Supplementary Material for "Learning Latent and Hierarchical Structures in Cognitive Diagnosis Models"

Chenchen Ma, Jing Ouyang, and Gongjun Xu

Department of Statistics, University of Michigan

In the supplementary material, we provide the proof of the main theorem, the derivations for the penalized EM algorithm and a sensitivity analysis of our algorithm with varying upper bounds for the number of latent classes.

1 Proof for Theorem 3

In this section, we provide the proof of Theorem 3.

Proof. We first introduce some notations. For two sequences $\{a_N\}$ and $\{b_N\}$, we denote $a_N \lesssim b_N$ if $a_N = O(b_N)$, and $a_N \asymp b_N$ if $a_N \lesssim b_N$ and $b_N \lesssim a_N$. We use $(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)$ to denote the true model parameter and use $(\hat{\boldsymbol{\pi}}^0, \hat{\boldsymbol{\Theta}}^0)$ to denote the oracle MLE obtained by assuming the number of latent attributes, the hierarchical structure, the Q-matrix and the item-level diagnostic models are known. Let $(\hat{\boldsymbol{\pi}}^*, \hat{\boldsymbol{\Theta}}^*)$ be the MLE obtained by directly optimizing log-likelihood (9) and $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}})$ be the estimator obtained by optimizing the regularized log-likelihood (10). We define $\hat{\boldsymbol{\pi}}_{\rho_N} := \{\hat{\pi}_m : \hat{\pi}_m > \rho_N,\ m \in [M]\}$ and $\hat{\boldsymbol{\Theta}}_{\rho_N} := \{\hat{\theta}_{j,m} : \hat{\pi}_m > \rho_N,\ j \in [J],\ m \in [M]\}$, the model parameters corresponding to the selected latent classes. Let M be the upper bound for the number of latent classes, M_0 be the true number of latent classes, and $\hat{M} = \big|\{m : \hat{\pi}_m > \rho_N,\ m \in [M]\}\big|$ be the estimated number of latent classes. Without loss of generality, let $\hat{\boldsymbol{\pi}}^0_{\text{full}} = (\hat{\boldsymbol{\pi}}^0, \mathbf{0}_{M-M_0})$. For the true item parameter matrix $\boldsymbol{\Theta}^0$, we defined the set of identical item parameter pairs $S^0 = \big\{(j, k_1, k_2) : \theta^0_{j,k_1} = \theta^0_{j,k_2}, 1 \le k_1 < k_2 \le M_0\big\}$. Similarly, for $(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}})$ we define $\hat{S} = \big\{(j, k_1, k_2) : \hat{\theta}_{j,k_1} = \hat{\theta}_{j,k_2},\ 1 \le k_1 < k_2 \le M,\ \hat{\pi}_{k_1} > \rho_N,\ \hat{\pi}_{k_2} > \rho_N\big\}$. We say $\hat{S} \sim S^0$ if there exists a column permutation σ of $\hat{\boldsymbol{\Theta}}$ such that $\hat{S}_\sigma = S^0$.

The probability $\mathbb{P}\big(\hat{M} \neq M_0\big)$ can be decomposed into two parts:

$$\mathbb{P}\big(\hat{M} \neq M_0\big) = \mathbb{P}\big(\hat{M} < M_0\big) + \big(\hat{M} > M_0\big). \tag{1}$$

Similarly, the probability $\mathbb{P}\big(\hat{S} \neq S^0\big)$ can be decomposed into three parts:

$$\mathbb{P}\big(\hat{S} \nsim S^0\big) = \mathbb{P}\big(\hat{M} < M_0\big) + \big(\hat{M} > M_0\big) + \mathbb{P}\big(\hat{S} \nsim S^0, \hat{M} = M_0\big). \tag{2}$$

In the following ,we will bound each part in (1) and (2) respectively. Therefore, we will consider three cases below:

1. overfitted case: $\hat{M} > M_0$,
2. underfitted case: $\hat{M} < M_0$,
3. $\hat{M} = M_0$ but $\hat{S} \nsim S^0$.

The objective function is

$$G_N(\boldsymbol{\pi}, \boldsymbol{\Theta}) = \frac{l_N(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{R})}{N} - \frac{\lambda_N^{(1)}}{N} \sum_{k=1}^{M} \log_{[\rho_N]} \pi_k - \frac{\lambda_N^{(2)}}{N} \sum_{j=1}^{J} \mathcal{J}_{\tau,\rho_N}(\boldsymbol{\theta}_j), \tag{3}$$

where $\log_{[\rho_N]} \pi_k = \log \pi_k \cdot \mathbb{I}\big(\pi_k > \rho_N\big) + \log \rho_N \cdot \mathbb{I}\big(\pi_k \leq \rho_N\big)$. Let $\log_{[\rho_N]}(\boldsymbol{\pi}) = \sum_{k=1}^{M} \log_{[\rho_N]} \pi_k$.

