Supplementary Appendix to "Identifying Latent Structures in Restricted Latent Class Models"

This supplementary appendix contains the proofs of the main results, including Theorem 1 and Propositions 1–3, in Section A.1, computational details in Section A.2, and additional simulation results in Section A.3.

A.1 Proofs of the Main Results

Proof of Theorem 1

To prove the identifiability result, directly working with $\mathrm{P}(\mathbf{R} = \mathbf{r} \mid Q, \Theta, \mathbf{p})$ is technically challenging. To better incorporate the induced restrictions by the Q-matrix, we consider a marginal probability approach as in Xu (2017). Although a similar technique is used, the problem settings and detailed arguments are significantly different from Xu (2017), which assumes Q-matrix is pre-specified, and it is more challenging to establish the identifiability of the Q-matrix. First, this is because estimation of the Q-matrix depends on the unknown model parameters, which themselves may not be identifiable without knowing the true Q-matrix. Second, the latent Q-matrix is binary and the discreteness nature of the problem makes the existing tools in Xu (2017) not directly applicable.

We introduce some notations. Define a T-matrix $T(Q, \Theta)$ as a $2^J \times 2^K$ matrix, where the entries are indexed by row index $\mathbf{r} \in \{0, 1\}^J$ and column index $\boldsymbol{\alpha}$. The $\mathbf{r} = (r_1, \cdots, r_J)^\top$th row and $\boldsymbol{\alpha}$th column element of $T(Q, \Theta)$, denoted by $t_{\mathbf{r},\boldsymbol{\alpha}}(Q, \Theta)$, is the marginal probability that a subject with attribute profile $\boldsymbol{\alpha}$ answers all items in subset $\{j : r_j = 1\}$ positively.

Thus $t_{\mathbf{r},\boldsymbol{\alpha}}(Q,\Theta)$ is the marginal probability that, given $Q,\Theta,\boldsymbol{\alpha}$, the random response $\mathbf{R}\succeq\mathbf{r}$, i.e.,

$$t_{\mathbf{r},\boldsymbol{\alpha}}(Q,\Theta)=\mathrm{P}(\mathbf{R}\succeq\mathbf{r}\mid Q,\Theta,\boldsymbol{\alpha}).$$

When $\mathbf{r}=0_{2^K}$,

$$t_{0,\boldsymbol{\alpha}}(Q,\Theta)=\mathrm{P}(\mathbf{R}\succeq 0)=1 \text{ for any } \boldsymbol{\alpha};$$

and for any $\mathbf{r}\neq 0$,

$$t_{\mathbf{r},\boldsymbol{\alpha}}(Q,\Theta)=\prod_{j:r_j=1}\mathrm{P}(R_j=r_j\mid Q,\Theta,\boldsymbol{\alpha})=\sum_{\mathbf{r}'\succeq\mathbf{r}}\mathrm{P}(\mathbf{R}=\mathbf{r}'\mid Q,\Theta,\boldsymbol{\alpha}).$$

In particular, for $\mathbf{r}=\mathbf{e}_j$ with $1\leq j\leq J$,

$$t_{\mathbf{e}_j,\boldsymbol{\alpha}}(Q,\Theta)=\mathrm{P}(R_j=1\mid Q,\Theta,\boldsymbol{\alpha})=\theta_{j,\boldsymbol{\alpha}}.$$

Let $T_{\mathbf{r},\star}(Q,\Theta)$ be the row vector corresponding to $\mathbf{r}$. Then we know that for $j=1,\cdots,J$, $T_{\mathbf{e}_j,\star}(Q,\Theta)=\Theta_{j,\star}$. In addition, for any $\mathbf{r}\neq 0$, we can write

$$T_{\mathbf{r},\star}(Q,\Theta)=\bigodot_{j:r_j=1}T_{\mathbf{e}_j,\star}(Q,\Theta), \tag{A.1}$$

where $\odot$ is the element-wise product of the row vectors. By definition, multiplying the T-matrix by the distribution of attribute profiles $\mathbf{p}$ results in a vector containing the marginal probabilities of successfully answering each subset of items correctly. The $\mathbf{r}$th entry of this vector is

$$T_{\mathbf{r},\star}(Q,\Theta)\mathbf{p}=\sum\nolimits_{\boldsymbol{\alpha}}t_{\mathbf{r},\boldsymbol{\alpha}}(Q,\Theta)p_{\boldsymbol{\alpha}}=\mathrm{P}(\mathbf{R}\succeq\mathbf{r}\mid Q,\Theta,\mathbf{p}).$$

There is a one-to-one mapping between the T-matrix and the vectors $\mathrm{P}(\mathbf{R}=\mathbf{r}\mid Q,\Theta,\mathbf{p})$,

$\mathbf{r} \in \{0,1\}^J$. Therefore, to show the identifiability of Q, we only need to prove that if

$$T(Q,\Theta)\mathbf{p} = T(\bar{Q},\bar{\Theta})\bar{\mathbf{p}}, \tag{A.2}$$

then we must have $Q \sim \bar{Q}$. We follow this argument and present the proof in the following four steps.

Step 1. Without loss of generality, we arrange the rows of Q such that it takes the form of condition C1. For notational convenience, we write $t_{\mathbf{e}_j,\boldsymbol{\alpha}}(Q,\Theta)$ and $t_{\mathbf{e}_j,\boldsymbol{\alpha}}(\bar{Q},\bar{\Theta})$ as $t_{\mathbf{e}_j,\boldsymbol{\alpha}}$ and $\bar{t}_{\mathbf{e}_j,\boldsymbol{\alpha}}$, respectively. By the definition of the T-matrix, $t_{\mathbf{e}_j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}}$ and $\bar{t}_{\mathbf{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{j,\boldsymbol{\alpha}}$ for any $j \in \{1,\cdots,J\}$ and $\boldsymbol{\alpha} \in \{0,1\}^K$.

Consider the $2^K \times 2^K$ T-matrix, $T(Q_{1:K,\star},\Theta_{1:K,\star})$, where $Q_{1:K,\star}$ is submatrix of Q with the first K rows and $\bar{\Theta}_{1:K}$ is the submatrix of Θ with the first K rows. Under C1, we know $Q_{1:K,\star} = \mathcal{I}_K$. Take

$$\tilde{\theta} = (\theta_{1,\mathbf{e}_1},\cdots,\theta_{K,\mathbf{e}_K})^\top.$$

Under (2), the row transformed T-matrix $T(Q_{1:K,\star},\Theta_{1:K} - \tilde{\theta}1^\top)$ takes an upper-left triangular form (up to column swapping) with nonzero diagonal elements as follows:

$$\begin{array}{c|cccccc}
 & \boldsymbol{\alpha} = 0_{2^K} & \mathbf{e}_1 & \cdots & \sum_{k=2}^K \mathbf{e}_k & 1_{2^K} \\
\hline
0_{2^K} & 1 & 1 & \cdots & 1 & 1 \\
\mathbf{e}_1 & (t_{\mathbf{e}_1,0} - \theta_{1,\mathbf{e}_1}) & 0 & \cdots & (t_{\mathbf{e}_1,\sum_{k=2}^K \mathbf{e}_k} - \theta_{1,\mathbf{e}_1}) & 0 \\
\vdots & \vdots & \vdots & \iddots & 0 & 0 \\
\sum_{k=2}^K \mathbf{e}_k & \prod_{k=2}^K (t_{\mathbf{e}_k,0} - \theta_{k,\mathbf{e}_k}) & \prod_{k=2}^K (t_{\mathbf{e}_k,\mathbf{e}_1} - \theta_{k,\mathbf{e}_k}) & 0 & 0 & 0 \\
1_{2^K} & \prod_{k=1}^K (t_{\mathbf{e}_k,0} - \theta_{k,\mathbf{e}_k}) & 0 & 0 & 0 & 0
\end{array}_{2^K \times 2^K}.$$

Therefore $T(Q_{1:K,\star},\Theta_{1:K} - \tilde{\theta}1^\top)$ is of full rank 2^K, which implies the full rank of $T(Q_{1:K,\star},\Theta_{1:K})$ by Proposition 3 of Xu (2017). Similarly we know $T(Q_{(K+1):(2K),\star},\Theta_{(K+1):(2K),\star})$ also has full

rank 2^K.

From above, for any attribute profile $\boldsymbol{\alpha}$, we can choose a 2^K-dimensional vector $u_{\boldsymbol{\alpha}}$ such that

$$u_{\boldsymbol{\alpha}}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star}) = (0^\top, \underbrace{1}_{column\ \boldsymbol{\alpha}}, 0^\top). \tag{A.3}$$

We next show $T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star})$ and $T(\bar{Q}_{(K+1):(2K),\star}, \bar{\Theta}_{(K+1):(2K),\star})$ also have full rank 2^K. Without loss of generality, we focus on $T(\bar{Q}_{(K+1):(2K),\star}, \bar{\Theta}_{(K+1):(2K),\star})$. If it is not of full rank, then there exists a 2^K-dimensional vector $x \neq 0$ such that

$$x^\top \cdot T(\bar{Q}_{(K+1):(2K),\star}, \bar{\Theta}_{(K+1):(2K),\star}) = 0^\top.$$

Since $T(Q_{(K+1):(2K),\star}, \Theta_{(K+1):(2K),\star})$ is of full rank, there must exist a nonzero element of vector $x^\top \cdot T(Q_{(K+1):(2K),\star}, \Theta_{(K+1):(2K),\star})$; we denote the corresponding column index as $\boldsymbol{\alpha}^*$. Then, for a vector $u_{\boldsymbol{\alpha}^*}$ chosen as in (A.3), we have

$$[\{u_{\boldsymbol{\alpha}^*}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star})\} \odot \{x^\top \cdot T(Q_{(K+1):(2K),\star}, \Theta_{(K+1):(2K),\star})\}] \cdot \mathbf{p} \neq 0$$

while

$$[\{u_{\boldsymbol{\alpha}^*}^\top \cdot T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star})\} \odot \{x^\top \cdot \bar{T}(Q_{(K+1):(2K),\star}, \bar{\Theta}_{(K+1):(2K),\star})\}] \cdot \bar{\mathbf{p}} = 0,$$

which contradicts Equation (A.2). This proves the full rank of $T(\bar{Q}_{(K+1):(2K),\star}, \bar{\Theta}_{(K+1):(2K),\star})$.

