Supplement to "Diving Deep in Diagnostic Modeling: DeepCDMs"

In this Supplementary Material, Section S.1 presents the proofs of the identifiability results of DeepCDMs, and Section S.2 provides the posterior computation details of the Gibbs sampling algorithms for DeepCDMs.

S.1 Proofs of the Identifiability Results

All of our identifiability proofs leverage a key technical insight about DeepCDMs – that is, identifiability can be examined and established in a layer-by-layer manner, from the bottom up, thanks to the probabilistic formulation of the directed graphical model. This insight was initially used in Gu and Dunson (2021) to establish identifiability of the deep Bayesian Pyramid model for multivariate categorical data.

Proof of Theorem 1. Recall the joint distribution of all the random variables in a DeepCDM (including a DeepDINA model and a Hybrid DeepCDM) is

$$\mathbb{P}(\mathbf{R}, \mathbf{A}^{(1)}, \dots, \mathbf{A}^{(D)}) = \mathbb{P}(\mathbf{R} \mid \mathbf{A}^{(1)}) \cdot \prod_{d=2}^{D} \mathbb{P}(\mathbf{A}^{(d-1)} \mid \mathbf{A}^{(d)}) \cdot \mathbb{P}(\mathbf{A}^{(D)}).$$

The marginal distribution of the observed vector $\mathbf{R}$ is obtained by marginalizing out all the latent variables $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(D)}$ in the above joint distribution. According to the definition of a general directed acyclic graph (DAG), the marginal distribution of each latent vector $\mathbf{A}^{(d)}$ for layer $d = 1, \dots, D-1$ can be written as

$$\begin{aligned} &\mathbb{P}(\mathbf{A}^{(d)} = \boldsymbol{\alpha}^{(d)}) \qquad \text{(S.1)} \\ = &\sum_{\boldsymbol{\alpha}^{(d+1)} \in \{0,1\}^{K_{d+1}}} \cdots \sum_{\boldsymbol{\alpha}^{(D)} \in \{0,1\}^{K_D}} \prod_{m=d+1}^{D} \mathbb{P}(\mathbf{A}^{(m-1)} = \boldsymbol{\alpha}^{(m-1)} \mid \mathbf{A}^{(m)} = \boldsymbol{\alpha}^{(m)}) \cdot \mathbb{P}(\mathbf{A}^{(D)} = \boldsymbol{\alpha}^{(D)}). \end{aligned}$$

Now we specifically marginalize out all latent variables except the shallowest layer $\mathbf{A}^{(1)}$ in the joint distribution,

$$
\begin{aligned}
&\mathbb{P}(\mathbf{R}=\boldsymbol{r}) \\
= &\sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \cdots \sum_{\boldsymbol{\alpha}^{(D)}\in\{0,1\}^{K_D}} \mathbb{P}(\mathbf{R}=\boldsymbol{r}, \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}, \ldots, \mathbf{A}^{(D)}=\boldsymbol{\alpha}^{(D)}) \\
= &\sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \mathbb{P}(\mathbf{R}=\boldsymbol{r} \mid \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}) \times \\
&\qquad \underbrace{\sum_{\boldsymbol{\alpha}^{(2)}\in\{0,1\}^{K_2}} \cdots \sum_{\boldsymbol{\alpha}^{(D)}\in\{0,1\}^{K_D}} \prod_{d=2}^{D} \mathbb{P}(\mathbf{A}^{(d-1)}=\boldsymbol{\alpha}^{(d-1)} \mid \mathbf{A}^{(d)}=\boldsymbol{\alpha}^{(d)}) \cdot \mathbb{P}(\mathbf{A}^{(D)}=\boldsymbol{\alpha}^{(D)})}_{\mathbb{P}(\mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)})} \\
= &\sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \mathbb{P}(\mathbf{R}=\boldsymbol{r} \mid \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}) \cdot \mathbb{P}(\mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}), \qquad \text{(S.2)}
\end{aligned}
$$

We introduce a notation $\boldsymbol{\pi}^{(1)} = \left(\pi^{(1)}_{\boldsymbol{\alpha}};\ \boldsymbol{\alpha}\in\{0,1\}^{K_1}\right)$ to collect the proportion parameters of the categorical distribution that $\mathbf{A}^{(1)}$ follows in (S.2):

$$
\mathbb{P}(\mathbf{A}^{(1)}=\boldsymbol{\alpha}) = \pi^{(1)}_{\boldsymbol{\alpha}}, \quad \forall \boldsymbol{\alpha}\in\{0,1\}^{K_1}. \qquad \text{(S.3)}
$$