First consider the overfitted case where $\hat{M} > M_0$. The event $\big\{G_N\big(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}\big) > G_N\big(\hat{\boldsymbol{\pi}}^0, \hat{\boldsymbol{\Theta}}^0\big)\big\}$ implies that

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^{N} \log \Big[\frac{\sum_{k=1}^{M} \hat{\pi}_k \prod_{j=1}^{J} \hat{\theta}_{j,k}^{R_{ij}} (1 - \hat{\theta}_{j,k})^{1-R_{ij}}}{\sum_{k=1}^{M} \hat{\pi}_k^0 \prod_{j=1}^{J} (\hat{\theta}_{j,k}^0)^{R_{ij}} (1 - \hat{\theta}_{j,k}^0)^{1-R_{ij}}} \Big] \\
&> \frac{\lambda_N^{(1)}}{N} \big\{ \log_{[\rho_N]}(\hat{\boldsymbol{\pi}}) - \log_{[\rho_N]}(\hat{\boldsymbol{\pi}}_{full}^0) \big\} + \frac{\lambda_N^{(2)}}{N} \Big\{ \sum_{j=1}^{J} \mathcal{J}_{\tau,\rho_N}(\hat{\boldsymbol{\theta}}_j) - \sum_{j=1}^{J} \mathcal{J}_{\tau,\rho_N}(\hat{\boldsymbol{\theta}}_j^0) \Big\} \qquad (4) \\
&:= J_1 + J_2.
\end{aligned}$$

For the RHS of (4), we have $J_1 \gtrsim N^{-1} \lambda_N^{(1)} |\log \rho_N|$ and $J_2 \gtrsim -N^{-1} \lambda_N^{(2)} \tau J M^2$. Since $\lambda_N^{(2)} \tau = o(\lambda_N^{(1)} |\log \rho_N|)$, we have RHS $\gtrsim N^{-1} \lambda_N^{(1)} |\log \rho_N|$.

For the LHS of (4), we have

$$\begin{aligned}\text{LHS of (4)} =& \frac{1}{N}\log\Big[\sum_{k=1}^{M}\hat{\pi}_k\prod_{j=1}^{J}\hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k})^{1-R_{ij}}\Big]-\frac{1}{N}\log\Big[\sum_{k=1}^{M}\hat{\pi}_k^0\prod_{j=1}^{J}(\hat{\theta}_{j,k}^0)^{R_{ij}}(1-\hat{\theta}_{j,k}^0)^{1-R_{ij}}\Big]\\ \le& \frac{1}{N}\log\Big[\sum_{k=1}^{M}\hat{\pi}_k^*\prod_{j=1}^{J}(\hat{\theta}_{j,k}^*)^{R_{ij}}(1-(\hat{\theta}_{j,k}^*))^{1-R_{ij}}\Big]-\frac{1}{N}\log\Big[\sum_{k=1}^{M}\hat{\pi}_k^0\prod_{j=1}^{J}(\hat{\theta}_{j,k}^0)^{R_{ij}}(1-\hat{\theta}_{j,k}^0)^{1-R_{ij}}\Big]\\ \lesssim& N^{-\delta},\end{aligned}$$

where the last inequality follows from Assumption 1. When $N^{1-\delta}/|\log(\rho_N)| = o\big(\lambda_N^{(1)}\big)$, we have $N^{-\delta} = o\big(N^{-1}\lambda_N^{(1)}|\log\rho_N|\big)$, which implies that the event described in (4) will happen with probability tending to zero. Therefore we have $\mathbb{P}\big(\hat{M} > M_0\big) \to 0$ as $N \to \infty$. That is to say, with the appropriate choice of tuning parameters, the extent that the log-penalty part favors a smaller model would dominate the extent that the likelihood part favors a larger model in the overfitted case.

Now consider the under-fitted case where $\hat{M} < M_0$. We need to bound

$$\mathbb{P}\Big(\sup_{\hat{M}<M_0}\big[G_N(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}})-G_N(\hat{\boldsymbol{\pi}}^0,\hat{\boldsymbol{\Theta}}^0)\big]>0\Big). \tag{5}$$

We follow a similar argument to Shen et al. (2012). More specifically, since

$$\mathbb{P}\Big(\sup_{\hat{M}<M_0}\big[G_N(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}})-G_N(\hat{\boldsymbol{\pi}}^0,\hat{\boldsymbol{\Theta}}^0)\big]>0\Big) \le \sum_{m=1}^{M_0-1}\mathbb{P}\Big(\sup_{\hat{M}=m}\big[G_N(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}})-G_N(\hat{\boldsymbol{\pi}}^0,\hat{\boldsymbol{\Theta}}^0)\big]>0\Big), \tag{6}$$

we will bound each term in the RHS of (6). By the large deviation inequality in Theorem 1 of Wong and Shen (1995), we have

$$\begin{aligned}&\mathbb{P}\Big(\sup_{h^2\left((\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}),(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)\right)\ge\epsilon_N^2}\big[\frac{1}{N}l_N\big(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}\big)-\frac{1}{N}l_N\big(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0\big)\big]>-\epsilon_N^2\Big)\\ \le&\mathbb{P}\Big(\sup_{h^2\left((\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}),(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)\right)\ge\epsilon_N^2}\big[\frac{1}{N}l_N\big(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}\big)-\frac{1}{N}l_N\big(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0\big)\big]>-\epsilon_N^2\Big)\le\exp(-N\epsilon_N^2),\end{aligned} \tag{7}$$

where $h^2\big((\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}),(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)\big) = \sum_{\boldsymbol{R}\in\{0,1\}^J}\big[\mathbb{P}(\boldsymbol{R}\mid\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}})^{1/2}-\mathbb{P}(\boldsymbol{R}\mid\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)^{1/2}\big]$ is the Hellinger distance. From the remark in Wong and Shen (1995), the inequality (7) holds for any $t > \epsilon_N$.