For $u_{\boldsymbol{\alpha}}$ in (A.3), $u_{\boldsymbol{\alpha}}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star}) \cdot \mathbf{p} \neq 0$. From Equation (A.2),

$$u_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star}) \cdot \bar{\mathbf{p}} \neq 0.$$

Therefore, there must exist at least one nonzero element in $u_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star})$; we index one such column with $\pi(\boldsymbol{\alpha})$ and write the nonzero value as $b_{u\boldsymbol{\alpha}}$. Then we know $b_{u\boldsymbol{\alpha}} \neq 0$ and $\bar{p}_{\pi(\boldsymbol{\alpha})} \neq 0$. Furthermore, there exists a set of $[\pi(\boldsymbol{\alpha}); \boldsymbol{\alpha} \in \{0,1\}^K]$ such that it is a one-to-one mapping from $[\boldsymbol{\alpha}; \boldsymbol{\alpha} \in \{0,1\}^K]$ to $[\pi(\boldsymbol{\alpha}); \boldsymbol{\alpha} \in \{0,1\}^K]$ due to the result that both

$T(Q_{1:K,\star}, \Theta_{1:K,\star})$ and $T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star})$ are of full rank. We choose this set of $\pi(\boldsymbol{\alpha})$'s in the following proof.

We further choose a vector $v_{\boldsymbol{\alpha}}$ such that

$$v_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star}) = (0^\top, \underbrace{1}_{column\ \pi(\boldsymbol{\alpha})}, 0^\top).$$

It follows that

$$\begin{aligned} &\{u_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star})\} \odot \{v_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star})\} \\ &\qquad = (0^\top, \underbrace{b_{u,\boldsymbol{\alpha}}}_{column\ \pi(\boldsymbol{\alpha})}, 0^\top) \neq 0^\top. \end{aligned} \tag{A.4}$$

Therefore, (A.4) $\cdot\, \bar{\mathbf{p}} = b_{u\boldsymbol{\alpha}} \bar{p}_{\pi(\boldsymbol{\alpha})} \neq 0$. From Equation (A.2),

$$[\{u_{\boldsymbol{\alpha}}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star})\} \odot \{v_{\boldsymbol{\alpha}}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\}] \cdot \mathbf{p} \neq 0;$$

therefore, the column $\boldsymbol{\alpha}$ of $v_{\boldsymbol{\alpha}}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star})$ is nonzero and we denote the value as $b_{v\boldsymbol{\alpha}}$. In particular, we have

$$\begin{aligned} &\{u_{\boldsymbol{\alpha}}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star})\} \odot \{v_{\boldsymbol{\alpha}}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\} \\ &\qquad = (0^\top, \underbrace{b_{v,\boldsymbol{\alpha}}}_{column\ \boldsymbol{\alpha}}, 0^\top) \neq 0^\top. \end{aligned} \tag{A.5}$$

Then for any item $j > 2K$, under Q and model parameters $(\Theta, \mathbf{p})$,

$$\begin{aligned} &T_{\mathbf{e}_j,\star}(Q, \Theta) \odot \{u_{\boldsymbol{\alpha}}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star})\} \\ &\odot \{v_{\boldsymbol{\alpha}}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\} \quad = \quad (0^\top, \underbrace{t_{\mathbf{e}_j,\boldsymbol{\alpha}} b_{v,\boldsymbol{\alpha}}}_{column\ \boldsymbol{\alpha}}, 0^\top). \end{aligned} \tag{A.6}$$

Similarly, under $\bar{Q}$ and model parameters $(\bar{\Theta}, \bar{\mathbf{p}})$,

$$\begin{aligned} & T_{\mathbf{e}_j,\star}(\bar{Q}, \bar{\Theta}) \odot \{u_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star})\} \\ & \quad \odot \{v_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star})\} \quad = \quad (0^\top, \underbrace{\bar{t}_{\mathbf{e}_j,\pi(\boldsymbol{\alpha})} b_{u,\boldsymbol{\alpha}}}_{column\ \pi(\boldsymbol{\alpha})}, \ 0^\top); \end{aligned} \tag{A.7}$$

From Equation (A.2), we have $(\text{A.5}) \cdot \mathbf{p} = (\text{A.4}) \cdot \bar{\mathbf{p}} \neq 0$ and $(\text{A.6}) \cdot \mathbf{p} = (\text{A.7}) \cdot \bar{\mathbf{p}}$. Therefore,

$$t_{\mathbf{e}_j,\boldsymbol{\alpha}} = \{(\text{A.6}) \cdot \mathbf{p}\}/\{(\text{A.5}) \cdot \mathbf{p}\} = \{(\text{A.7}) \cdot \bar{\mathbf{p}}\}/\{(\text{A.4}) \cdot \bar{\mathbf{p}}\} = \bar{t}_{\mathbf{e}_j,\pi(\boldsymbol{\alpha})}.$$

From assumption C2, we know $\pi(0) = 0$. Furthermore, for $\boldsymbol{\alpha} \succeq \boldsymbol{\alpha}'$ and $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$, the column $\boldsymbol{\alpha}'$ of $v_{\boldsymbol{\alpha}}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})$ must be zero. Since otherwise, from a similar argument as above, we have $\{u_{\boldsymbol{\alpha}'}^\top \cdot T(Q_{1:K,\star}, \Theta_{1:K,\star})\} \odot \{v_{\boldsymbol{\alpha}}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\}$ has only one nonzero element and it corresponds to column $\boldsymbol{\alpha}'$; and $\{u_{\boldsymbol{\alpha}'}^\top \cdot T(\bar{Q}_{1:K,\star}, \bar{\Theta}_{1:K,\star})\} \odot \{v_{\boldsymbol{\alpha}}^\top \cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star})\}$ has only one nonzero element and it corresponds to column $\pi(\boldsymbol{\alpha})$. Then for any $j > 2K$, from Equation (A.2), we would have

$$t_{\mathbf{e}_j,\boldsymbol{\alpha}'} = \bar{t}_{\mathbf{e}_j,\pi(\boldsymbol{\alpha})} = t_{\mathbf{e}_j,\boldsymbol{\alpha}},$$

which contradicts assumption C2.

Step 2. Assumption C2 implies that for any $h \in \{1, \ldots, K\}$, there is a 2^K-dimensional vector m_h such that

$$m_h^\top \cdot T(Q_{(2K+1):J,\star}, \Theta_{(2K+1):J,\star}) = (\underbrace{0}_{column\ 0}, *, \ldots, *, \underbrace{1}_{column\ \mathbf{e}_h}, *, \ldots, *),$$

where '$*$' denotes some unspecified value. From Step 1, we know

$$m_h^\top \cdot T(\bar{Q}_{(2K+1):J,\star}, \bar{\Theta}_{(2K+1):J,\star}) = (\underbrace{0}_{column\ 0}, *, \ldots, *, \underbrace{1}_{column\ \pi(\mathbf{e}_h)}, *, \ldots, *).$$

Also for the v vector chosen in Step 1, we know the $\mathbf{e}_h$th element of $v_{\mathbf{e}_h}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})$ must be nonzero. For simplicity, we write the nonzero value as b_{vh}.

We choose a J-dimensional vector

$$\theta^* \;=\; \big(t_{\mathbf{e}_1,\mathbf{e}_1}, \ldots, t_{\mathbf{e}_{h-1},\mathbf{e}_{h-1}}, 0, t_{\mathbf{e}_{h+1},\mathbf{e}_{h+1}}, \ldots, t_{\mathbf{e}_K,\mathbf{e}_K},\; 0_{J-K}^\top \big)^\top.$$

From Proposition 3 of Xu (2017), (A.2) implies

$$T(Q, \Theta - \theta^* 1^\top)\mathbf{p} = T(\bar{Q}, \bar{\Theta} - \theta^* 1^\top)\bar{\mathbf{p}}.$$

In addition,

$$\begin{aligned} & T_{\sum_{k=1}^K \mathbf{e}_k - \mathbf{e}_h, \star}(Q, \Theta - \theta^* 1^\top) \\ = \; & \Big\{ \underbrace{\prod_{\substack{k=1,\ldots,K;\\ k\neq h}} (t_{\mathbf{e}_k,0} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ 0},\; 0^\top,\; \underbrace{\prod_{\substack{k=1,\ldots,K;\\ k\neq h}} (t_{\mathbf{e}_k,\mathbf{e}_h} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_h},\; 0^\top \Big\}. \end{aligned} \tag{A.8}$$

Therefore, we have the following two equations:

$$\begin{aligned} & T_{\sum_{k=1}^K \mathbf{e}_k - \mathbf{e}_h, \star}(Q, \Theta - \theta^* 1^\top) \odot \{m_h^\top \cdot T(Q_{(2K+1):J,\star}, \Theta_{(2K+1):J,\star})\} \\ & \odot \{v_{\mathbf{e}_h}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\} \;=\; \Big\{0^\top,\; \underbrace{b_{vh} \cdot \prod_{\substack{k=1,\ldots,K;\\ k\neq h}} (t_{\mathbf{e}_k,\mathbf{e}_h} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_h},\; 0^\top \Big\}; \end{aligned} \tag{A.9}$$

$$\begin{aligned} & T_{\sum_{k=1}^K \mathbf{e}_k, \star}(Q, \Theta - \theta^* 1^\top) \odot \{m_h^\top \cdot T(Q_{(2K+1):J,\star}, \Theta_{(2K+1):J,\star})\} \\ & \odot \{v_{\mathbf{e}_h}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\} \;=\; \Big\{0^\top,\; \underbrace{t_{\mathbf{e}_h,\mathbf{e}_h} \cdot b_{vh} \cdot \prod_{\substack{k=1,\ldots,K;\\ k\neq h}} (t_{\mathbf{e}_k,\mathbf{e}_h} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_h},\; 0^\top \Big\}. \end{aligned} \tag{A.10}$$

Similarly for $(\bar{Q}, \bar{\Theta}, \bar{\mathbf{p}})$, we have equations:

$$T_{\sum_{k=1}^{K}\mathbf{e}_k-\mathbf{e}_h,\star}(\bar{Q}, \bar{\Theta}-\theta^*1^\top)\odot\{m_h^\top\cdot T(\bar{Q}_{(2K+1):J,\star}, \bar{\Theta}_{(2K+1):J,\star})\} \tag{A.11}$$
$$\odot\{v_{\mathbf{e}_h}^\top\cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star})\} \;=\; \Big\{0^\top,\; \underbrace{\prod_{\substack{k=1,\ldots,K;\\ k\neq h}}(\bar{t}_{\mathbf{e}_k,\mathbf{e}_h}-t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \pi(\mathbf{e}_h)},\; 0^\top\Big\};$$

$$T_{\sum_{k=1}^{K}\mathbf{e}_k,\star}(\bar{Q}, \bar{\Theta}-\theta^*1^\top)\odot\{m_h^\top\cdot T(\bar{Q}_{(2K+1):J,\star}, \bar{\Theta}_{(2K+1):J,\star})\} \tag{A.12}$$
$$\odot\{v_{\mathbf{e}_h}^\top\cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star})\} \;=\; \Big\{0^\top,\; \underbrace{\bar{t}_{\mathbf{e}_h,\pi(\mathbf{e}_h)}\cdot\prod_{\substack{k=1,\ldots,K;\\ k\neq h}}(\bar{t}_{\mathbf{e}_k,\mathbf{e}_h}-t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \pi(\mathbf{e}_h)},\; 0^\top\Big\}.$$