Then $\boldsymbol{\pi}^{(1)}$ lives in the $(2^{K_1}-1)$-dimensional probability simplex. Then based solely on $\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}$, the probability mass function of the random vector $\mathbf{R}$ can be written as follows for each $\boldsymbol{r}\in\{0,1\}^J$,

$$
\mathbb{P}(\mathbf{R}=\boldsymbol{r} \mid \boldsymbol{\pi}^{(1)}, \boldsymbol{\theta}^{(1)}, \mathbf{Q}^{(1)}) = \sum_{\boldsymbol{\alpha}^{(1)}\in\{0,1\}^{K_1}} \pi^{(1)}_{\boldsymbol{\alpha}^{(1)}} \prod_{j=1}^{J} \mathbb{P}(R_j=r_j \mid \mathbf{A}^{(1)}=\boldsymbol{\alpha}^{(1)}, \boldsymbol{\theta}^{(1)}, \mathbf{Q}^{(1)}), \quad \text{(S.4)}
$$

where the notation $\boldsymbol{\theta}^{(1)}$ collects all the continuous parameters needed to specify the conditional distribution of $\mathbf{R}\mid\mathbf{A}^{(1)}$ under $\mathbf{Q}^{(1)}$. For example, under the DeepDINA model, $\boldsymbol{\theta}^{(1)}$ denotes the collection of $\boldsymbol{s}^{(1)}$ and $\boldsymbol{g}^{(1)}$. Note that (S.4) gives a restricted latent class model (equivalently, a CDM) for $\mathbf{R}$ with 2^{K_1} latent classes, subject to the constraints induced by the $J\times K_1$ $\mathbf{Q}$-matrix $\mathbf{Q}^{(1)}$. Similarly, according to the general marginal distribution of $\mathbf{A}^{(d)}$

in (S.1), we also have

$$\mathbb{P}(\mathbf{A}^{(d)} \mid \mathbf{A}^{(d+1)}) = \sum_{\boldsymbol{\alpha}^{(d+1)} \in \{0,1\}^{K_{d+1}}} \mathbb{P}(\mathbf{A}^{(d)} \mid \mathbf{A}^{(d+1)} = \boldsymbol{\alpha}^{(d+1)}, \mathbf{Q}^{(d+1)}, \boldsymbol{\theta}^{(d+1)}) \cdot \mathbb{P}(\mathbf{A}^{(d+1)} = \boldsymbol{\alpha}^{(d+1)}),$$

which is another cognitive diagnostic model for the "response vector" being $\mathbf{A}^{(d)}$ and the "latent attribute vector" being $\mathbf{A}^{(d+1)}$ under the $\mathbf{Q}$-matrix $\mathbf{Q}^{(d+1)}$, where $d = 2, \ldots, D$.

Now consider the DeepDINA model setting in Theorem 1. When $\mathbf{R} \mid \mathbf{A}^{(1)}$ follows the DINA model, then as long as $\mathbf{Q}^{(1)}$ satisfies the C-R-D conditions in Gu and Xu (2021), then $\mathbf{Q}^{(1)}$ itself and the continuous parameters $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\pi}^{(1)}$ are identifiable. Note that the statement that $\boldsymbol{\pi}^{(1)}$ is identifiable means the marginal distribution of $\mathbf{A}^{(1)}$ is identifiable, which implies $\mathbf{A}^{(1)}$ can be treated as if it is observed when studying the identifiability of $\mathbf{Q}^{(2)}$, $\boldsymbol{\theta}^{(2)}$, and the marginal distribution of $\mathbf{A}^{(2)}$. Therefore, if $\mathbf{Q}^{(2)}$ also satisfies the C-R-D conditions, then $\mathbf{Q}^{(2)}$, $\boldsymbol{\theta}^{(2)}$, and the marginal distribution of $\mathbf{A}^{(2)}$ are identifiable. Now it is easy to see that we can proceed in a layerwise manner from bottom up, and examining whether $\mathbf{Q}^{(1)}$, $\mathbf{Q}^{(2)}$, ..., $\mathbf{Q}^{(D)}$ satisfy the identifiability conditions successively. Specifically, under a DeepDINA model, as long as all the $\mathbf{Q}^{(d)}$ satisfy the C-R-D conditions, then all the $\mathbf{Q}$-matrices and all the continuous parameters $(\boldsymbol{s}^{(d)}, \boldsymbol{g}^{(d)})$, $d = 1, \ldots, D$ and $\boldsymbol{\pi}^{\text{deep}}$ are strictly identifiable. This proves the sufficiency part in Theorem 1.