To use this large deviation inequality, we need to introduce the notion of bracketing Hellinger metric entropy $H(t, \mathcal{B}_m)$, which characterizes the size of the local parameter space. Consider the local parameter space $\mathcal{B}_m = \big\{(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}) : \hat{M} = m \le M_0,\ h^2\big((\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}), (\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)\big) \le 2\epsilon_N^2\big\}$, then $H(t, \mathcal{B}_m)$ is defined as the logarithm of the cardinality of the t-bracketing of $\mathcal{B}_m$ of the smallest size. Specifically, following the definition in Shen et al. (2012), consider a bracket covering $S(t, m) = \{f_1^l, f_1^u, \cdots, f_m^l, f_m^u\}$ such that $\max_{1\le j\le m} ||f_j^u - f_j^l||_2 \le t$ and for any $f \in \mathcal{B}_m$, there is some j such that $f_j^l \le f \le f_j^u$ almost surely. Then $H(t, \mathcal{B}_m)$ is defined as $\log\big(\min\{m : S(t, m)\}\big)$. Following Lemma 3 in Gu and Xu (2019), for any $2^{-4}\epsilon < t < \epsilon$, there is

$$H(t, \mathcal{B}_m) \lesssim M_0 \log M \log(2\epsilon/t). \tag{8}$$

Next we need to verify the conditions in Wong and Shen (1995). Let's take $\epsilon_N = \sqrt{M_0 \log M/N}$ and verify the entropy integral condition in Theorem 1 of Wong and Shen (1995) for such ϵ_N. The integral of bracketing Hellinger metric entropy on the interval $[2^{-8}\epsilon_N^2, \sqrt{2}\epsilon_N]$ satisfies the following inequality

$$\begin{aligned}
\int_{2^{-8}\epsilon_N^2}^{\sqrt{2}\epsilon_N} H^{1/2}(t, \mathcal{B}_m)dt &\le \int_{2^{-8}\epsilon_N^2}^{\sqrt{2}\epsilon_N} \sqrt{M_0 \log M \log(2\epsilon_N/t)}dt \\
&= \sqrt{M_0 \log M} \int_{\sqrt{\log\sqrt{2}}}^{\sqrt{\log\frac{2^9}{\epsilon_N}}} 4\epsilon_N u^2 e^{-u^2} du \\
&= \sqrt{M_0 \log M} \cdot 2\epsilon_N \int_{\log\sqrt{2}}^{\log\frac{2^9}{\epsilon_N}} \sqrt{u} e^{-u} du \\
&\lesssim \sqrt{N}\epsilon_N^2.
\end{aligned}$$

Note that $\epsilon_N = o(1)$ as $N \to \infty$.

Following the proof in Gu and Xu (2019), there exists a constant c_0, for some small constant $t > \epsilon_N$, we have

$$C_{\min}(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0) := \inf_{(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}): \hat{M} \le M_0} \left\{ \frac{h^2\big((\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}), (\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)\big)}{\max\big(M_0 - \hat{M}, 1\big)} \right\} \ge c_0 \gtrsim t^2 > \epsilon_N^2.$$

Moreover, for $\hat{M} = m < M_0$, there is $h^2((\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}), (\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)) \ge \big(M_0 - m\big) C_{\min}(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)$. In order

to have the probability of the event (4) go to zero in the under-fitted case, the log-penalty term should not be too large such that the likelihood part is dominated by the log-penalty term that favors a smaller model. Here we take $\lambda_N^{(1)} = o(N \log \rho_N|^{-1})$. Then for (6) we have

$$
\begin{aligned}
&\text{RHS of (6)} \\
&\leq \sum_{m=1}^{M_0-1} \mathbb{P}\Big(\sup_{h^2((\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}),(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)) \geq (M_0-m)C_{\min}(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0), \hat{M}=m} \big[G_N(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}) - G_N(\hat{\boldsymbol{\pi}}^0,\hat{\boldsymbol{\Theta}}^0) \big] > 0 \Big) \\
&\leq \sum_{m=1}^{M_0-1} \mathbb{P}\Big(\sup_{h^2((\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}),(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)) \geq (M_0-m)C_{\min}(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0), \hat{M}=m} \big[l_N(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}) - l_N(\hat{\boldsymbol{\pi}}^0,\hat{\boldsymbol{\Theta}}^0) \big] > -\frac{\lambda_N^{(1)} M_0 |\log \rho_N|}{N} \Big) \\
&\leq \sum_{m=1}^{M_0-1} \mathbb{P}\Big(\sup_{h^2((\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}),(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)) \geq (M_0-m)C_{\min}(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0), \hat{M}=m} \big[l_N(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}) - l_N(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0) \big] > -\frac{\lambda_N^{(1)} M_0 |\log \rho_N|}{N} \Big) \\
&\leq \sum_{m=1}^{M_0-1} \mathbb{P}\Big(\sup_{h^2((\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}),(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0)) \geq (M_0-m)C_{\min}(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0), \hat{M}=m} \big[l_N(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}}) - l_N(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0) \big] > -(M_0-m)C_{\min}(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0) \Big) \\
&\leq \sum_{m=1}^{M_0-1} \exp\big(-c_2 N (M_0-m) C_{\min}(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0) \big) \\
&\leq c_3 \exp\big(-c_2 N C_{\min}(\boldsymbol{\pi}^0,\boldsymbol{\Theta}^0) \big).
\end{aligned}
$$

Therefore we have $\mathbb{P}\big(\hat{M} < M_0\big) \to 0$ as $N \to \infty$. So far we have proved (12) in Theorem 3,

$$\mathbb{P}\big(\hat{M} \neq M_0\big) = \mathbb{P}\big(\hat{M} < M_0\big) + \mathbb{P}\big(\hat{M} > M_0\big) \longrightarrow 0.$$