Under C2, (A.9) $\neq 0$. From Equation (A.2), we have (A.10) $\cdot\,\mathbf{p}$ = (A.12) $\cdot\,\bar{\mathbf{p}}$ and (A.9) $\cdot\,\mathbf{p}$ = (A.11) $\cdot\,\bar{\mathbf{p}} \neq 0$. Therefore, for $1 \le h \le K$,

$$t_{\mathbf{e}_h,\mathbf{e}_h} = \{(\text{A.10})\cdot\mathbf{p}\}/\{(\text{A.9})\cdot\mathbf{p}\} = \{(\text{A.12})\cdot\bar{\mathbf{p}}\}/\{(\text{A.11})\cdot\bar{\mathbf{p}}\} = \bar{t}_{\mathbf{e}_h,\pi(\mathbf{e}_h)}$$

Similarly, $t_{\mathbf{e}_{K+h},\mathbf{e}_h} = \bar{t}_{\mathbf{e}_{K+h},\pi(\mathbf{e}_h)}$.

Furthermore, we can choose a vector m_0 such that the following equations hold:

$$m_0^\top\cdot T(Q_{(2K+1):J,\star}, \Theta_{(2K+1):J,\star}) = (\underbrace{1}_{column\ 0}, *, \ldots, *, \quad \underbrace{0}_{column\ \mathbf{e}_h}, \; *, \ldots, *);$$

$$m_0^\top\cdot T(\bar{Q}_{(2K+1):J,\star}, \bar{\Theta}_{(2K+1):J,\star}) = (\underbrace{1}_{column\ 0}, *, \ldots, *, \quad \underbrace{0}_{column\ \pi(\mathbf{e}_h)}, \; *, \ldots, *).$$

Then a similar argument gives $t_{\mathbf{e}_h,0} = \bar{t}_{\mathbf{e}_h,0}$ for any $h \in \{1, \ldots, 2K\}$.

Step 3. For vector $v_{\mathbf{e}_h+\mathbf{e}_j}$ chosen as in Step 1, we know

$$v_{\mathbf{e}_h+\mathbf{e}_j}^\top\cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star}) = (0^\top, \quad \underbrace{1}_{column\ \pi(\mathbf{e}_h+\mathbf{e}_j)}, \; 0^\top).$$

From Step 1, we know the $(\mathbf{e}_h + \mathbf{e}_j)$th element of $v_{\mathbf{e}_h+\mathbf{e}_j}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})$ must be nonzero; for simplicity, we denote the value as b_{vhj}. In addition, in the last paragraph of Step 1 we proved that the $\mathbf{e}_h$th and $\mathbf{e}_j$th elements of $v_{\mathbf{e}_h+\mathbf{e}_j}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})$ are zero.

For any $j \neq h$ such that $1 \leq j \leq K$, we take a J-dimensional vector

$$\theta^* = \big(\, t_{\mathbf{e}_1,\mathbf{e}_1}, \ldots, t_{\mathbf{e}_{h-1},\mathbf{e}_{h-1}}, t_{\mathbf{e}_h,0}, t_{\mathbf{e}_{h+1},\mathbf{e}_{h+1}}, \ldots, t_{\mathbf{e}_{j-1},\mathbf{e}_{j-1}}, t_{\mathbf{e}_j,0}, t_{\mathbf{e}_{j+1},\mathbf{e}_{j+1}}, \ldots, t_{\mathbf{e}_K,\mathbf{e}_K}, 0_{J-K}^\top \, \big)^\top.$$

We have

$$\begin{aligned}
& T_{\sum_{k=1}^K \mathbf{e}_k - \mathbf{e}_j, \star}(Q, \Theta - \theta^* 1^\top) \\
= \; & \Big\{ 0^\top, \underbrace{(t_{\mathbf{e}_h,\mathbf{e}_h} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K,\\ k\neq h,j}} (t_{\mathbf{e}_k,\mathbf{e}_h} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_h}, 0^\top, \underbrace{(t_{\mathbf{e}_h,\mathbf{e}_j} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K,\\ k\neq h,j}} (t_{\mathbf{e}_k,\mathbf{e}_j} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_j}, \\
& \underbrace{0^\top, (t_{\mathbf{e}_h,\mathbf{e}_h+\mathbf{e}_j} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K,\\ k\neq h,j}} (t_{\mathbf{e}_k,\mathbf{e}_h+\mathbf{e}_j} - t_{\mathbf{e}_k,\mathbf{e}_k}),\ 0^\top}_{column\ \mathbf{e}_h+\mathbf{e}_j} \Big\}.
\end{aligned}$$

Therefore for $(Q, \Theta, \mathbf{p})$, we have

$$\begin{aligned}
& T_{\sum_{k=1}^K \mathbf{e}_k - \mathbf{e}_j, \star}(Q, \Theta - \theta^* 1^\top) \odot \{ v_{\mathbf{e}_h+\mathbf{e}_j}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star}) \} \\
= & \underbrace{\Big\{ 0^\top, b_{vhj} \cdot (t_{\mathbf{e}_h,\mathbf{e}_h+\mathbf{e}_j} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K,\\ k\neq h,j}} (t_{\mathbf{e}_k,\mathbf{e}_h+\mathbf{e}_j} - t_{\mathbf{e}_k,\mathbf{e}_k}),\ 0^\top \Big\}}_{column\ \mathbf{e}_h+\mathbf{e}_j}
\end{aligned}$$

and

$$
\begin{aligned}
&T_{\sum_{k=1}^{K}\mathbf{e}_k,\star}(Q,\Theta-\theta^*1^\top)\odot\{v_{\mathbf{e}_h+\mathbf{e}_j}^\top\cdot T(Q_{(K+1):2K,\star},\Theta_{(K+1):2K,\star})\}\\
=&\Bigg\{0^\top,\underbrace{(t_{\mathbf{e}_j,\mathbf{e}_h+\mathbf{e}_j}-t_{\mathbf{e}_j,0})\cdot b_{vhj}\cdot(t_{\mathbf{e}_h,\mathbf{e}_h+\mathbf{e}_j}-t_{\mathbf{e}_h,0})\prod_{\substack{k=1,\cdots,K,\\ k\neq h,j}}(t_{\mathbf{e}_k,\mathbf{e}_h+\mathbf{e}_j}-t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_h+\mathbf{e}_j},\ 0^\top\Bigg\}.
\end{aligned}
$$

Similarly for $(\bar{Q},\bar{\Theta},\bar{\mathbf{p}})$, we have

$$
\begin{aligned}
&T_{\sum_{k=1}^{K}\mathbf{e}_k-\mathbf{e}_j,\star}(\bar{Q},\bar{\Theta}-\theta^*1^\top)\odot\{v_{\mathbf{e}_h+\mathbf{e}_j}^\top\cdot T(\bar{Q}_{(K+1):2K,\star},\bar{\Theta}_{(K+1):2K,\star})\}\\
=&\Bigg\{0^\top,\underbrace{(\bar{t}_{\mathbf{e}_h,\pi(\mathbf{e}_h+\mathbf{e}_j)}-t_{\mathbf{e}_h,0})\prod_{\substack{k=1,\cdots,K,\\ k\neq h,j}}(\bar{t}_{\mathbf{e}_k,\pi(\mathbf{e}_h+\mathbf{e}_j)}-t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \pi(\mathbf{e}_h+\mathbf{e}_j)},\ 0^\top\Bigg\};
\end{aligned}
$$

and

$$
\begin{aligned}
&\{v_{\mathbf{e}_h+\mathbf{e}_j}^\top\cdot T(\bar{Q}_{(K+1):2K,\star},\bar{\Theta}_{(K+1):2K,\star})\}\odot T_{\sum_{k=1}^{K}\mathbf{e}_k,\star}(\bar{Q},\bar{\Theta}-\theta^*1^\top)\\
=&\Bigg\{0^\top,\underbrace{(\bar{t}_{\mathbf{e}_j,\pi(\mathbf{e}_h+\mathbf{e}_j)}-t_{\mathbf{e}_j,0})(\bar{t}_{\mathbf{e}_h,\pi(\mathbf{e}_h+\mathbf{e}_j)}-t_{\mathbf{e}_h,0})\prod_{\substack{k=1,\cdots,K,\\ k\neq h,j}}(\bar{t}_{\mathbf{e}_k,\pi(\mathbf{e}_h+\mathbf{e}_j)}-t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \pi(\mathbf{e}_h+\mathbf{e}_j)},\ 0^\top\Bigg\}.
\end{aligned}
$$

From (A.2), the above equations imply

$$
t_{\mathbf{e}_j,\mathbf{e}_h+\mathbf{e}_j}=\bar{t}_{\mathbf{e}_j,\pi(\mathbf{e}_h+\mathbf{e}_j)}
$$

for $1\le h,j\le K$ and $j\neq h$. Similarly $t_{\mathbf{e}_{K+j},\mathbf{e}_h+\mathbf{e}_j}=\bar{t}_{\mathbf{e}_{K+j},\pi(\mathbf{e}_h+\mathbf{e}_j)}$. Furthermore, a similar argument implies for any $\boldsymbol{\alpha}\succeq\mathbf{e}_j$ and $\boldsymbol{\alpha}\neq\mathbf{e}_j$,

$$
t_{\mathbf{e}_j,\boldsymbol{\alpha}}=\bar{t}_{\mathbf{e}_j,\pi(\boldsymbol{\alpha})}\text{ and }t_{\mathbf{e}_{K+j},\boldsymbol{\alpha}}=\bar{t}_{\mathbf{e}_{K+j},\pi(\boldsymbol{\alpha})}.
$$

Step 4. Consider any j and h such that $K+1\le j\le 2K$ and $1\le h\le K$. Take a