To show the necessity part in Theorem 1, we only need to note that if $\mathbf{Q}^{(d)}$ fails to satisfy the C-R-D conditions, then certain parameters in $\boldsymbol{\pi}^{(d)}$ and $\boldsymbol{\theta}^{(d)}$ will not identifiable, indicating the non-identifiability of the DeepDINA model. This proves the necessity of the proposed identifiability conditions and completes the proof of Theorem 1. □

Proof of Theorem 2 and Proposition 1. We use the same insight elaborated in the proof of Theorem 1: the layerwise proof argument of identifiability. Specifically, the marginal distribution of $\mathbf{R}$ in (S.2), the marginal distribution of $\mathbf{A}^{(1)}$ in (S.3), and the conditional distribution of $\mathbf{R}$ given $\mathbf{A}^{(1)}$ in (S.4) all hold generally for an arbitrary DeepCDM and a Hybrid DeepCDM. Therefore, we still start with the bottom two layers and examine whether $\mathbf{Q}^{(1)}$ satisfies the identifiability conditions for a general CDM; if so, we then examine $\mathbf{Q}^{(2)}$, so on and so forth. First, we consider the case that condition (S) holds; that is, each $\mathbf{Q}^{(d)}$

can be written as $\mathbf{Q}^{(d)} = [\mathbf{I}_{K_d},\ \mathbf{I}_{K_d},\ \mathbf{I}_{K_d},\ (\mathbf{Q}^{(d)*})^\top]^\top$ after some column/row permutation. In this case, following a similar argument as the proof of Theorem 4 in Gu and Dunson (2021) but constraining to considering binary responses, we obtain the strict identifiability of $(\boldsymbol{\theta}^{(d)}, \mathbf{Q}^{(d)})$ for $d = 1, \ldots, D$ and that of $\boldsymbol{\pi}^{\text{deep}}$. Second, we consider the case that condition (S*) holds, then following a similar argument as the proof of Theorem 1 in Culpepper (2019) but constraining to considering binary responses, we also obtain the strict identifiability of all the parameters and $\mathbf{Q}$-matrices in a general DeepCDM. This proves Theorem 2.

Further note that the above layerwise proof strategy does not require each layer in a DeepCDM to conform to the same diagnostic model. This means in a Hybrid CDM where some layers follow the DINA (or DINO) model and some layers follow the main-effect or all-effect diagnostic models, we can examine their corresponding $\mathbf{Q}$-matrices using the respective identifiability conditions in Theorems 1 or 2 to assess identifiability. For example, if the marginal distribution of $\mathbf{A}^{(d)}$ is already identified, then $\mathbf{A}^{(d)} \mid \mathbf{A}^{(d+1)}$ follows the DINA model, then $\mathbf{Q}^{(d+1)}$ only needs to satisfy the weaker C-R-D conditions to proceed to the deeper layer. This proves Proposition 1. $\square$

Proof of Theorem 3. Similarly as the proofs of strict identifiability results, we still use the layerwise identifiability argument. In the literature, Theorem 4 in Gu and Xu (2021) established generic identifiability for single-latent-layer main-effect/all-effect CDMs (also see Gu and Xu (2020) and Chen et al. (2020)) under the considered conditions (G1) and (G2) in its single-layer form ($D = 1$); in that theorem, the Lebesgue measure-zero subset of the parameter space where identifiability may break down only concerns the item parameters. That means, in the context of a DeepCDM consisting of main-effect or all-effect layers, as long as the item parameters $\boldsymbol{\theta}^{(1)} \in \Omega_{\text{main}}(\boldsymbol{\beta}^{(1)}; \mathbf{Q}^{(1)})$ do not fall within the layer-specific unidentifiable subset $\mathcal{N}^{(1)}$ which has measure zero in $\Omega_{\text{main}}(\boldsymbol{\beta}^{(1)}; \mathbf{Q}^{(1)})$, then $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\pi}^{(1)}$, and $\mathbf{Q}^{(1)}$ are identifiable. This implies that as long as the between-layer continuous parameters $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(D)}$ do not fall within the finite union of the measure-zero subsets of the parameter space $\cup_{d=1}^{D} \Omega_{\text{main}}(\boldsymbol{\beta}^{(d)}; \mathbf{Q}^{(d)})$, then the entire main-effect or all-effect DeepCDM is identifiable. This proves the generic identifiability conclusion in Theorem 3 under conditions (G1) and (G2). $\square$