Finally we consider the third case where $\hat{M} = M_0$ but $\hat{S} \nsim S^0$. The argument is similar to the proof of Proposition 2 in Xu and Shang (2018). We first show $(\hat{\boldsymbol{\pi}}_{\rho_N}, \hat{\boldsymbol{\Theta}}_{\rho_N})$ converge to $(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)$ with rate $N^{-1/2}$. For $(\boldsymbol{\pi}, \boldsymbol{\Theta})$ with $(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N})$ in a small neighborhood of $(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)$,

$$
\begin{aligned}
G'_N(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}) :=& \frac{l_N(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}; \mathcal{R})}{N} - \frac{\lambda_N^{(1)}}{N} \sum_{k:\pi_k > \rho_N} \log \pi_k - \frac{\lambda_N^{(2)}}{N} \sum_{j=1}^{J} \mathcal{J}_{\tau,\rho_N}(\boldsymbol{\theta}_j) \\
=& \frac{l_N(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}; \mathcal{R})}{N} - O(\lambda_N^{(1)} N^{-1} |\log \rho_N|) - O(\lambda_N^{(2)} \tau N^{-1}),
\end{aligned}
$$

converges uniformly to the same limit of $l_N(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}; \mathcal{R})/N$ by the uniform law of large number, since $\lambda_N^{(1)} N^{-1} |\log \rho_N| \to 0$ and $\lambda_N^{(2)} \tau N^{-1} \to 0$. We use $G_0(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N})$ to denote the limit process, which is the expectation of the negative log-likelihood of a single observation. By Taylor's

expansion, we have $G_0(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}) - G_0(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0) = O(\big|\big|(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N})) - (\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)\big|\big|^2)$.

For the log-likelihood function $l_N(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}; \mathcal{R}) = \sum_{i=1}^N \log\big(\sum_{k=1}^M \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})\big)$, we have

$$
\begin{aligned}
&\frac{1}{N}\big|l_N(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}; \mathcal{R}) - l_N(\hat{\boldsymbol{\pi}}_{\rho_N}, \hat{\boldsymbol{\Theta}}_{\rho_N}; \mathcal{R})\big| \\
\leq &\frac{1}{N}\sum_{i=1}^N \Big|\log\big(\sum_{k=1}^M \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})\big) - \log\big(\sum_{k:\hat{\pi}_k>\rho_N} \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})\big)\Big| \\
\leq &\frac{1}{N}\sum_{i=1}^N \frac{\big|\big(\sum_{k=1}^M \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})\big) - \big(\sum_{k:\hat{\pi}_k>\rho_N} \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})\big)\big|}{\sqrt{\big(\sum_{k=1}^M \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})\big) \times \big(\sum_{k:\hat{\pi}_k>\rho_N} \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})\big)}} \qquad (9) \\
\leq &\frac{1}{N}\sum_{i=1}^N \frac{(M-\hat{M})\rho_N}{\sum_{k:\hat{\pi}_k>\rho_N} \hat{\pi}_k \prod_{j=1}^J \hat{\theta}_{j,k}^{R_{ij}}(1-\hat{\theta}_{j,k}^{1-R_{ij}})} \\
= &O(\rho_N) = O(N^{-d}),\ d \geq 1, \qquad (10)
\end{aligned}
$$

where inequality (9) follows from an upper bound for log function. Specifically, for $x \geq 1$, we know $\log x \leq (x-1)/\sqrt{x}$, and thus for $0 < x \leq y$, we have $\log y - \log x \leq (y-x)/\sqrt{xy}$. From (10), $G'_N(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Theta}}) = G'_N(\hat{\boldsymbol{\pi}}_{\rho_N}, \hat{\boldsymbol{\Theta}}_{\rho_N}) + O(N^{-d}) \geq G'_N(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)$ and thus $G'_N(\hat{\boldsymbol{\pi}}_{\rho_N}, \hat{\boldsymbol{\Theta}}_{\rho_N}) > G'_N(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0) - O(N^{-d}) \geq G'_N(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0) - O(N^{-1})$. Since $N^{-1/2}\lambda_N^{(1)} \to 0$ and $N^{-1/2}\lambda_N^{(2)}\tau \to 0$, then for sufficiently small ζ, by Taylor's expansion,

$$
\mathbb{E}\Big(\sup_{||(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}) - (\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)|| \leq \zeta} G'_N(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}; \mathcal{R}) - G_0(\boldsymbol{\pi}_{\rho_N}, \boldsymbol{\Theta}_{\rho_N}) - G'_N(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0; \mathcal{R}) + G_0(\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0)\Big) = O(\zeta N^{-1/2}).
$$

By Theorem 3.2.5 in Van Der Vaart and Wellner (1996), we have $(\hat{\boldsymbol{\pi}}_{\rho_N}, \hat{\boldsymbol{\Theta}}_{\rho_N}) - (\boldsymbol{\pi}^0, \boldsymbol{\Theta}^0) = O_p(N^{-1/2})$.

We next show selection consistency of S^0. If true item parameters $\theta^0_{j,k_1} \neq \theta^0_{j,k_2}$, then from the above convergence result, we know $\hat{\theta}_{j,k_1} \to \theta^0_{j,k_1}$ and $\hat{\theta}_{j,k_2} \to \theta^0_{j,k_2}$, and thus $\hat{\theta}_{j,k_1} \neq \hat{\theta}_{j,k_2}$ in probability. If true item parameters $\theta^0_{j,k_1} = \theta^0_{j,k_2}$ but $\hat{\theta}_{j,k_1} \neq \hat{\theta}_{j,k_2}$, by the Karush-Kuhn-Tucker (KKT) conditions, we have $N^{-1/2}\partial l_N(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{R})/\partial\theta_{j,k_1}|_{(\boldsymbol{\pi},\boldsymbol{\Theta})=(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}})} = N^{-1/2}\lambda_N^{(2)} \to \infty$ in probability. However $N^{-1/2}\partial l_N(\boldsymbol{\pi}, \boldsymbol{\Theta}; \mathcal{R})/\partial\theta_{j,k_1}|_{(\boldsymbol{\pi},\boldsymbol{\Theta})=(\hat{\boldsymbol{\pi}},\hat{\boldsymbol{\Theta}})} = O_p(1)$. Therefore, if $\theta^0_{j,k_1} = \theta^0_{j,k_2}$, we have $\hat{\theta}_{j,k_1} = \hat{\theta}_{j,k_2}$ in probability, which proved the selection consistency that $\mathbb{P}(\hat{S} \nsim S^0) \to 0$ as $N \to \infty$. ■