J-dimensional vector

$$\theta^* = \left(t_{\mathbf{e}_1,\mathbf{e}_1}, \ldots, t_{\mathbf{e}_{h-1},\mathbf{e}_{h-1}}, t_{\mathbf{e}_h,0}, t_{\mathbf{e}_{h+1},\mathbf{e}_{h+1}}, \ldots, t_{\mathbf{e}_K,\mathbf{e}_K},\ 0^\top_{J-K} \right)^\top.$$

From Step 3, we have the following equations:

$$T_{\sum_{k=1}^K \mathbf{e}_k,\star}(Q, \Theta - \theta^* 1^\top) = \Bigg\{ 0^\top, \underbrace{(t_{\mathbf{e}_h,\mathbf{e}_h} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K;\\ k\neq h}} (t_{\mathbf{e}_k,\mathbf{e}_h} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_h},\ 0^\top \Bigg\};$$

$$T_{\mathbf{e}_j+\sum_{k=1}^K \mathbf{e}_k,\star}(Q, \Theta - \theta^* 1^\top) = \Bigg\{ 0^\top, \underbrace{t_{\mathbf{e}_j,\mathbf{e}_h}(t_{\mathbf{e}_h,\mathbf{e}_h} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K;\\ k\neq h}} (t_{\mathbf{e}_k,\mathbf{e}_h} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \mathbf{e}_h},\ 0^\top \Bigg\};$$

$$T_{\sum_{k=1}^K \mathbf{e}_k,\star}(\bar{Q}, \bar{\Theta} - \theta^* 1^\top\} = \Bigg\{ 0^\top, \underbrace{(\bar{t}_{\mathbf{e}_h,\pi(\mathbf{e}_h)} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K;\\ k\neq h}} (\bar{t}_{\mathbf{e}_k,\pi(\mathbf{e}_h)} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \pi(\mathbf{e}_h)},\ 0^\top \Bigg\};$$

$$T_{\mathbf{e}_j+\sum_{k=1}^K \mathbf{e}_k,\star}(\bar{Q}, \bar{\Theta} - \theta^* 1^\top) = \Bigg\{ 0^\top, \underbrace{\bar{t}_{\mathbf{e}_j,\pi(\mathbf{e}_h)}(\bar{t}_{\mathbf{e}_h,\pi(\mathbf{e}_h)} - t_{\mathbf{e}_h,0}) \prod_{\substack{k=1,\cdots,K;\\ k\neq h}} (\bar{t}_{\mathbf{e}_k,\pi(\mathbf{e}_h)} - t_{\mathbf{e}_k,\mathbf{e}_k})}_{column\ \pi(\mathbf{e}_h)},\ 0^\top \Bigg\}.$$

Then Equation (A.2) implies $t_{\mathbf{e}_j,\mathbf{e}_h} = \bar{t}_{\mathbf{e}_j,\pi(\mathbf{e}_h)}$. Similarly $t_{\mathbf{e}_{j-K},\mathbf{e}_h} = \bar{t}_{\mathbf{e}_{j-K},\pi(\mathbf{e}_h)}$.

For any j, h_1 and h_2 such that $K+1 \leq j \leq 2K$ and $1 \leq h_1 \neq h_2 \leq K$, take

$$\theta^* = (\theta^*_1, \cdots, \theta^*_K, 0^\top_{J-K})$$

such that $\theta_i^* = t_{\mathbf{e}_i,\mathbf{e}_i}$ for $i \notin \{h_1, h_2\}$ and $\theta_i^* = t_{\mathbf{e}_i,0}$ for $i \in \{h_1, h_2\}$. Then we have

$$
\begin{aligned}
& t_{\mathbf{e}_j,\mathbf{e}_{h_1}+\mathbf{e}_{h_2}} \\
= & \frac{[T_{\mathbf{e}_j+\sum_{k=1}^K \mathbf{e}_{k,\star}}(Q, \Theta - \theta^* 1^\top) \odot \{v_{\mathbf{e}_{h_1}+\mathbf{e}_{h_2}}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\}]\mathbf{p}}{[T_{\sum_{k=1}^K \mathbf{e}_{k,\star}}(Q, \Theta - \theta^* 1^\top) \odot \{v_{\mathbf{e}_{h_1}+\mathbf{e}_{h_2}}^\top \cdot T(Q_{(K+1):2K,\star}, \Theta_{(K+1):2K,\star})\}]\mathbf{p}} \\
= & \frac{[T_{\mathbf{e}_j+\sum_{k=1}^K \mathbf{e}_{k,\star}}(\bar{Q}, \bar{\Theta} - \theta^* 1^\top) \odot \{v_{\mathbf{e}_{h_1}+\mathbf{e}_{h_2}}^\top \cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star})\}]\bar{\mathbf{p}}}{[T_{\sum_{k=1}^K \mathbf{e}_{k,\star}}(\bar{Q}, \bar{\Theta} - \theta^* 1^\top) \odot \{v_{\mathbf{e}_{h_1}+\mathbf{e}_{h_2}}^\top \cdot T(\bar{Q}_{(K+1):2K,\star}, \bar{\Theta}_{(K+1):2K,\star})\}]\bar{\mathbf{p}}} \\
= & \bar{t}_{\mathbf{e}_j,\pi(\mathbf{e}_{h_1}+\mathbf{e}_{h_2})}.
\end{aligned}
$$

Similarly, $t_{\mathbf{e}_{j-K},\mathbf{e}_{h_1}+\mathbf{e}_{h_2}} = \bar{t}_{\mathbf{e}_{j-K},\pi(\mathbf{e}_{h_1}+\mathbf{e}_{h_2})}$. Furthermore, a similar argument implies for any $\boldsymbol{\alpha} \not\succeq \mathbf{e}_j$, $t_{\mathbf{e}_j,\boldsymbol{\alpha}} = \bar{t}_{\mathbf{e}_j,\pi(\boldsymbol{\alpha})}$ and $t_{\mathbf{e}_{j-K},\boldsymbol{\alpha}} = \bar{t}_{\mathbf{e}_{j-K},\pi(\boldsymbol{\alpha})}$.

The results in Steps 1–4 imply $T(Q, \Theta)$ is the same as $T(\bar{Q}, \bar{\Theta})$ up to the chosen mapping $\pi(\cdot)$. This also indicates that such a one-to-one mapping $\pi(\cdot)$ is unique under (2). This further implies $Q \sim \bar{Q}$.

Proof of Proposition 1

The proof follows from a similar argument in Shen et al (2012). We first define a complexity measure for a family of probability density (mass) functions $\mathcal{F}$. Let $H(\cdot, \mathcal{F})$ be the bracketing Hellinger metric entropy of $\mathcal{F}$. Specifically, for a finite set of pairs of functions $S(\epsilon, n) = \{(f_1^l, f_1^u), \cdots, (f_n^l, f_n^u)\}$, it is called a Hellinger ϵ-bracketing of $\mathcal{F}$ if the L_2-norm $\|(f_i^l)^{1/2} - (f_i^u)^{1/2}\| \leq \epsilon$ for any $i = 1 \cdots, n$, and for any $f \in \mathcal{F}$, there exist an i such that $f_i^l \leq f \leq f_i^u$. Then the bracketing Hellinger metric entropy of $H(\cdot, \mathcal{F})$ is defined to be the logarithm of the cardinality of the ϵ-bracketing with the smallest size, i.e., $H(\cdot, \mathcal{F}) = \log \min\{n : S(\epsilon, n)\}$.

To show the consistency result, we apply Theorem 2 in Shen et al (2012) by checking the required technical conditions.

First, we need to show the size of the parameter space is well controlled under the Hellinger metric. Specifically, we show that for some constant c, any $\epsilon < 1$ and any t such that $\epsilon/2^4 < t \leq \epsilon$, $H(t, \mathcal{B}_S) \leq c \log(J2^K)|S| \log(2\epsilon/t)$, where S and $\mathcal{B}_S = \mathcal{F}_S \cap \{h(\eta, \eta_0) \leq$

$2\epsilon\}$ are defined as follows. The $\eta = (B, \mathbf{p})$ denotes model parameters B and $\mathbf{p}$, and η_0 denotes the true model parameters. The $h(\eta, \eta_0)$ is defined as the Hellinger distance between the two probability distribution mass functions $P(\mathbf{R} \mid \eta)$ and $P(\mathbf{R} \mid \eta_0)$ under η and η_0, respectively. The S is an an index set indicating the nonzero elements of parameter vector B and the cardinality of S, denoted by $|S|$, is $\leq M_0$, where M_0 denotes the number of nonzero coefficients in the true parameter vector B_0. The set $\mathcal{F}_S$ is the set of square root probability mass functions $\{P^{1/2}(\mathbf{R} \mid \eta) : B_{S^c} = 0\}$ under the sparsity structure introduced by S^c.

To prove the above result, we only need to show that $H(t, \mathcal{B}_S) = O(1)\log(2\epsilon/t)$ uniformly for any S, ϵ and t. For any η_1 and η_2 in the small neighborhood $\mathcal{B}_S$ of η_0, we write

$$\begin{aligned} h^2(\eta_1, \eta_2) &= \sum_{\mathbf{r} \in \{0,1\}^J} [P^{1/2}(\mathbf{R} = \mathbf{r} \mid \eta_1) - P^{1/2}(\mathbf{R} = \mathbf{r} \mid \eta_2)]^2 \\ &= \sum_{\mathbf{r}} \Big\{ \Big[\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{1,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_1) \Big]^{1/2} - \Big[\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{2,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_2) \Big]^{1/2} \Big\}^2 \\ &= \sum_{\mathbf{r}} \left\{ \frac{\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{1,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_1) - \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{2,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_2)}{\Big[\sum_{\boldsymbol{\alpha}} p_{1,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_1) \Big]^{1/2} + \Big[\sum_{\boldsymbol{\alpha}} p_{2,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_2) \Big]^{1/2}} \right\}^2. \end{aligned}$$

Under the assumption of C3, the denominator terms in the above display are bounded and therefore,

$$h^2(\eta_1, \eta_2) = \Theta(1) \sum_{\mathbf{r}} \left\{ \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{1,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_1) - \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{2,\boldsymbol{\alpha}} P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_2) \right\}^2.$$

Here we write $a_N = \Theta(1)$ if $0 < \liminf_{N \to \infty} a_N \leq \limsup_{N \to \infty} a_n < \infty$. Under the identifiability conditions,

$$h^2(\eta_1, \eta_2) = \Theta(1) \|T(B_1)\mathbf{p}_1 - T(B_2)\mathbf{p}_2\| = \Theta(1)\|\eta_1 - \eta_2\|^2.$$

This implies that $H(t, \mathcal{B}_A)$ is controlled by the size of the corresponding local parameter

space of η under the L_2 norm. We know that the t-bracketing entropy under the L_2 norm is controlled by $O(\log(\epsilon/t))$; therefore, $H(t, \mathcal{B}_A) = O(1)\log(2\epsilon/t)$ holds.

