{G} P^*_{jky} Q^*_{jky} q_{gr} \Big[\frac{r_{jgky}}{P^2_{jky|q_g}} (P^*_{j(k-1)y|q_g} Q^*_{j(k-1)y|q_g} - P^*_{jky|q_g} Q^*_{jky|q_g}) \\ &\quad + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} (P^*_{jky|q_g} Q^*_{jky|q_g} - P^*_{j(k+1)y|q_g} Q^*_{j(k+1)y|q_g}) \Big] \\ &= \sum_{y}^{3} \sum_{g=1}^{G} \omega_{jky} q_{gr} \Big[\frac{r_{jgky}}{P^2_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y}) \Big] \end{aligned} \tag{68}$$

where

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

$$\frac{\partial^2 \log M}{\partial \gamma_{jry}^2} = \frac{\partial^2 \log M}{\partial \gamma_{jry} \partial a_{jr}} = \sum_{k=1}^{p} \sum_{g=1}^{G} - \frac{r_{jgky} q_{gr}^2 (\omega_{j(k-1)y} - \omega_{jky})}{P^2_{jky|q_g}} \tag{69}$$

$$\frac{\partial^2 \log M}{\partial \gamma_{jry} \partial d_{jk}} = \sum_{g=1}^{G} \omega_{jky} q_{gr} [\frac{r_{jgky}}{P^2_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})] \tag{70}$$

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

$$E(\frac{\partial^2 \log M}{\partial a_{jr}^2}) = \sum_{y}^{3} \sum_{k=1}^{p} \sum_{g=1}^{G} - \frac{n_{gy} q_{gr}^2 (\omega_{j(k-1)y} - \omega_{jky})}{P_{jky|q_g}}, \tag{71}$$

$$E(\frac{\partial^2 \log M}{\partial d_{jk}^2}) = \sum_{y}^{3} \sum_{g=1}^{G} - n_{gy} (\frac{1}{P_{jky|q_g}} + \frac{1}{P_{j(k+1)y|q_g}}) \omega^2_{jky}, \tag{72}$$

$$E(\frac{\partial^2 \log M}{\partial d_{jk} \partial d_{j,k+1}}) = \sum_{y}^{3} \sum_{g=1}^{G} \frac{n_{gy}}{P_{j(k+1)y|q_g}} \omega^2_{jky} \omega^2_{j(k+1)y}, \tag{73}$$

and

$$E(\frac{\partial^2 \log M}{\partial a_{jr} \partial d_{jk}}) = \sum_{y}^{3} \sum_{g=1}^{G} n_{gy} \omega_{jky} q_{gr} [\frac{1}{P_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{1}{P_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})]. \tag{74}$$

$$E(\frac{\partial^2 \log M}{\partial \gamma_{jry}^2}) = E(\frac{\partial^2 \log M}{\partial \gamma_{jry} \partial a_{jr}}) = \sum_{k=1}^{p} \sum_{g=1}^{G} - \frac{n_{gy} q_{gr}^2 (\omega_{j(k-1)y} - \omega_{jky})}{P_{jky|q_g}}. \tag{75}$$

$$E(\frac{\partial^2 \log M}{\partial \gamma_{jry} \partial d_{jk}}) = \sum_{g=1}^{G} n_{gy} \omega_{jky} q_{gr} [\frac{1}{P_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{1}{P_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})]. \tag{76}$$

4.2.1 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

$$\begin{aligned} \hat{\mathbf{\Gamma}} &= \text{argmin}\{-\log M(\mathbf{\Gamma}) + \eta ||\mathbf{\Gamma}||_1\} \\ &= \text{argmin}\{-\log M(\mathbf{\Gamma}_0) - \partial_{\Gamma} \log M(\mathbf{\Gamma}_0)(\mathbf{\Gamma} - \mathbf{\Gamma}_0) - \frac{\partial^2_{\Gamma} \log M(\mathbf{\Gamma}_0)}{2} (\mathbf{\Gamma} - \mathbf{\Gamma}_0)^2 + \eta ||\mathbf{\Gamma}||_1\} \\ &= - \frac{\text{soft}(\partial_{\mathbf{\Gamma}} \log M - \mathbf{\Gamma}_j^{(t-1)} * \partial^2_{\mathbf{\Gamma}} \log M, \eta)}{\partial^2_{\mathbf{\Gamma}} \log M} \end{aligned} \tag{77}$$

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

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

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

and

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

Then we update $P^*_{jky|q_g}$ and $Q^*_{jky|q_g}$ 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

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

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

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

4.3 Simulation

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 difficulty parameter (-1) on the 4 DIF items.

Generated parameters.