2 Derivations of PEM Algorithm

In this section, we give detailed derivations of the penalized EM algorithm in Section 4.1. First let's introduce a new variable $\boldsymbol{d} = (d_{jkl}, j = 1, \ldots, J, 1 \le k < l \le M)$ to be the differences of the item parameters for each item. Then our problem becomes

$$\begin{aligned} \min_{\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{d}} \quad & G(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{d}) \\ \text{s.t.} \quad & d_{jkl} = \theta_{jk} - \theta_{jl} \\ & j = 1, \ldots, J, \ 1 \le k < l \le M. \end{aligned} \tag{11}$$

By using the difference convex property of the truncated Lasso penalty, we can decompose the objective function into two parts:

$$G(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{d}) = G_1(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{d}) - G_2(\boldsymbol{d}), \tag{12}$$

where

$$G_1(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{d}) = -\frac{1}{N} Q(\boldsymbol{\pi}, \boldsymbol{\Theta} | \boldsymbol{\pi}^{(c)}, \boldsymbol{\Theta}^{(c)}) + \tilde{\lambda}_1 \sum_{k=1}^{M} \log \pi_k + \tilde{\lambda}_2 \sum_{j=1}^{J} \sum_{1 \le k < l \le M} |d_{jkl}|, \tag{13}$$

$$G_2(\boldsymbol{d}) = \tilde{\lambda}_2 \sum_{j=1}^{J} \sum_{1 \le k < l \le M} \big(|d_{jkl} - \tau| \big)_+. \tag{14}$$

Then we construct a sequence of upper approximation of $G(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{d})$ iteratively by replacing $G_2(\boldsymbol{d})$ at iteration $c+1$ with its piecewise affine minorization:

$$G_2^{(c)}(\boldsymbol{d}) = G_2(\hat{\boldsymbol{d}}^{(c)}) + \tilde{\lambda}_2 \sum_{j=1}^{J} \sum_{1 \le k < l \le M} \big(|d_{jkl}| - |\hat{d}_{jkl}^{(c)}| \big) \cdot \mathbb{I}\big(|\hat{d}_{jkl}^{(c)}| \ge \tau \big), \tag{15}$$

at the current estimate $\hat{\boldsymbol{d}}^{(c)}$, which lead to an upper convex approximation:

$$\begin{aligned} G^{(c+1)}(\boldsymbol{\pi}, \boldsymbol{\Theta}, \boldsymbol{d}) = & -\frac{1}{N} Q(\boldsymbol{\pi}, \boldsymbol{\Theta} | \boldsymbol{\pi}^{(c)}, \boldsymbol{\Theta}^{(c)}) + \tilde{\lambda}_1 \sum_{k=1}^{M} \log \pi_k \\ & + \tilde{\lambda}_2 \sum_{j=1}^{J} \sum_{1 \le k < l \le M} |d_{jkl}| \cdot \mathbb{I}\big(|\hat{d}_{jkl}^{(c)}| < \tau \big) \end{aligned}$$

$$+\tilde{\lambda}_2\tau\sum_{j=1}^{J}\sum_{1\le k<l\le M}\mathbb{I}\big(|\hat{d}_{jkl}^{(c)}|\ge\tau\big).$$

Now we can apply ADMM. At iteration $c+1$, the augmented Lagrangian is

$$L_\gamma(\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{d},\boldsymbol{y})=G^{(c+1)}(\boldsymbol{\pi},\boldsymbol{\Theta},\boldsymbol{d})+\sum_{j=1}^{J}\sum_{1\le k<l\le M}y_{jkl}\cdot\big(d_{jkl}-(\theta_{jk}-\theta_{jl})\big)+\frac{\gamma}{2}\sum_{j=1}^{J}\sum_{1\le k<l\le M}\big|d_{jkl}-(\theta_{jk}-\theta_{jl})\big|^2, \tag{16}$$

where y_{jkl}'s are the dual variables and γ is a nonnegative penalty parameter. Then ADMM (Boyd et al., 2011) consists of the following iterations:

$$\begin{aligned}
\boldsymbol{\pi}^{(c+1)}&=\underset{\boldsymbol{\pi}}{\operatorname{argmin}}\ L_\gamma(\boldsymbol{\pi},\boldsymbol{\Theta}^{(c)},\boldsymbol{d}^{(c)},\boldsymbol{y}^{(c)}),\\
\boldsymbol{\Theta}^{(c+1)}&=\underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\ L_\gamma(\boldsymbol{\pi}^{(c+1)},\boldsymbol{\Theta},\boldsymbol{d}^{(c)},\boldsymbol{y}^{(c)}),\\
\boldsymbol{d}^{(c+1)}&=\underset{\boldsymbol{d}}{\operatorname{argmin}}\ L_\gamma(\boldsymbol{\pi}^{(c+1)},\boldsymbol{\Theta}^{(c+1)},\boldsymbol{d},\boldsymbol{y}^{(c)}),\\
y_{jkl}^{(c+1)}&=y_{jkl}^{(c)}+\gamma(d_{jkl}^{(c+1)}-\big(\theta_{jk}^{(c+1)}-\theta_{jl}^{(c+1)}\big)),\ j=1,...,J,1\le k<l\le M.
\end{aligned}$$