We further verify the following condition, which indicates that the truncated L_1 penalty approximates the L_0 function well enough. For the truncated L_1 penalty, define $B_{\tau+} = \{\beta_{j,k_1\cdots k_h} I(\beta_{j,k_1\cdots k_h} \geq \tau),\ \text{for any } 1 \leq j \leq J,\ 1 \leq h \leq K \text{ and } 1 \leq k_1 < \cdots < k_h \leq K\}$ and $\eta_{\tau+} = \{B_{\tau+}, \mathbf{p}\}$. We aim to prove the following inequality hols for small τ and some constants $d_1, d_2, d_3 > 0$,

$$-\log\{1 - h^2(\eta, \eta_0)\} \geq -d_1 \log\{1 - h^2(\eta_{\tau,+}, \eta_0)\} - d_3(J2^K)\tau^{d_2}, \tag{A.13}$$

which is the Assumption B in Shen et al. (2012). To show (A.13), note that

$$\begin{aligned}
& |h^2(\eta, \eta_0) - h^2(\eta_{\tau,+}, \eta_0)| \\
\leq\ & \sum_{\mathbf{r}} |P^{1/2}(\mathbf{R} = \mathbf{r} \mid \eta) - P^{1/2}(\mathbf{R} = \mathbf{r} \mid \eta_{\tau,+})| \\
& \times |P^{1/2}(\mathbf{R} = \mathbf{r} \mid \eta) + P^{1/2}(\mathbf{R} = \mathbf{r} \mid \eta_{\tau,+}) - 2P^{1/2}(\mathbf{R} = \mathbf{r} \mid \eta_0)| \\
=\ & O(1) \sum_{\mathbf{r}} |P(\mathbf{R} = \mathbf{r} \mid \eta) - P(\mathbf{R} = \mathbf{r} \mid \eta_{\tau,+})| \\
=\ & O(1) \sum_{\mathbf{r}} \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} |P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta) - P(\mathbf{R} = \mathbf{r} \mid \boldsymbol{\alpha}, \eta_{\tau,+})| \\
=\ & O(1) J 2^K \tau
\end{aligned}$$

where $O(1)$ does not depend on τ and the above result holds uniformly for η in a neighborhood of η_0. Further note that for small τ,

$$|\log\{1 - h^2(\eta, \eta_0)\} - \log\{1 - h^2(\eta_{\tau,+}, \eta_0)\}| \leq O(1)|h^2(\eta, \eta_0) - h^2(\eta_{\tau,+}, \eta_0)|.$$

Therefore, (A.13) holds with $d_1 = d_2 = 1$ and d_3 is taken as some big constant.

Another key condition that is used in Shen et al (2012) is that

$$\inf_{\{\eta:|S|\leq M_0, S\nsim S_0\}} \frac{-\log\{1-h^2(\eta,\eta_0)\}}{\max\{|S_0\backslash S|,1\}} \geq d_0 \frac{\log(J2^K)}{N}, \tag{A.14}$$

where d_0 is a constant and $|S_0\backslash S|$ denotes the size of the set difference between S_0 and S. Note that $\inf_{\{\eta:|S|\leq M_0, S\nsim S_0\}} \frac{-\log\{1-h^2(\eta,\eta_0)\}}{\max\{|S_0\backslash S|,1\}} \geq (2M_0)^{-1} \inf_{\{\eta:|S|\leq M_0, S\nsim S_0\}} h^2(\eta,\eta_0)$. From the proof of the identifiability result (Theorem 1), we can obtain that under the identifiability conditions, there exists a constant $\delta > 0$ such that $\inf_{\{\eta:|S|\leq M_0, S\nsim S_0\}} h^2(\eta,\eta_0) \geq \delta$. Therefore, (A.14) holds.

With the above conditions satisfied, we apply Theorem 2 in Shen et al. (2012) and obtain that for all small τ, there exists a constant such that

$$P(\hat{B} \neq \hat{B}_0) \leq \exp\{-c_1 N + c_2\};$$

this further implies that

$$P(\hat{Q} \nsim Q_0) \leq \exp\{-c_1 N + c_2\}$$

and $\sqrt{N}(\hat{\eta}_{S_0} - \eta_{0,S_0})$ weakly converges to the limiting distribution of the oracle maximum likelihood estimator $\sqrt{N}(\hat{\eta}_{0,S_0} - \eta_{0,S_0})$, which is a normal distribution with mean zero and covariance being the inverse of the Fisher information matrix.

Proof of Proposition 2

From the proof of the identifiability result, we can obtain the consistency of the model parameter estimators that $\|\hat{\eta} - \eta_0\| \to 0$ in probability. In particular, when $N^{-1/2}\lambda \to 0$, the likelihood-based estimators $\hat{\eta}$ satisfies

$$\sum_{\mathbf{r}} \left| N^{-1} \sum_{i=1}^{N} I(\mathbf{R}_i = \mathbf{r}) - \mathrm{P}(\mathbf{R} = \mathbf{r} \mid \hat{\eta}) \right| \to 0.$$

By the law of large number,

$$N^{-1}\sum_{i=1}^{N} I(\mathbf{R}_i = \mathbf{r}) \to \mathrm{P}(\mathbf{R} = \mathbf{r} \mid \eta_0)$$

almost surely for any response vector $\mathbf{r}$ as the sample size $N \to \infty$. Therefore we have $\sum_{\mathbf{r}} |\mathrm{P}(\mathbf{R} = \mathbf{r} \mid \eta_0) - \mathrm{P}(\mathbf{R} = \mathbf{r} \mid \hat{\eta})| \to 0$ almost surely. Then the proof of the identifiability result can be applied and we have the consistency of the model parameters up to column swapping.

We further show the convergence rate of $\hat{\eta} - \eta_0$ is $N^{-1/2}$. Under the condition that $N^{-1/2}\lambda \to 0$, for η in a small neighborhood of η_0,

$$\mathbb{M}_N(\eta; \mathcal{R}) := \frac{1}{N}\Big[- l_N(\eta; \mathcal{R}) + \lambda \sum_{j=1}^{J} \sum_{\substack{1 \le h \le K \\ 1 \le k_1 < \cdots < k_h \le K}} J_\tau(\beta_{j,k_1 \cdots k_h})\Big]$$

converges uniformly to the same limit of $-\frac{1}{N} l_N(\eta; \mathcal{R})$ by the uniform law of large number. We use $\mathbb{M}_0(\eta)$ to denote the limit process, which is the expectation of the negative log-likelihood of a single observation. By Taylor's expansion and C3, $\mathbb{M}_0(\eta) - \mathbb{M}_0(\eta_0) = \Theta(\|\eta - \eta\|^2)$. Furthermore, Taylor's expansion implies that for sufficiently small δ,

$$E \sup_{\|\eta - \eta\| \le \delta} |\mathbb{M}_N(\eta; \mathcal{R}) - \mathbb{M}_N(\eta_0; \mathcal{R}) - \mathbb{M}_0(\eta) + \mathbb{M}_0(\eta_0)| = O(\delta N^{-1/2}).$$

Therefore, Theorem 3.2.5 in van der Vaart and Wellner (1996) gives the convergence rate $\hat{\eta} - \eta_0 = O_p(N^{-1/2})$.

We next show the normality result. We reparameterize $\eta = \eta_0 + N^{-1/2}u$ for some real vector u. Then for any true $\beta_{0;j,k_1 \cdots k_h} = 0$ and the corresponding u element $u_{j,k_1 \cdots k_h}$, $J_\tau(\beta_{0;j,k_1 \cdots k_h} + N^{-1/2}u_{j,k_1 \cdots k_h}) - J_\tau(\beta_{0;j,k_1 \cdots k_h}) = \min\{N^{-1/2}\lambda\tau^{-1}|u_{j,k_1 \cdots k_h}|, \lambda\}$ diverges to ∞ under the assumption that $\lambda \to \infty$ and $N^{-1/2}\lambda\tau^{-1} \to \infty$. For any $\beta_{0;j,k_1 \cdots k_h} \neq 0$ and the corresponding $u_{j,k_1 \cdots k_h}$, we have $J_\tau(\beta_{0;j,k_1 \cdots k_h} + N^{-1/2}u_{j,k_1 \cdots k_h}) - J_\tau(\beta_{0;j,k_1 \cdots k_h}) \to 0$ in probability. Therefore, $N[\mathbb{M}_N(\eta_0 + N^{-1/2}u; \mathcal{R}) - \mathbb{M}_N(\eta_0; \mathcal{R})]$ converges to the limit distribution of

$-l_N(\eta_0+N^{-1/2}u;\mathcal{R})+l_N(\eta_0;\mathcal{R})$ if $u_{j,k_1\cdots k_h}=0$ for any $\beta_{0;j,k_1\cdots k_h}=0$, and diverges otherwise. Following the epi-convergence results of Geyer (1994), we then have the asymptotic normality of $N^{-1/2}(\hat{\eta}_{S_0}-\eta_{0,S_0})$ and its asymptotic equivalent to the distribution of $N^{-1/2}(\hat{\eta}_{0,S_0}-\eta_{0,S_0})$.

We further show the selection consistency. From the above result, if $\beta_{0;j,k_1\cdots k_h}\neq 0$, then $\hat{\beta}_{j,k_1\cdots k_h}\to\beta_{0;j,k_1\cdots k_h}\neq 0$ in probability. On the other hand, if $\beta_{0;j,k_1\cdots k_h}=0$ and $\hat{\beta}_{j,k_1\cdots k_h}\neq 0$, by the Karush-Kuhn-Tucker (KKT) conditions and under the assumptions in the proposition, we know $N^{-1/2}\partial l_N(\eta;\mathcal{R})/\partial\beta_{j,k_1\cdots k_h}|_{\eta=\hat{\eta}}=N^{-1/2}\lambda/\tau\to\infty$ in probability; however, this contradict the fact that $N^{-1/2}\partial l_N(\eta;\mathcal{R})/\partial\beta_{j,k_1\cdots k_h}|_{\eta=\hat{\eta}}=O_p(1)$. Therefore, we have if $\beta^0_{j,k_1\cdots k_h}=0$, $\hat{\beta}^0_{j,k_1\cdots k_h}=0$ in probability. This gives $P(\hat{S}\nsim S_0)\to 0$ and $P(\hat{Q}\nsim Q_0)\to 0$.