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

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

$d_1 \sim N(0, 1)$

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

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

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

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

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

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

4.3.1 Results of 16 Replications

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

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Power	1	0.5	1
Type I	0.049	0.01875	0.022

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

Table 2. *Type I error and Power of mirt LRT*

Group	Omnibus DIF
Power	0.9875
Type I	0.0357

mirt LRT can only do the omnibus DIF test.

Both regularization and mirt LRT can detect DIF maginitude 1 with power 100%. Our regularization method has slightly lower Type I error.

Table 3. *Item parameter estimates by regularization*

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	-0.0160795	0.00515	0.02939
RMSE	0.141	0.145	0.146

Table 4. *Item parameter estimates by mirt LRT*

Item Parameters	$\boldsymbol{a}_1$	$\boldsymbol{a}_2$	$\boldsymbol{d}$
Bias	0.0129	0.00515	-0.00693
RMSE	0.142	0.123	0.0686

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

Table 5. *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)	0.255	0.302	0.208
Regularization (exclude false negative)	0.173	0.104	0.208
mirt LRT (include false negative)	0.174	0.169	0.179

Table 6. *Absolute bias for DIF magnitude recoveries that were* $non-DIF$

Group	Omnibus DIF	Group with DIF=0.5	Group with DIF=1
Regularization	0.336	0.33	0.341
mirt LRT	0.15696	0.1516	0.1622

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 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 $\boldsymbol{\beta} = (\mathbf{0}, ..., \mathbf{0}, \boldsymbol{\beta}_{q+1}, ..., \boldsymbol{\beta}_m)$.

5.1 E step

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

$$L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\theta}_i \mid \mathbf{y}, \boldsymbol{u}_i) = \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}}. \tag{78}$$

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

$$\begin{aligned} &L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) = L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\theta}_i \mid \mathbf{y}, \mathbf{u}_i) f(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{\theta}_i) \\ &= \prod_{j=1}^{m} \prod_{k=1}^{p} P_{jk}(\boldsymbol{\theta}_i)^{x_{ijk}} (2\pi)^{-p/2} |\mathbf{\Sigma}_y|^{-1/2} \exp(-0.5(\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)' \mathbf{\Sigma}_y^{-1} (\boldsymbol{\theta}_i - \boldsymbol{\mu}_y)). \end{aligned} \tag{79}$$

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

$$m(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i) = \int L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{\theta}_i) \partial \boldsymbol{\theta}_i \tag{80}$$

Then

$$h(\boldsymbol{\theta}_i \mid \boldsymbol{u}_i, \boldsymbol{y}_i, \mathbf{A}^{(t-1)}, \mathbf{D}^{(t-1)}, \mathbf{\Gamma}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \Sigma_y^{(t-1)}) = \frac{L(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{u}_i, \boldsymbol{\theta}_i) f(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{\theta}_i)}{m(\mathbf{A}, \mathbf{D}, \mathbf{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{y}, \boldsymbol{u}_i)} \tag{81}$$

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

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

$$\begin{aligned} &E[log\{L(\boldsymbol{A}, \boldsymbol{D}, \mathbf{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}, \mathbf{\Sigma} \mid \boldsymbol{Y}, \boldsymbol{U}, \mathbf{\Theta})\} | \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \mathbf{\Gamma}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{Y}, \mathbf{U}, \boldsymbol{\mu}^{(t-1)}, \mathbf{\Sigma}^{(t-1)}] \\ &= \sum_{i}^{N} \{ \int \log L(\boldsymbol{A}, \boldsymbol{D}, \mathbf{\Gamma}, \boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{u}_i, \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \mathbf{\Gamma}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}^{(t-1)}, \mathbf{\Sigma}^{(t-1)}) \partial \boldsymbol{\theta}_i \\ &+ \int \log f(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y \mid \boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i | \boldsymbol{u}_i, \boldsymbol{A}^{(t-1)}, \boldsymbol{D}^{(t-1)}, \mathbf{\Gamma}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\mu}_y^{(t-1)}, \mathbf{\Sigma}_y^{(t-1)}) \partial \boldsymbol{\theta}_i \} \end{aligned} \tag{82}$$

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

$$
\begin{aligned}
&E[logL(\boldsymbol{A},\boldsymbol{D},\boldsymbol{\Gamma},\beta,\boldsymbol{\mu},\boldsymbol{\Sigma}\mid \boldsymbol{Y},\boldsymbol{U})]\\
&=\sum_{i}^{N}\sum_{g}^{G}\mathrm{log}L(\boldsymbol{A},\boldsymbol{D},\boldsymbol{\Gamma},\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{\Gamma}^{(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{\Gamma}^{(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}\mathrm{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{\Gamma}^{(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{\Gamma}^{(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}\mathrm{log}P_{ijk|q_g}\frac{L(\boldsymbol{q}_g\mid \boldsymbol{u}_i,\boldsymbol{A}^{(t-1)},\boldsymbol{D}^{(t-1)},\boldsymbol{\Gamma}^{(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{\Gamma}^{(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}\mathrm{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{\Gamma}^{(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{\Gamma}^{(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{83}
$$

Then we can define two artificial terms.