Using the scaled Lagrangian multiplier $\mu_{jkl}=y_{jkl}/\gamma$ and defining the residual $r_{jkl}=d_{jkl}-(\theta_{jk}-\theta_{jl})$, we have:

$$\begin{aligned}
&y_{jkl}\cdot\big(d_{jkl}-(\theta_{jk}-\theta_{jl})\big)+\frac{\gamma}{2}\big|d_{jkl}-(\theta_{jk}-\theta_{jl})\big|^2\\
=&y_{jkl}\cdot r_{jkl}+\frac{\gamma}{2}r_{jkl}^2\\
=&\frac{\gamma}{2}\big(r_{jkl}+(1/\gamma)y_{jkl}\big)^2-\frac{1}{2\gamma}\mu_{jkl}^2\\
=&\frac{\gamma}{2}\big(r_{jkl}+\mu_{jkl}\big)^2-\frac{1}{2\gamma}\mu_{jkl}^2.
\end{aligned}$$

Then using the scaled dual variable, we can express ADMM as:

$$\begin{aligned}
\boldsymbol{\pi}^{(c+1)}&=\underset{\boldsymbol{\pi}}{\operatorname{argmin}}\ G^{(c+1)}(\boldsymbol{\pi},\boldsymbol{\Theta}^{(c)},\boldsymbol{d}^{(c)}),\\
\boldsymbol{\Theta}^{(c+1)}&=\underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\ G^{(c+1)}(\boldsymbol{\pi}^{(c+1)},\boldsymbol{\Theta},\boldsymbol{d}^{(c)})+\frac{\gamma}{2}\sum_{j=1}^{J}\sum_{1\le k<l\le M}\big(d_{jkl}^{(c)}-(\theta_{jk}^{(c)}-\theta_{jl}^{(c)})+\mu_{jkl}^{(c)}\big),\\
\boldsymbol{d}^{(c+1)}&=\underset{\boldsymbol{d}}{\operatorname{argmin}}\ G^{(c+1)}(\boldsymbol{\pi}^{(c+1)},\boldsymbol{\Theta}^{(c+1)},\boldsymbol{d})+\frac{\gamma}{2}\sum_{j=1}^{J}\sum_{1\le k<l\le M}\big(d_{jkl}-(\theta_{jk}^{(c+1)}-\theta_{jl}^{(c+1)})+\mu_{jkl}^{(c)}\big),
\end{aligned}$$

$$\mu_{jkl}^{(c+1)} = \mu_{jkl}^{(c)} + d_{jkl}^{(c+1)} - \big(\theta_{jk}^{(c+1)} - \theta_{jl}^{(c+1)}\big),\ j = 1, ..., J, 1 \le k < l \le M.$$

Specifically, we get the following updates:

(1)

$$\pi_k^{(c+1)} = \frac{\sum_{i=1}^N s_{ik}^{(c+1)}/N - \tilde{\lambda}_1}{1 - M\tilde{\lambda}_1}, \quad \text{where } s_{ik}^{(c+1)} = \frac{\pi_k^{(c)} \varphi_k(\boldsymbol{R}_i; \boldsymbol{\Theta}_k^{(c)})}{\sum_{k'}^{(c)} \pi_{k'}^{(c)} \varphi_{k'}^{(c)}(\boldsymbol{R}_i; \boldsymbol{\theta}_{k'}^{(c)})},$$

(2)

$$\begin{aligned}\hat{\theta}_{jk}^{(c+1)} = \underset{\theta_{jk}}{\operatorname{argmin}} \Big\{ & - \frac{\sum_{i=1}^N s_{ik}^{(c)} R_{ij}}{N} \log \theta_{jk} - \frac{\sum_{i=1}^N s_{ik}^{(c)} (1 - R_{ij})}{N} \log(1 - \theta_{jk}) \\ & + \frac{\gamma}{2} \sum_{l>k} \big(\hat{d}_{jkl}^{(c)} - (\theta_{jk} - \hat{\theta}_{jl}^{(c)}) + \hat{\mu}_{jkl}^{(c)}\big)^2 \\ & + \frac{\gamma}{2} \sum_{l<k} \big(\hat{d}_{jlk}^{(c)} - (\hat{\theta_{jl}}^{(c+1)} - \theta_{jk}) + \hat{\mu}_{jlk}^{(c)}\big)^2 \Big\}\end{aligned}$$

(3)

$$\hat{d}_{jkl}^{(c+1)} = \begin{cases} \hat{\theta}_{jk}^{(c+1)} - \hat{\theta}_{jl}^{(c+1)} - \hat{\mu}_{jkl}^{(c)}, & \text{if } |\hat{d}_{jkl}^{(c)}| \ge \tau \\ \mathrm{ST}\big(\hat{\theta}_{jk}^{(c+1)} - \hat{\theta}_{jl}^{(c+1)} - \hat{\mu}_{jkl}^{(c)}; \tilde{\lambda}_2/\gamma\big), & \text{if } |\hat{d}_{jkl}^{(c)}| < \tau, \text{where } \mathrm{ST}(x;\gamma) = (|x| - \gamma)_+ x/|x| \end{cases},$$

(4)

$$\hat{\mu}_{jkl}^{(c+1)} = \hat{\mu}_{jkl}^{(c)} + \hat{d}_{jkl}^{(c+1)} - \big(\hat{\theta}_{jk}^{(c+1)} - \hat{\theta}_{jl}^{(c+1)}\big).$$

Note that the objective in step (2) is convex in θ_{jk}, therefore we use gradient descent to perform the minimization.