Proof of Proposition 3

We prove the consistency of the IC procedure to select the true Q-matrix. Note that $\hat{S}_{(\lambda,\tau)}$ is an index set of a finite dimensional vector $\hat{B}_{(\lambda,\tau)}$ and therefore there are finite number of possible S's. Further from Proposition 2, we know there exists one pair of (λ_N,τ_N) such that $P(\hat{S}_{(\lambda_N,\tau_N)}\sim S_0)\to 1$. Therefore to prove the consistency, we only need to show that

$$IC(S_0,c_N)<IC(S,c_N)$$

for any $S\nsim S_0$

This follows from the classic argument of proving the consistency of an information criterion (e.g., Nishii, 1988). First consider the case when S_0 is a subset of S. From the identifiability result, we have the consistency of the model parameter estimators that $\|\hat{\eta}^*_S-\eta_0\|\to 0$ in probability. Therefore using Taylor's expansion, we can obtain that $\hat{\eta}^*_S-\eta_0=O_p(N^{-1/2})$ and $l_N(\hat{\eta}^*_S;\mathcal{R})-l_N(\eta_0;\mathcal{R})=O_p(1)$. This further implies that

$$l_N(\hat{\eta}^*_S;\mathcal{R})-l_N(\hat{\eta}_0;\mathcal{R})=O_p(1)$$

and therefore $IC(S, c_N) - IC(S_0, c_N) \to \infty$ if $c_N \to \infty$.

On the other hand, if the index set S_0 is not a subset of S, by the proof of the identifiability theorem, there exists some $\delta > 0$ such that

$$N^{-1}\{l_N(\hat{\eta}_0; \mathcal{R}) - l_N(\hat{\eta}_S^*; \mathcal{R})\} = \Theta(1)\|T(\hat{B}_S^*)\hat{\mathbf{p}}_S^* - T(Q, \hat{B}_0)\hat{\mathbf{p}}_0\| \geq \delta.$$

This implies that $IC(S, c_N) - IC(S_0, c_N) \to \infty$ if $c_N/N \to 0$. This completes the proof.

A.2 Computation method

The log-likelihood function $l_N(B, \mathbf{p}; \mathcal{R})$ in Equation (7) is the marginal log-likelihood taking the form

$$l_N(B, \mathbf{p}; \mathcal{R}) = \sum_{i=1}^{N} \log \left\{ \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{R_{ij}} (1 - \theta_{j,\boldsymbol{\alpha}})^{(1-R_{ij})} \right\},$$

where $\theta_{j,\boldsymbol{\alpha}}$ is a function of $\boldsymbol{\beta}_j$ specified by the diagnostic model assumption. The log-likelihood $l_N(B, \mathbf{p}; \mathcal{R})$ involves the summation over the latent variables $\boldsymbol{\alpha}$ and direct optimization of it over B is computationally challenging.

Instead we use the EM algorithm. In particular, the log-likelihood of the complete data, denoted by $(\mathcal{R}, \mathbf{A}) = (\mathbf{R}_i, \boldsymbol{\alpha}_i; i = 1, \cdots, n)$, as a function of $B = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J)$ is:

$$l_N(B; \mathcal{R}, \mathbf{A}) = \sum_{i=1}^{N} \sum_{j=1}^{J} \big[R_{ij} \log \theta_{j,\boldsymbol{\alpha}_i} + (1 - R_{ij}) \log(1 - \theta_{j,\boldsymbol{\alpha}_i}) \big].$$

In the E-step, we compute the expectation of $l_N(B; \mathcal{R}, \mathbf{A})$ with respect to the posterior distributions of $\boldsymbol{\alpha}_i$, $i = 1, \cdots, N$. Let $(B^*, \mathbf{p}^*)$ be the parameter values in the previous step estimation (or the initial value in first step). For the ith subject, we write the posterior distribution of $\boldsymbol{\alpha}_i$ as $P^*_{i\boldsymbol{\alpha}} = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha} \mid \mathbf{R}_i, B^*, \mathbf{p}^*)$, for any $\boldsymbol{\alpha} \in \{0,1\}^K$. Then the conditional expectation of the complete data log-likelihood $l_N(B, \mathbf{p}; \mathcal{R}, \mathbf{A})$ is:

$$\begin{aligned} Q(B \mid \mathcal{R}, B^*, \mathbf{p}^*) &:= \sum_{j=1}^{J} Q_j(\boldsymbol{\beta}_j \mid \mathcal{R}, B^*, \mathbf{p}^*) \\ &:= \sum_{i=1}^{N} \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} P^*_{i\boldsymbol{\alpha}} \sum_{j=1}^{J} \big[R_{ij} \log \theta_{j,\boldsymbol{\alpha}} + (1 - R_{ij}) \log(1 - \theta_{j,\boldsymbol{\alpha}}) \big]. \end{aligned}$$

In the M-step, we find the maximizer of the TLP regularized $Q(B \mid \mathcal{R}, B^*, \mathbf{p}^*)$, i.e., $-Q(B \mid \mathcal{R}, B^*, \mathbf{p}^*) + \lambda \sum_{j=1}^{J} \sum_{\substack{1 \le h \le K \\ 1 \le k_1 < \cdots < k_h \le K}} J_\tau(\beta_{j,k_1 \cdots k_h})$. This can be done by minimizing each $\boldsymbol{\beta}_j$

independently with respect to the following objective function

$$\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} H_j(\boldsymbol{\beta}_j), \tag{A.15}$$

where

$$H_j(\boldsymbol{\beta}_j) := -Q_j(\boldsymbol{\beta}_j \mid \mathcal{R}, B^*, \mathbf{p}^*) + \lambda \sum_{\substack{1\le h\le K \\ 1\le k_1<\cdots<k_h\le K}} J_\tau(\beta_{j,k_1\cdots k_h}).$$

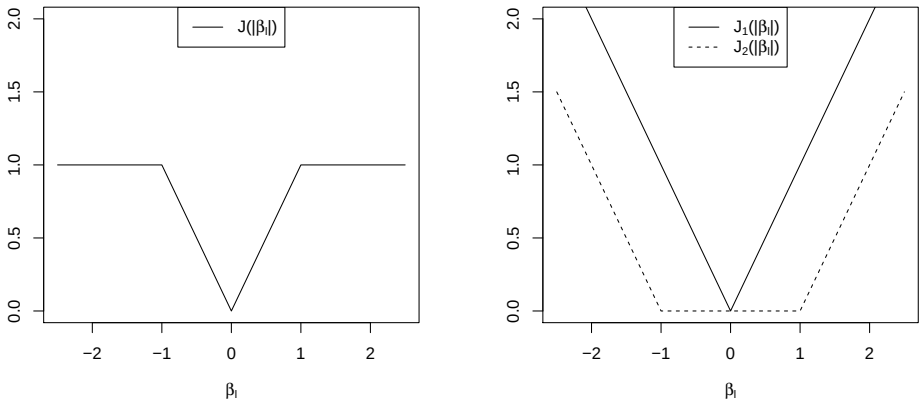

Figure A.1: Truncated L_1 function and difference of convex decomposition when $\tau = 1$

To solve the optimization problem in (A.15), we first implement the Difference of Convex (DC) decompositions for the nonconvex penalty function $J_\tau(\cdot)$ as shown in the right plot of Figure A.1, i.e., $J_\tau(\beta) = J_{1,\tau}(\beta) - J_{2,\tau}(\beta)$ where $J_{1,\tau}(\beta) = \frac{|\beta|}{\tau}$ and $J_{2,\tau}(\beta) = \max(\frac{|\beta|}{\tau} - 1, 0)$. We then apply the algorithm proposed by Shen et al. (2012) and write the objective function $H_j(\boldsymbol{\beta}_j)$ as the difference of two convex functions $H_j(\boldsymbol{\beta}_j) = H_{j,1}(\boldsymbol{\beta}_j) - H_{j,2}(\boldsymbol{\beta}_j)$ by defining

$$H_{j,1}(\boldsymbol{\beta}_j) = -Q_j(\boldsymbol{\beta}_j \mid \mathcal{R}, B^*, \mathbf{p}^*) + \lambda \sum_{\substack{1\le h\le K \\ 1\le k_1<\cdots<k_h\le K}} J_{1,\tau}(\beta_{j,k_1\cdots k_h})$$

and

$$H_{j,2}(\boldsymbol{\beta}) = \lambda \sum_{\substack{1\le h\le K \\ 1\le k_1<\cdots<k_h\le K}} J_{2,\tau}(\beta_{j,k_1\cdots k_h}).$$

We start the iterative minimization of $H_j(\boldsymbol{\beta}_j)$ by first solving the initial Lasso problem $\hat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta}_j} H_{j,1}(\boldsymbol{\beta}_j)$. Then at the $m \geq 1$ step, we solve for the following weighted Lasso problem:

$$\hat{\boldsymbol{\beta}}_j^{(m)} = \arg\min_{\boldsymbol{\beta}_j} \Big\{ -Q_j(\boldsymbol{\beta}_j \mid \mathcal{R}, B^*, \mathbf{p}^*) + \frac{\lambda}{\tau} \sum_{\substack{1\leq h\leq K \\ 1\leq k_1<\cdots<k_h\leq K}} |\beta_{j,k_1\cdots k_h}| I(|\hat{\beta}^{(m-1)}_{j,k_1\cdots k_h}| \leq \tau) \Big\}.$$

The update is repeated until the objective function converges. To solve the initial Lasso and the weighted Lasso problems in the maximization step, we use the coordinate descent method similar to the algorithm solving regularized generalized linear models discussed by Friedman et al. (2010). We will then check the KKT conditions, and if they are not satisfied we go back and update the entire parameter again. In the M-step, we update $\boldsymbol{\beta}_j$ from $j = 1$ to J sequentially and the EM update is performed until convergence.

With the above EM algorithm, we can get a set of candidate estimators $\hat{B} = (\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_J)$ for a set of tuning parameters (τ, λ). For each $\hat{B}$, we obtain the corresponding candidate Q-matrix. Then we refit these Q-restricted models and select the Q-matrix that has the minimal BIC value as the final estimator.