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

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

and

$$r_{jgky} = r_{jgk1} = \sum_{i=1}^{N_1}x_{ijk}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_1^{(t-1)},\boldsymbol{\Sigma}_1^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_1^{(t-1)},\boldsymbol{\Sigma}_1^{(t-1)})}.$$

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

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

and

$$r_{jgky} = r_{jgk2} = \sum_{i=N_1+1}^{N_1+N_2}x_{ijk}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_2^{(t-1)},\boldsymbol{\Sigma}_2^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_2^{(t-1)},\boldsymbol{\Sigma}_2^{(t-1)})}.$$

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

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

and

$$r_{jgky} = r_{jgk3} = \sum_{i=N_1+N_2+1}^{N_1+N_2+N3}x_{ijk}\frac{L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_3^{(t-1)},\boldsymbol{\Sigma}_3^{(t-1)})}{\sum_g^G L(\boldsymbol{q}_g\mid \mathbf{y}_i,\boldsymbol{u}_i,\mathbf{A}^{(t-1)},\mathbf{D}^{(t-1)},\boldsymbol{\Gamma}^{(t-1)},\boldsymbol{\beta}^{(t-1)},\boldsymbol{\mu}_3^{(t-1)},\boldsymbol{\Sigma}_3^{(t-1)})}.$$

$n_g = n_{g1} + n_{g2} + n_{g3}$ 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 $\boldsymbol{q}_g$.

$$E[log\{L(\mathbf{A}, \mathbf{D}, \boldsymbol{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{U}, \boldsymbol{\Theta})] = \sum_y^3 \sum_g^G \sum_j^m \sum_k^p (r_{jgky} \text{log} P_{jky|q_g}) + \sum_y^3 \sum_g^G n_g \text{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{84}$$

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

$$\log M_j = \sum_y^3 \sum_g^G \sum_k^p (r_{jgky} \text{log} P_{jky|q_g}) + \sum_y^3 \sum_g^G n_g \text{log} f(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y \mid \boldsymbol{q}_g) \tag{85}$$

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

$$-\log M + \eta \sum_j^m ||(\boldsymbol{\Gamma}_j, \boldsymbol{\beta}_j)||_2 \tag{86}$$

For each item, we minimize

$$-\log M_j + \eta ||(\boldsymbol{\Gamma}_j, \boldsymbol{\beta}_j)||_2 \tag{87}$$

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

$$S_\eta(\boldsymbol{\tau}) = -\log M + \eta \sum_j^m ||\boldsymbol{\tau}||_2. \tag{88}$$

For each item j,

$$S_\eta(\boldsymbol{\tau}_j) = -\log M_j + \eta ||\boldsymbol{\tau}_j||_2. \tag{89}$$

5.2 M step

Same as before, 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.

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

$$\hat{\boldsymbol{\mu}}_2 = \frac{\sum_{g=1}^G n_{g2} \boldsymbol{q}_g}{N_2}, \tag{90}$$

and

$$\hat{\boldsymbol{\mu}}_3 = \frac{\sum_{g=1}^{G} n_{g3}\boldsymbol{q}_g}{N_3}. \tag{91}$$

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

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{\sum_{g=1}^{G} n_{g1}\boldsymbol{q}_g\boldsymbol{q}_g'}{N_1}, \tag{92}$$

$$\hat{\boldsymbol{\Sigma}}_2 = \frac{\sum_{g=1}^{G} n_{g2}(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{93}$$

and

$$\hat{\boldsymbol{\Sigma}}_3 = \frac{\sum_{g=1}^{G} n_{g3}(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{94}$$

Standardized quadrature points.

$$\boldsymbol{q}_g^* = \frac{q_g}{\sqrt{\mathrm{diag}\hat{\boldsymbol{\Sigma}}_1}}. \tag{95}$$