S.2 Details for the Gibbs Sampling Algorithms

S.2.1 Gibbs Sampler for Two-latent-layer DeepDINA

For $i \in [N]$, $j \in [J]$, and $k \in [K_1]$, introduce binary ideal response indicators $\xi_{1,ij}$ and $\xi_{2,ik}$:

$$\xi_{1,ij} = \prod_{k=1}^{K_1} \left(a_{i,k}^{(1)}\right)^{q_{j,k}^{(1)}}, \quad \xi_{2,ik} = \prod_{m=1}^{K_2} \left(a_{i,m}^{(2)}\right)^{q_{k,m}^{(2)}}. \tag{S.5}$$

Denote $s_j^{(1)}$, $g_j^{(1)}$, $s_k^{(2)}$, and $g_k^{(2)}$ by $s_{1,j}$, $g_{1,j}$, $s_{2,k}$, and $g_{2,k}$, respectively. Under the priors specified in the main text, the posterior distribution in the two-latent-layer DeepDINA can be written as

$$\begin{aligned}
&p(\boldsymbol{\theta}^{(1)}_{\text{DINA}}, \boldsymbol{\theta}^{(2)}_{\text{DINA}}, \boldsymbol{\pi}^{\text{deep}}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)} \mid \mathbf{R}, \mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}) \\
\propto &\prod_{i=1}^{N}\prod_{j=1}^{J} \left[(1-s_{1,j})^{\xi_{1,ij}} g_{1,j}^{1-\xi_{1,ij}}\right]^{r_{i,j}} \left[s_{1,j}^{\xi_{1,ij}} (1-g_{1,j})^{1-\xi_{1,ij}}\right]^{1-r_{i,j}} \\
&\times \prod_{i=1}^{N}\prod_{\ell=1}^{2^{K_2}} \left\{ \pi_\ell \prod_{k=1}^{K_1} \left[(1-s_{2,k})^{\xi_{2,ik}} g_{2,k}^{1-\xi_{2,ik}}\right]^{a_{i,k}^{(1)}} \left[s_{2,k}^{\xi_{2,ik}} (1-g_{2,k})^{1-\xi_{2,ik}}\right]^{1-a_{i,k}^{(1)}} \right\}^{\mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell)} \\
&\times \prod_{j=1}^{J} [s_{1,j}^{a_s-1}(1-s_{1,j})^{b_s-1} g_{1,j}^{a_g-1}(1-g_{1,j})^{b_g-1} \mathbb{1}(g_{1,j} < 1-s_{1,j})] \\
&\times \prod_{k=1}^{K_1} [s_{2,k}^{a_s-1}(1-s_{2,k})^{b_s-1} g_{2,k}^{a_g-1}(1-g_{2,k})^{b_g-1} \mathbb{1}(g_{2,k} < 1-s_{2,k})] \times \prod_{\ell=1}^{2^{K_2}} \pi_\ell^{\delta-1}
\end{aligned}$$

Based on the above posterior, the full conditional distributions of the quantities $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, $\boldsymbol{\pi}^{\text{deep}}$, $\mathbf{A}^{(1)}$, $\mathbf{A}^{(2)}$ are as follows.