Remark 1 (Initial values selection)**.** *Under the saturated model setting without restrictions of the Q-matrix, the number of item parameters is $J \times 2^K$. In this case the EM algorithm may stops at some local maximum especially when the penalty is large. One solution to this problem is to use different starting points. However, the method becomes less efficient and often requires much computation time when the number of attributes K becomes larger. Alternatively, we propose a fast pre-screening method to get a reasonable starting point. We choose a set of tuning parameter λ's and conduct the initial screening in the following steps:*

1. *For each λ, solve the following L_1 regularized likelihood for the additive LCDM (main*

effect model) using the above EM algorithm, i.e.,

$$(\hat{\boldsymbol{\beta}}_1, ...\hat{\boldsymbol{\beta}}_J) = \arg\min_{\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_J} -\sum_{i=1}^{N}\sum_{\boldsymbol{\alpha}\in\{0,1\}^K}\left\{p_{\boldsymbol{\alpha}}\prod_{j=1}^{J}\theta_{j,\boldsymbol{\alpha}}^{R_{ij}}(1-\theta_{j,\boldsymbol{\alpha}})^{1-R_{ij}}\right\} + \lambda\sum_{j=1}^{J}\sum_{k=1}^{K}|\beta_{j,k}|;$$

where $\text{logit}(\theta_{j,\boldsymbol{\alpha}}) = \beta_{j0} + \sum_{k=1}^{K}\beta_{jk}\alpha_k$. *Note the total number of item parameters of this model is only* $J \times (K+1)$.

2. *For the candidate Q-matrices* $\hat{Q}(\lambda)$, *choose the Q-matrix with least number of 1's and no all-0 rows or columns. Then we refit the Q-restricted model under the selected Q via EM and use the estimated parameters as starting value of the proposed method.*

Note that, though fast and straightforward, the above method cannot be used to estimate Q-matrix directly and usually will not be consistent, especially when the true model has large interaction effects like the DINA model. However, we find it useful in providing good initial values by separating those items with fewer attributes from those with multiple ones. In practice, the method can also be used to provide multiple initial points by repeatedly fitting resampled data.

A3 Additional Simulation Results

A3.1 Additional Simulation for Section 5.1

Estimation results of the parameters $\Theta = (\theta_{j,\boldsymbol{\alpha}})_{J\times 2^K}$ are presented in Figure A.2 with $K \in \{3, 4, 5\}$, $\rho = 0$ and $N = 1000$. For the TLP method, the $\hat{\theta}$'s are calculated from the refitted $\hat{\beta}$ values under the estimated model structure. For the true model, the θ's are estimated under the true Q-matrix and the true diagnostic model assumption. We report the box plots of the squared-root mean squared errors (RMSEs) of $\hat{\theta}$'s. Figure A.2 and Table 3 in the main text show that the proposed method gives similar estimation results to those under the true model, which is consistent with the theoretical results in Propositions 1 and 2.

A3.2 Additional Simulation for Section 5.2

Table A.1 presents the simulation results for $K = 3$ and misspecification rate is 20%. The cases with higher dimension of latent attributes $K = 4$ and $K = 5$ are presented in Table A.2. In particular, we use the following design with $\rho = 0$:

$$\begin{pmatrix} K=4 \\ K=5 \end{pmatrix} \bigotimes \begin{pmatrix} \text{DINA} \\ \text{LCDM} \end{pmatrix} \bigotimes \begin{pmatrix} 10\% \text{ Misspecification} \\ 20\% \text{ Misspecification} \end{pmatrix} \bigotimes \begin{pmatrix} N=1000 \\ N=2000 \end{pmatrix}.$$

We can see that the proposed method performs well and similar conclusions to the case with $K = 3$ and missepcification level being 10% can be obtained.

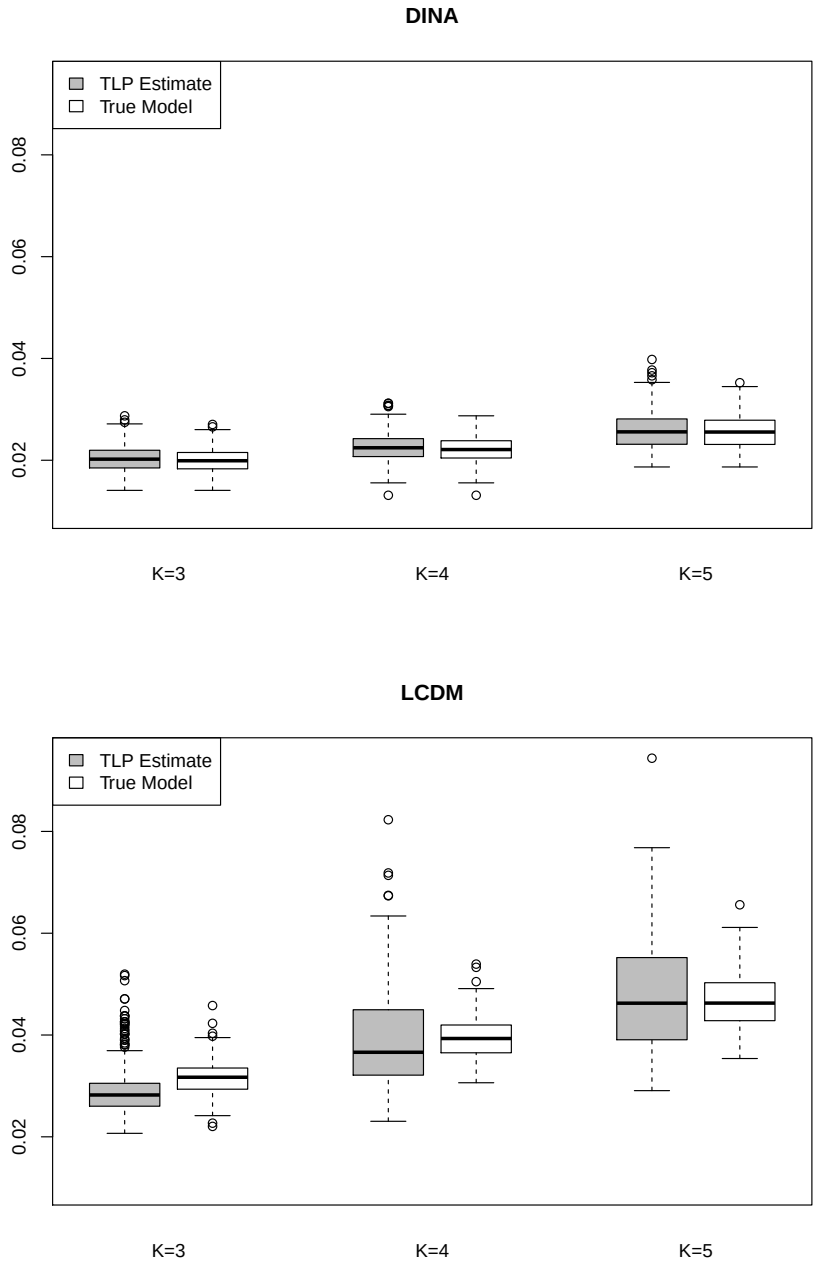

Figure A.2: Box plots of RMSEs of $\hat{\theta}$'s.

			DINA					LCDM				
				Entry		Vector			Entry		Vector	
ρ	N		Total	TPR	FPR	TPR	FPR	Total	TPR	FPR	TPR	FPR
0	500	Proposed	0.997 (0.994)	0.992 (0.988)	0.000 (0.001)	0.985 (0.975)	0.001 (0.002)	0.951 (0.969)	0.886 (0.937)	0.007 (0.005)	0.833 (0.900)	0.020 (0.014)
		GMDI	0.842	0.223	0.000	0.210	0.000	0.833	0.185	0.000	0.168	0.001
	1000	Proposed	0.998 (0.997)	0.993 (0.992)	0.000 (0.000)	0.990 (0.988)	0.000 (0.001)	0.988 (0.991)	0.958 (0.983)	0.000 (0.002)	0.938 (0.970)	0.000 (0.004)
		GMDI	0.856	0.298	0.000	0.282	0.001	0.855	0.290	0.000	0.278	0.001
	2000	Proposed	0.997 (0.996)	0.991 (0.990)	0.000 (0.001)	0.985 (0.983)	0.000 (0.001)	0.996 (0.995)	0.992 (0.991)	0.001 (0.001)	0.980 (0.980)	0.000 (0.002)
		GMDI	0.862	0.315	0.000	0.310	0.000	0.863	0.323	0.000	0.312	0.000
0.15	500	Proposed	0.995 (0.995)	0.985 (0.986)	0.000 (0.001)	0.975 (0.978)	0.001 (0.001)	0.950 (0.965)	0.876 (0.927)	0.006 (0.005)	0.825 (0.877)	0.019 (0.013)
		GMDI	0.840	0.214	0.000	0.205	0.001	0.835	0.202	0.000	0.180	0.001
	1000	Proposed	0.999 (0.998)	0.996 (0.994)	0.000 (0.000)	0.993 (0.988)	0.000 (0.000)	0.990 (0.995)	0.969 0.987)	0.000 (0.001)	0.953 (0.975)	0.001 (0.001)
		GMDI	0.852	0.294	0.001	0.270	0.003	0.853	0.284	0.000	0.270	0.001
	2000	Proposed	0.998 (0.996)	0.994 (0.992)	0.000 (0.001)	0.990 (0.985)	0.000 (0.002)	0.999 (0.998)	0.998 (0.994)	0.000 (0.000)	0.995 (0.988)	0.000 (0.000)
		GMDI	0.860	0.312	0.001	0.307	0.003	0.862	0.307	0.000	0.310	0.000
0.25	500	Proposed	0.995 (0.993)	0.986 (0.983)	0.001 (0.001)	0.978 (0.973)	0.001 (0.002)	0.954 (0.965)	0.887 (0.922)	0.005 (0.004)	0.833 (0.877)	0.016 (0.013)
		GMDI	0.840	0.215	0.001	0.210	0.003	0.832	0.193	0.001	0.170	0.003
	1000	Proposed	1.000 (0.998)	0.999 (0.993)	0.000 (0.000)	0.998 (0.988)	0.000 (0.000)	0.986 (0.992)	0.957 (0.983)	0.001 (0.001)	0.932 (0.968)	0.001 (0.002)
		GMDI	0.851	0.282	0.001	0.270	0.004	0.851	0.271	0.001	0.268	0.003
	2000	Proposed	0.998 (0.998)	0.994 (0.994)	0.000 (0.000)	0.990 (0.990)	0.000 (0.001)	0.998 (0.998)	0.996 (0.994)	0.000 (0.000)	0.990 (0.988)	0.000 (0.000)
		GMDI	0.860	0.301	0.001	0.305	0.002	0.861	0.305	0.000	0.307	0.001

Table A.1: High misspecification rate detection results with $K = 3$. For the proposed method, results after the first $J_0 = 4$ steps are presented in brackets. "Total" is the proportion of correctly estimated items with the initial baseline 0.8. "TPR" is true positive rate and "FPR" is the false positive rate.