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

$$\hat{\boldsymbol{\Sigma}}_1^* = \frac{\sum_{g=1}^{G} n_{g1}\boldsymbol{q}_g^*\boldsymbol{q}_g^{*'}}{N_1}, \tag{96}$$

$$\hat{\boldsymbol{\Sigma}}_2^* = \frac{\sum_{g=1}^{G} n_{g2}(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_2)'}{N_2}, \tag{97}$$

and

$$\hat{\boldsymbol{\Sigma}}_3^* = \frac{\sum_{g=1}^{G} n_{g3}(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)(\boldsymbol{q}_g^* - \hat{\boldsymbol{\mu}}_3)'}{N_3}. \tag{98}$$

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

$$\frac{\partial \log M}{\partial a_{jr}} = \sum_{y}^{3}\sum_{k=1}^{p}\sum_{g=1}^{G}\left(\frac{r_{jgky}q_{gr}}{P_{jky|q_g}}(\omega_{j,(k-1),y} - \omega_{jky})\right) \tag{99}$$

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

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

$$\frac{\partial \log M}{\partial d_{jk}} = \sum_{y}^{3}\sum_{g}^{G}\omega_{jky}\left(\frac{r_{jg,(k+1),y}}{P_{j,(k+1),y|q_g}} - \frac{r_{jgky}}{P_{jky|q_g}}\right) \tag{100}$$

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

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

$$\begin{aligned}\frac{\partial \log M}{\partial \gamma_{jry}} &= \sum_{g}^{G}\sum_{k}^{p}\frac{r_{jgky}q_{gr}[P^{*}_{j(k-1)y|q_g}(1-P^{*}_{j(k-1)y|q_g})-P^{*}_{jky|q_g}(1-P^{*}_{jky|q_g})]}{P_{jky|q_g}}\\ &= \sum_{k}^{p}\sum_{g}^{G}(\frac{r_{jgky}q_{gr}}{P_{jky|q_g}}(\omega_{j(k-1)y}-\omega_{jky}))\end{aligned}\tag{101}$$

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

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

$$\frac{\partial \log M}{\partial \beta_{jky}} = \sum_{g}^{G}\omega_{jky}(\frac{r_{jg(k+1)y}}{P_{j(k+1)y|q_g}} - \frac{r_{jgky}}{P_{jky|q_g}})\tag{102}$$

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

The second partial derivatives in the Hessian matrix are given by

$$\begin{aligned}\frac{\partial^2 \log M}{\partial a^2_{jr}} &= \sum_{y}^{3}\sum_{k=1}^{p}\sum_{g=1}^{G}-\frac{r_{jgky}q^2_{gr}(P^{*}_{j(k-1)y|q_g}Q^{*}_{j(k-1)y|q_g}-P^{*}_{jky|q_g}Q^{*}_{jky|q_g})^2}{P^2_{jky|q_g}}\\ &= \sum_{y}^{3}\sum_{k=1}^{p}\sum_{g=1}^{G}-\frac{r_{jgky}q^2_{gr}(\omega_{j(k-1)y}-\omega_{jky})}{P^2_{jky|q_g}}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \log M}{\partial d^2_{jk}} &= \sum_{y}^{3}\sum_{g=1}^{G}-(\frac{r_{jgky}}{P^2_{jky|q_g}}+\frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}})P^{*2}_{jky|q_g}(1-P^{*}_{jky|q_g})^2\\ &= \sum_{y}^{3}\sum_{g=1}^{G}-(\frac{r_{jgky}}{P^2_{jky|q_g}}+\frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}})\omega^2_{jky}\end{aligned}$$

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

and

$$\begin{aligned}\frac{\partial^2 \log M}{\partial a_{jr}\partial d_{jk}} &= \sum_{y}^{3}\sum_{g=1}^{G}P^{*}_{jky}Q^{*}_{jky}q_{gr}[\frac{r_{jgky}}{P^2_{jky|q_g}}(P^{*}_{j(k-1)y|q_g}Q^{*}_{j(k-1)y|q_g}-P^{*}_{jky|q_g}Q^{*}_{jky|q_g})\\ &\quad+\frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}}(P^{*}_{jky|q_g}Q^{*}_{jky|q_g}-P^{*}_{j(k+1)y|q_g}Q^{*}_{j(k+1)y|q_g})]\end{aligned}$$

$$= \sum_{y}^{3}\sum_{g=1}^{G}\omega_{jky}q_{gr}[\frac{r_{jgky}}{P^2_{jky|q_g}}(\omega_{j(k-1)y}-\omega_{jky})+\frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}}(\omega_{jky}-\omega_{j(k+1)y})]$$

where

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

$$\frac{\partial^2 \log M}{\partial \gamma^2_{jry}} = \frac{\partial^2 \log M}{\partial \gamma_{jry} \partial a_{jr}} = \sum_{k=1}^{p} \sum_{g=1}^{G} -\frac{r_{jgky} q^2_{gr} (\omega_{j(k-1)y} - \omega_{jky})}{P^2_{jky|q_g}}$$