(1) Sample $s_{1,j}^{(1)}$ and $g_{1,j}^{(1)}$ from truncated Beta distributions:

$$\begin{aligned}
s_j^{(1)} &\sim \text{Beta}\left(1 + \sum\nolimits_{i=1}^{N} (1-r_{ij})\xi_{1,ij},\ 1 + \sum\nolimits_{i=1}^{N} r_{ij}\xi_{1,ij}\right) \cdot \mathbb{1}(s_j^{(1)} < 1 - g_j^{(1)}); \\
g_j^{(1)} &\sim \text{Beta}\left(1 + \sum\nolimits_{i=1}^{N} r_{ij}(1-\xi_{1,ij}),\ 1 + \sum\nolimits_{i=1}^{N} (1-r_{ij})(1-\xi_{1,ij})\right) \cdot \mathbb{1}(g_j^{(1)} < 1 - s_j^{(1)}).
\end{aligned}$$

(2) Sample $s_{2,k}^{(2)}$ and $g_{2,k}^{(2)}$ from truncated Beta distributions:

$$s_k^{(2)} \sim \text{Beta}\left(1 + \sum\nolimits_{i=1}^N (1 - a_{ik}^{(1)})\xi_{2,ik},\ 1 + \sum\nolimits_{i=1}^N a_{ik}^{(1)}\xi_{2,ik}\right) \cdot \mathbb{1}(s_k^{(2)} < 1 - g_k^{(2)});$$
$$g_k^{(2)} \sim \text{Beta}\left(1 + \sum\nolimits_{i=1}^N a_{ik}^{(1)}(1 - \xi_{2,ik}),\ 1 + \sum\nolimits_{i=1}^N (1 - a_{ik}^{(1)})(1 - \xi_{2,ik})\right) \cdot \mathbb{1}(g_k^{(2)} < 1 - s_k^{(2)}).$$

(3) Sample $\boldsymbol{\pi}^{\text{deep}}$ from the Dirichlet distribution:

$$\boldsymbol{\pi}^{\text{deep}} \sim \text{Dirichlet}\left(\delta_1 + \sum\nolimits_{i=1}^N \mathbb{1}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_1), \ldots, \delta_{2^{K_2}} + \sum\nolimits_{i=1}^N \mathbb{1}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{2^{K_2}})\right).$$

(4) Sample each entry $a_{i,k}^{(1)}$ from the Bernoulli distribution with the following probability:

$$\begin{aligned}\mathbb{P}(a_{i,k}^{(1)} = 1 \mid -) &= \mathbb{P}(a_{i,k}^{(1)} = 1 \mid \boldsymbol{r}_i, \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) \\ &= \frac{\mathbb{P}(a_{i,k}^{(1)} = 1 \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)} = 1, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})}{\sum_{x=0,1} \mathbb{P}(a_{i,k}^{(1)} = x \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)} = x, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})},\end{aligned}$$

where the conditional distributions $\mathbb{P}(a_{i,k}^{(1)} = x \mid \boldsymbol{a}_i^{(2)}, \boldsymbol{\theta}^{(2)})$ and $\mathbb{P}(\boldsymbol{r}_i \mid a_{i,k}^{(1)} = x, \boldsymbol{a}_{i,-k}^{(1)}, \boldsymbol{\theta}^{(1)})$ just directly follow the likelihood defined under the DeepDINA model in Section 4.1 of the main text, and they are both DINA.

(5) Sample each pattern $\boldsymbol{a}_i^{(2)}$ from the categorical distribution with $|\{0,1\}^{K_2}| = 2^{K_2}$ components with the following probabilities:

$$\begin{aligned}\mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell \mid -) &= \mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell \mid \boldsymbol{a}_i^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\pi}^{\text{deep}}); \\ &= \frac{\mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell, \boldsymbol{\theta}^{(2)})}{\sum_{\ell'=1}^{2^{K_2}} \mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})},\end{aligned}$$

where the $\mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})$ and $\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})$ also directly follow the definition of DeepDINA, with the former being a Dirichlet distribution and the latter following a DINA model conditional distribution.

Overall, our Gibbs sampler cycles through the above five steps iteratively to approximate the posterior distributions of all the quantities.