3 PEM Algorithm with Missing Values

In this section, we present the penalized EM algorithm with missing values. Here we use a mask matrix $M \in \{0, 1\}^{N \times J}$ to indicate the locations of the missing values, where $M_{i,j} = 0$ means the ith subject's response to the jth item is missing. The details of the algorithm is summarized in Algorithm 1.

Algorithm 1: Penalized EM with missing data

Data: Binary response matrix $\mathcal{R} = (R_{i,j})_{N\times J}$ and the mask matrix $\boldsymbol{M} = (M_{ij})_{N\times J}$ indicating missing values.

Set hyperparameters $\tilde{\lambda}_1,\ \tilde{\lambda}_2,\ \tau,\ \gamma$ and ρ.

Set an upper bound of the number of latent classes L.

Initialize parameters $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$, and the conditional expectations $\boldsymbol{s}$.

while *not converged* **do**

In the $(c+1)$th iteration,

for $(i,k) \in [N]\times[L]$ **do**

$$s_{ik}^{(c+1)} = \frac{\pi_k^{(c)}\varphi_k(\boldsymbol{R}_i;\boldsymbol{\theta}_k^{(c)})}{\sum_{k'}^{(c)}\pi_{k'}^{(c)}\varphi_{k'}^{(c)}(\boldsymbol{R}_i;\boldsymbol{\theta}_{k'}^{(c)})}, \quad \varphi(\boldsymbol{r}_i;\boldsymbol{\theta}_k) = \prod_{j=1}^{J}\left(\theta_{jk}^{R_{ij}}(1-\theta_{kj})^{1-R_{ij}}\right)^{m_{ij}}$$

for $k \in [L]$ *and* $\pi_k^{(c)} > \rho$ **do**

$$\pi_k^{(c+1)} = \frac{\sum_{i=1}^{N} s_{ik}^{(c+1)}/N - \tilde{\lambda}_1}{1-L\tilde{\lambda}_1}.$$

for $(j,k) \in [J]\times[L]$ *and* $\pi_k^{(c+1)} > \rho$ **do**

$$\begin{aligned}\theta_{jk}^{(c+1)} = \underset{\theta_{jk}}{\operatorname{argmin}}\Big\{ &-\frac{\sum_{i=1}^{N} s_{ik}^{(c)} R_{ij} m_{ij}}{\sum_{i=1}^{N} m_{ij}}\log\theta_{jk} - \frac{\sum_{i=1}^{N} s_{ik}^{(c)}(1-_{ij})m_{ij}}{\sum_{i=1}^{N} m_{ij}}\log(1-\theta_{jk}) \\ &+ \frac{\gamma}{2}\sum_{l>k}\left(\hat{d}_{jkl}^{(c)} - (\theta_{jk} - \hat{\theta}_{jl}^{(c)}) + \hat{\mu}_{jkl}^{(c)}\right)^2 \\ &+ \frac{\gamma}{2}\sum_{l<k}\left(\hat{d}_{jlk}^{(c)} - (\hat{\theta_{jl}}^{(c+1)} - \theta_{jk}) + \hat{\mu}_{jlk}^{(c)}\right)^2\Big\}\end{aligned}$$

for $j \in [J], k, l \in [L],\ k < l$ *and* $\pi_k^{(c+1)} > \rho,\ \pi_l^{(c+1)} > \rho$ **do**

$$\hat{d}_{jkl}^{(c+1)} = \begin{cases}\hat{\theta}_{jk}^{(c+1)} - \hat{\theta}_{jl}^{(c+1)} - \hat{\mu}_{jkl}^{(c)}, & \text{if } |\hat{d}_{jkl}^{(c)}| \geq \tau \\ \mathrm{ST}\big(\hat{\theta}_{jk}^{(c+1)} - \hat{\theta}_{jl}^{(c+1)} - \hat{\mu}_{jkl}^{(c)}; \tilde{\lambda}_2/\gamma\big), & \text{if } |\hat{d}_{jkl}^{(c)}| < \tau\end{cases},$$

$$\hat{\mu}_{jkl}^{(c+1)} = \hat{\mu}_{jkl}^{(c)} + \hat{d}_{jkl}^{(c+1)} - \big(\hat{\theta}_{jk}^{(c+1)} - \hat{\theta}_{jl}^{(c+1)}\big).$$