$$\frac{\partial^2 \log M}{\partial \gamma_{jry} \partial d_{jk}} = \frac{\partial^2 \log M}{\partial a_{jr} \partial \beta_{jky}} = \frac{\partial^2 \log M}{\partial \gamma_{jry} \partial \beta_{jky}} = \sum_{g=1}^{G} \omega_{jky} q_{gr} [\frac{r_{jgky}}{P^2_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})]$$

where

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

$$\frac{\partial^2 \log M}{\partial \beta^2_{jky}} = \frac{\partial^2 \log M}{\partial \beta_{jky} \partial d_{jk}} = \sum_{g=1}^{G} -(\frac{r_{jgky}}{P^2_{jky|q_g}} + \frac{r_{jg(k+1)y}}{P^2_{j(k+1)y|q_g}}) P^{*2}_{jky|q_g} (1 - P^*_{jky|q_g})^2$$

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

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

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

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

and

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

$$E(\frac{\partial^2 \log M}{\partial \gamma^2_{jry}}) = E(\frac{\partial^2 \log M}{\partial \gamma_{jry} \partial a_{jr}}) = \sum_{k=1}^{p} \sum_{g=1}^{G} -\frac{n_{gy} q^2_{gr} (\omega_{j(k-1)y} - \omega_{jky})}{P_{jky|q_g}}$$

$$E(\frac{\partial^2 \log M}{\partial \gamma_{jry} \partial d_{jk}}) = E(\frac{\partial^2 \log M}{\partial a_{jr} \partial \beta_{jky}}) = E(\frac{\partial^2 \log M}{\partial \gamma_{jry} \partial \beta_{jky}}) = \sum_{g=1}^{G} n_{gy} \omega_{jky} q_{gr} [\frac{1}{P_{jky|q_g}} (\omega_{j(k-1)y} - \omega_{jky}) + \frac{1}{P_{j(k+1)y|q_g}} (\omega_{jky} - \omega_{j(k+1)y})].$$

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

5.2.1 Block co-ordinate gradient descent

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

Then, our objective function is

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

Using a second-order Taylor series expansion at $\hat{\boldsymbol{\tau}}^{(t-1)}$ (and replacing the Hessian of the marginal log-likelihood $\log M(\cdot)$ by a suitable matrix $H^{(t-1)}$) we define

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

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

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

We run a block co-ordinate gradient descent algorithm for each group (item) with all other groups fixed. For item j, our algorithm is given by following.

1. Calculate $P^*_{jky|q_g}, Q^*_{jky|q_g}$ and $||\boldsymbol{\tau}_j||_2$.
2. The parameter a_{jr} and d_{jk} can be updated by

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

and

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

3. Denote u to be the subgradient of $||\boldsymbol{\tau}_j^{(t-1)} + \boldsymbol{\epsilon}_j||_2$. We have

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

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

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

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

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

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

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

Otherwise,

Then the subgradient equation is

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

$$-\nabla \log M_j - \boldsymbol{\epsilon}_j^T H_{jj}^{(t-1)} + \eta \frac{(\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j)(-H_{jj}^{(t-1)})}{||\hat{\boldsymbol{\tau}}_j^{(t-1)} + \boldsymbol{\epsilon}_j||_2(-H_{jj}^{(t-1)})} = 0$$

$$-\nabla \log M_j - \boldsymbol{\epsilon}_j^T H_{jj}^{(t-1)} + \eta \frac{\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_{jj}^{(t-1)}}{||\nabla \log M_j - \hat{\boldsymbol{\tau}}_j^{(t-1)} H_{jj}^{(t-1)}||_2} = 0$$

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

$$\nabla \log M_j = (\frac{\partial \log M}{\partial \gamma_{jry}}, \frac{\partial \log M}{\partial \beta_{jky}}), r = 1, ..., q; k = 1, ..., p-1; y = 2, 3.$$

If $\boldsymbol{\epsilon}_j^{(t-1)} \neq 0$, performing a Backtracking-Armijo line search: let $\alpha^{(t-1)}$ be the largest value in $\{\alpha^{(0)}\delta^l\}_{l\geq 0}$ s.t.

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

where $\alpha^{(0)} = 1$, $\delta = 0.5$ and $\sigma = 0.1$, and $\Delta^{(t-1)}$ is the improvement in the objective function $S_\eta(\cdot)$ when using a linear approximation for the log-likelihood, i.e.

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

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

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 $\hat{\boldsymbol{\Sigma}_1}$.