S.2.2 Gibbs Sampler for Hybrid GDINA-DINA

Recall that we will focus on those $\theta^{(1)}_{j,S}$ parameters for the shallower GDINA layer during the Gibbs sampling, which denote conditional positive response probabilities:

$$\theta^{(1)}_{j,S} = \sum_{S' \subseteq S} \beta^{(1)}_{j,S'} = \mathbb{P}(r_{i,j} = 1 \mid \boldsymbol{a}_i^{(1)\top} \boldsymbol{q}^{(1)}_{j,S} = \boldsymbol{q}^{(1)\top}_{j,S} \boldsymbol{q}^{(1)}_{j,S}).$$

Introduce binary indicators for the GDINA layer as

$$\xi_{1,ij,S} = \mathbb{1}\left(\boldsymbol{a}_i^{(1)\top} \boldsymbol{q}^{(1)}_{j,S} = \boldsymbol{q}^{(1)\top}_{j,S} \boldsymbol{q}^{(1)}_{j,S}\right), \quad i \in [N],\ j \in [J],\ S \subseteq \mathcal{K}_j,$$

where the notation $\mathcal{K}_j = \{k \in [K_1] :\ q^{(1)}_{j,k} = 1\}$ was defined in the main text. For the deeper DINA layer, we still introduce binary ideal response indicators $\xi_{2,ik}$ for $k \in [K_1]$ similarly as the previous (S.5). Under the priors specified in the main text, the posterior distribution in the Hybrid GDINA-DINA can be written as

$$\begin{aligned}
& p(\boldsymbol{\theta}^{(1)}_{\text{GDINA}}, \boldsymbol{\theta}^{(2)}_{\text{DINA}}, \boldsymbol{\pi}^{\text{deep}}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)} \mid \mathbf{R}, \mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}) \\
\propto & \prod_{i=1}^{N} \prod_{j=1}^{J} \prod_{S \subseteq \mathcal{K}_j} \left[\left(\theta^{(1)}_{j,S}\right)^{r_{i,j}\xi_{1,ij,S}} \left(1 - \theta^{(1)}_{j,S}\right)^{(1-r_{i,j})\xi_{1,ij,S}} \right] \\
\times & \prod_{i=1}^{N} \prod_{\ell=1}^{2^{K_2}} \left\{ \pi_\ell \prod_{k=1}^{K_1} \left[(1 - s_{2,k})^{\xi_{2,ik}} g_{2,k}^{1-\xi_{2,ik}} \right]^{a^{(1)}_{i,k}} \left[s_{2,k}^{\xi_{2,ik}} (1 - g_{2,k})^{1-\xi_{2,ik}} \right]^{1-a^{(1)}_{i,k}} \right\}^{\mathbb{1}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell)} \\
\times & \prod_{j=1}^{J} \prod_{S \subseteq \mathcal{K}_j} \left[(\theta^{(1)}_{j,S})^{a_\theta - 1} (1 - \theta^{(1)}_{j,S})^{a_\theta - 1} \mathbb{1}(\theta^{(1)}_{j,S} > \theta^{(1)}_{j,\varnothing} \text{ if } S \text{ is a singleton set}) \right] \\
\times & \prod_{k=1}^{K_1} [s_{2,k}^{a_s - 1} (1 - s_{2,k})^{b_s - 1} g_{2,k}^{a_g - 1} (1 - g_{2,k})^{b_g - 1} \mathbb{1}(g_{2,k} < 1 - s_{2,k})] \times \prod_{\ell=1}^{2^{K_2}} \pi_\ell^{\delta - 1}.
\end{aligned}$$

Our Gibbs sampler will cycle through the following steps iteratively.

(1) Sample each $\theta^{(1)}_{j,S}$ from the (truncated) Beta distribution:

$$\theta^{(1)}_{j,S} \sim \text{Beta}\left(a_\theta + \sum_{i=1}^{N} r_{i,j}\xi_{1,ij,S},\ b_\theta + \sum_{i=1}^{N} (1 - r_{i,j})\xi_{1,ij,S}\right) \mathbb{1}(\theta^{(1)}_{j,S} > \theta^{(1)}_{j,\varnothing} \text{ if } S \text{ is a singleton set}).$$

(2) Sample $s_{2,k}^{(2)}$ and $g_{2,k}^{(2)}$ from truncated Beta distributions:

$$s_k^{(2)} \sim \text{Beta}\left(1+\sum\nolimits_{i=1}^{N}(1-a_{ik}^{(1)})\xi_{2,ik},\ 1+\sum\nolimits_{i=1}^{N}a_{ik}^{(1)}\xi_{2,ik}\right)\cdot \mathbb{1}(s_k^{(2)}<1-g_k^{(2)});$$
$$g_k^{(2)} \sim \text{Beta}\left(1+\sum\nolimits_{i=1}^{N}a_{ik}^{(1)}(1-\xi_{2,ik}),\ 1+\sum\nolimits_{i=1}^{N}(1-a_{ik}^{(1)})(1-\xi_{2,ik})\right)\cdot \mathbb{1}(g_k^{(2)}<1-s_k^{(2)}).$$