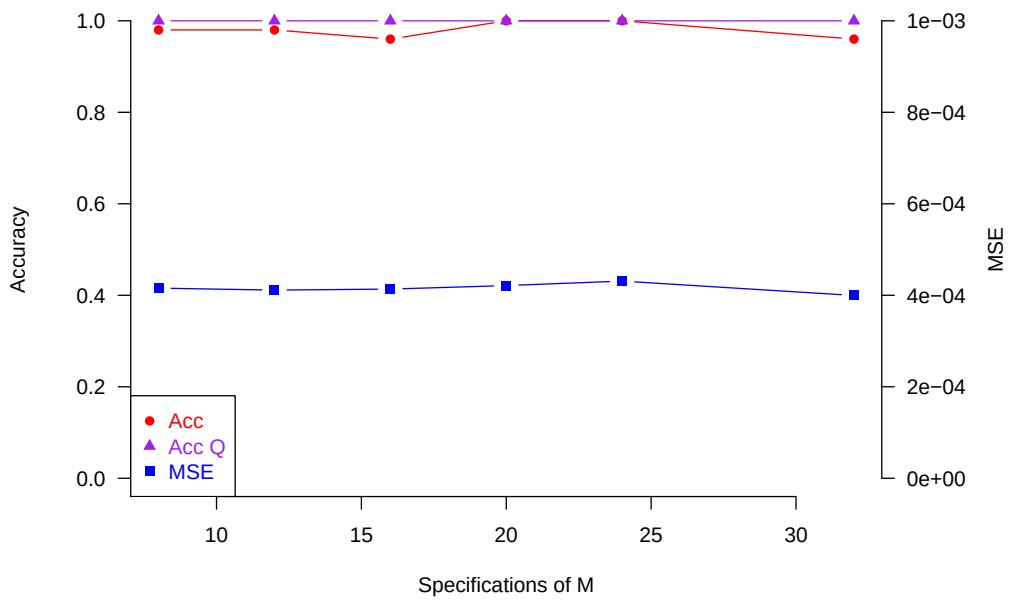
Output: $\left\{\hat{\boldsymbol{\pi}},\ \hat{\boldsymbol{\Theta}},\ \hat{\boldsymbol{s}}\right\}$

4 Sensitivity Analysis

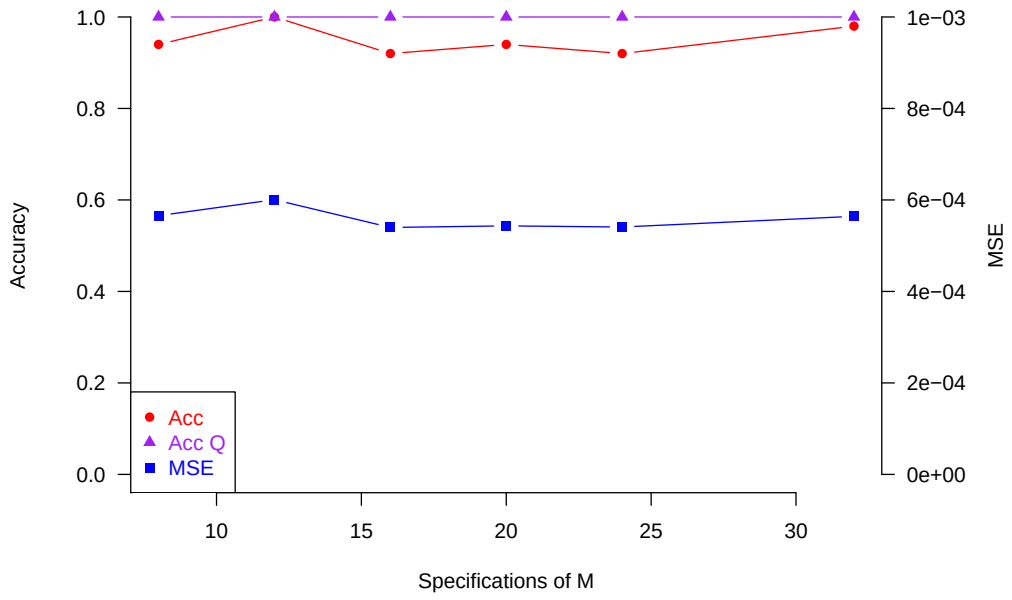
In this section, we conduct the sensitivity analysis of our algorithm by investigating the effects of different inputs of M, the upper bound of the number of latent classes, on the simulation results. In particular, we focus on two simulation settings: (1) DINA model with linear hierarchical structure, $N = 500$ and $r = 0.1$; (2) GDINA model with linear hierarchical structure, $N = 1000$ and $r = 0.1$. Both two settings have $K = 4$ latent attributes and $J = 30$ test items, and run

for 50 repetitions. We keep the parameter generation process and the hyperparameter tuning strategy consistent with the simulation studies in the main article. In this sensitivity analysis, we fit our proposed method with various $M = \{8, 12, 16, 20, 24, 32\}$ in the two simulations settings. The evaluation results in DINA and GDINA settings are based on metrics Acc($\hat{M}$), Acc($\hat{\boldsymbol{P}}$), Acc($\hat{\mathcal{E}}$), MSE($\hat{\boldsymbol{\Theta}}$) and Acc($\hat{\boldsymbol{Q}}$). Consistent with the simulation studies in the main article, the Acc($\hat{M}$), Acc($\hat{\boldsymbol{P}}$) and Acc($\hat{\mathcal{E}}$) are calculated for all the cases; MSE($\hat{\boldsymbol{\Theta}}$) is calculated for the cases when the number of latent classes is correctly selected; Acc($\hat{\boldsymbol{Q}}$) is calculated for the cases when the hierarchical structure is successfully recovered. The results are plotted in Figure 1.

From the simulation results in Figure 1, we see our proposed method is robust to the different specifications of M, in terms of all metrics. Among cases with different M, our method achieves a high accuracy in estimating the number of latent classes, and in recovering the partial orders, the hierarchical structures, the item parameter matrix, and the Q-matrix. In terms of computation time, the average running time is 0.36 seconds and 1.12 seconds for DINA and GDINA, respectively, per repetition per set of tuning hyperparameters.



(a)



(b)

Figure 1: Sensitivity analysis results. (a) DINA results; (b) GDINA results. The red curve captures the Acc($\hat{M}$), Acc($\hat{\boldsymbol{P}}$), Acc($\hat{\mathcal{E}}$), the blue curve captures MSE($\hat{\boldsymbol{\Theta}}$) and the purple curve captures the Acc($\hat{\boldsymbol{Q}}$) for various M.

References

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122.

Gu, Y. and Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, 20(2019).

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

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, 113(523):1284–1295.