Low Misspecification

Settings			DINA					LCDM				
				Entry		Vector			Entry		Vector	
K	N		Total	TPR	FPR	TPR	FPR	Total	TPR	FPR	TPR	FPR
4	1000	Proposed	0.997 (0.996)	0.984 (0.984)	0.000 (0.001)	0.975 (0.975)	0.001 (0.002)	0.987 (0.988)	0.966 (0.960)	0.002 (0.001)	0.930 (0.915)	0.007 (0.004)
		GMDI	0.923	0.263	0.000	0.235	0.000	0.919	0.225	0.000	0.195	0.000
	2000	Proposed	0.996 (0.995)	0.976 (0.976)	0.000 (0.001)	0.960 (0.960)	0.001 (0.001)	0.997 (0.996)	0.985 (0.988)	0.000 (0.001)	0.965 (0.965)	0.000 (0.001)
		GMDI	0.930	0.360	0.000	0.305	0.000	0.925	0.292	0.000	0.255	0.000
5	1000	Proposed	0.994 (0.994)	0.965 (0.981)	0.000 (0.001)	0.940 (0.950)	0.000 (0.002)	0.977 (0.981)	0.907 (0.942)	0.002 (0.001)	0.865 (0.860)	0.011 (0.006)
		GMDI	0.917	0.222	0.000	0.180	0.001	0.914	0.155	0.000	0.145	0.001
	2000	Proposed	0.996 (0.994)	0.980 (0.971)	0.000 (0.001)	0.960 (0.940)	0.000 (0.001)	0.996 (0.995)	0.980 (0.982)	0.000 (0.000)	0.960 (0.955)	0.000 (0.001)
		GMDI	0.925	0.307	0.000	0.255	0.000	0.921	0.238	0.000	0.220	0.001

High Misspecification

Settings			DINA					LCDM				
				Entry		Vector			Entry		Vector	
K	N		Total	TPR	FPR	TPR	FPR	Total	TPR	FPR	TPR	FPR
4	1000	Proposed	0.994 (0.992)	0.986 (0.986)	0.000 (0.001)	0.973 (0.968)	0.001 (0.002)	0.976 (0.979)	0.942 (0.960)	0.003 (0.003)	0.917 (0.917)	0.010 (0.006)
		GMDI	0.845	0.257	0.000	0.223	0.000	0.830	0.186	0.000	0.158	0.002
	2000	Proposed	0.995 (0.995)	0.989 (0.990)	0.000 (0.000)	0.978 (0.980)	0.001 (0.002)	0.987 (0.985)	0.979 (0.966)	0.002 (0.002)	0.953 (0.935)	0.004 (0.003)
		GMDI	0.855	0.312	0.000	0.280	0.002	0.841	0.260	0.002	0.235	0.008
5	1000	Proposed	0.985 (0.985)	0.965 (0.969)	0.001 (0.002)	0.940 (0.943)	0.004 (0.005)	0.973 (0.973)	0.955 (0.954)	0.003 (0.002)	0.912 (0.892)	0.012 (0.007)
		GMDI	0.827	0.162	0.001	0.145	0.003	0.817	0.121	0.001	0.100	0.004
	2000	Proposed	0.995 (0.990)	0.984 (0.981)	0.000 (0.002)	0.978 (0.958)	0.001 (0.002)	0.985 (0.981)	0.968 (0.963)	0.001 (0.002)	0.940 (0.922)	0.004 (0.005)
		GMDI	0.836	0.213	0.001	0.193	0.003	0.828	0.169	0.001	0.160	0.005

Table A.2: Misspecification detection results with $K = 4$ and 5. Results after the first J_0 steps are presented in brackets. "Total" is the proportion of correctly estimated items with the initial baseline 0.9 or 0.8 for low or high misspecification. "TPR" is true positive rate and "FPR" is the false positive rate.

Bootstrap bagging method

We illustrate the performance of proposed bootstrap bagging method. We consider the case with $K = 3$ and $N = 500$ under the LCDM. For each replicate in previous simulation, we perform bootstrap bagging estimation.

Figure A.3 summarizes the bootstrap aggregating with different threshold values from 0 to 1. On the left, the blue curve represents the averaged TPR, i.e., the proportions of misspecified entries that are correctly detected; the red curve is FPR, i.e., the proportion of correct entries being falsely detected. The right plot shows the averaged numbers of the true positive (TP) and the false positive (FP). In both plots, the dashed lines are the results using the proposed method without bootstrap aggregating. Figure A.3 shows the TPR (FPR) achieves maximum (minimum) at about $s = 0.5$ and therefore validates the choice of the threshold $s = 0.5$. Furthermore, the bootstrap method with $s = 0.5$ improves the stepwise detection method by reducing the overall false detection rates/numbers, especially the number of false positive entries.

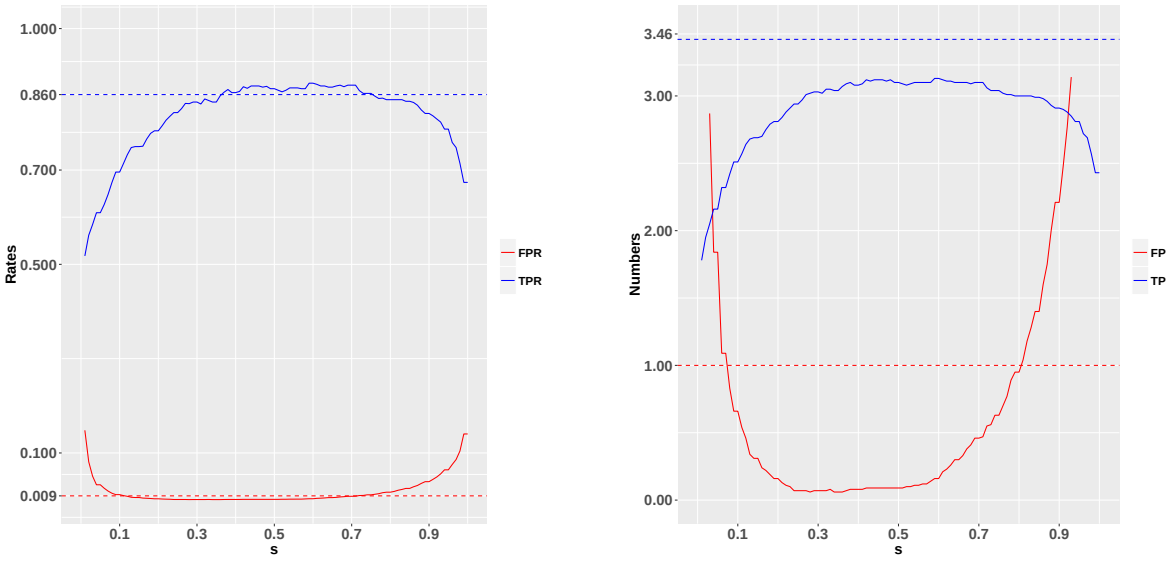

Figure A.3: Bootstrap aggregating results. The x-axis is the threshold of bootstrap method. TPR is the true positive rate; FPR is the false positive rate; TP is the true positive number of entries; FPR is the false positive number of entries.

To further illustrate the improvement, Table A.3 presents the result from one replication

where the sequential method has overestimated the misspecified entries. The right is the true Q-matrix. The left presents the misspecified initial Q-matrix; by design the Q-vectors for Item 2 and 19 are incorrect and the misspecified entries are identified using subscript "$*$". The left also shows the entries detected by the sequential estimation approach are highlighted in blue. We can see that it detects all misspecified entries but has two false discoveries (the blue entries without "$*$", i.e, item 12 with attribute 1 and item 18 with attribute 2). The bootstrap aggregated Q-matrix in the middle indicates only the starred entries should be changed, where recall that a "1" entry is detected if the bootstrap average is < 0.5 and a "0" entry is detected if > 0.5. This agrees with the true Q-matrix.

	Initial Q-matrix			Bootstrap Aggregation			True Q-matrix		
Item	Attr1	Attr2	Attr3	Attr1	Attr2	Attr3	Attr1	Attr2	Attr3
1	1	0	0	1.00	0.00	0.00	1	0	0
2	0	**0***	**1***	0.01	**1.00***	**0.10***	0	1	0
3	0	0	1	0.00	0.00	1.00	0	0	1
4	1	0	0	1.00	1.00	0.02	1	0	0
5	0	1	0	0.00	1.00	0.00	0	1	0
6	0	0	1	0.00	0.00	1.00	0	0	1
7	1	0	0	1.00	0.00	0.00	1	0	0
8	0	1	0	0.00	1.00	0.02	0	1	0
9	0	0	1	0.00	0.01	1.00	0	0	1
10	1	1	0	0.99	1.00	0.02	1	1	0
11	1	0	1	1.00	0.00	1.00	1	0	1
12	**0**	1	1	**0.07**	1.00	1.00	0	1	1
13	1	1	0	1.00	1.00	0.00	1	1	0
14	1	0	1	1.00	0.01	1.00	1	0	1
15	0	1	1	0.00	1.00	1.00	0	1	1
16	1	1	0	1.00	0.99	0.00	1	1	0
17	1	0	1	1.00	0.00	1.00	1	0	1
18	1	**1**	1	0.88	**0.79**	0.98	1	1	1
19	**0***	1	**0***	**0.95***	0.94	**0.89***	1	1	1
20	1	1	1	1.00	0.94	1.00	1	1	1

Table A.3: Bootstrap aggregating for a case with $K = 3$, $N = 500$, $\rho = 0$ under LCDM. The left is the initial Q-matrix with Item 2 and 19 misspecified. The middle matrix is the bootstrap aggregated Q-matrix. Entries in blue are detections from the stepwise method; and entries with "$*$" are true misspecified entries (and also detections with bootstrap significance).

References

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, 33(1), 1–22.

Geyer, C. J. (1994), "On the asymptotics of constrained M-estimation," *The Annals of Statistics*, 1993–2010.

Nishii, R. (1988), "Maximum likelihood principle and model selection when the true model is unspecified," *Journal of Multivariate Analysis*, 27, 392–403.

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

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer, New York.

Xu, G. (2017), "Identifiability of restricted latent class models with binary responses," *The Annals of Statistics*, 45, 675–707.