(3) Sample $\boldsymbol{\pi}^{\text{deep}}$ from the Dirichlet distribution:

$$\boldsymbol{\pi}^{\text{deep}} \sim \text{Dirichlet}\left(\delta_1+\sum\nolimits_{i=1}^{N}\mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_1),\ldots,\delta_{2^{K_2}}+\sum\nolimits_{i=1}^{N}\mathbb{1}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{2^{K_2}})\right).$$

(4) Sample each entry $a_{i,k}^{(1)}$ from the Bernoulli distribution with the following probability:

$$\mathbb{P}(a_{i,k}^{(1)}=1\mid -)=\frac{\mathbb{P}(a_{i,k}^{(1)}=1\mid \boldsymbol{a}_i^{(2)},\boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i\mid a_{i,k}^{(1)}=1,\boldsymbol{a}_{i,-k}^{(1)},\boldsymbol{\theta}^{(1)})}{\sum_{x=0,1}\mathbb{P}(a_{i,k}^{(1)}=x\mid \boldsymbol{a}_i^{(2)},\boldsymbol{\theta}^{(2)})\mathbb{P}(\boldsymbol{r}_i\mid a_{i,k}^{(1)}=x,\boldsymbol{a}_{i,-k}^{(1)},\boldsymbol{\theta}^{(1)})},$$

where the conditional distributions $\mathbb{P}(a_{i,k}^{(1)}=x\mid \boldsymbol{a}_i^{(2)},\boldsymbol{\theta}^{(2)})$ and $\mathbb{P}(\boldsymbol{r}_i\mid a_{i,k}^{(1)}=x,\boldsymbol{a}_{i,-k}^{(1)},\boldsymbol{\theta}^{(1)})$ follow the likelihood under the DINA and GDINA, respectively.

(5) Sample each pattern $\boldsymbol{a}_i^{(2)}$ from the categorical distribution with $|\{0,1\}^{K_2}|=2^{K_2}$ components with the following probabilities:

$$\begin{aligned}\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell\mid -)&=\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell\mid \boldsymbol{a}_i^{(1)},\boldsymbol{\theta}^{(2)},\boldsymbol{\pi}^{\text{deep}});\\ &=\frac{\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell\mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)}\mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_\ell,\boldsymbol{\theta}^{(2)})}{\sum_{\ell'=1}^{2^{K_2}}\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'}\mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)}\mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'},\boldsymbol{\theta}^{(2)})},\end{aligned}$$

where the $\mathbb{P}(\boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'}\mid \boldsymbol{\pi}^{\text{deep}})$ and $\mathbb{P}(\boldsymbol{a}_i^{(1)}\mid \boldsymbol{a}_i^{(2)}=\boldsymbol{\alpha}_{\ell'},\boldsymbol{\theta}^{(2)})$ also directly follow the definition of DeepDINA, with the former being a Dirichlet distribution and the latter following a DINA model conditional distribution.

S.2.3 Gibbs Sampler for Two-latent-layer DeepLLM

The posterior distribution of the two-latent-layer DeepLLM can be written as

$$
\begin{aligned}
&p(\boldsymbol{\beta}^{(1)}_{\mathrm{LLM}}, \boldsymbol{\beta}^{(2)}_{\mathrm{LLM}}, \boldsymbol{\pi}^{\mathrm{deep}}, \mathbf{A}^{(1)}, \mathbf{A}^{(2)} \mid \mathbf{R}, \mathbf{Q}^{(1)}, \mathbf{Q}^{(2)}) \\
\propto &\prod_{i=1}^{N} \left\{ \prod_{j=1}^{J} \frac{\exp\left(r_{i,j}\left(\beta^{(1)}_{j,0} + \sum_{k=1}^{K_1} q^{(1)}_{j,k}\beta^{(1)}_{j,k}a^{(1)}_{i,k}\right)\right)}{1+\exp\left(\beta^{(1)}_{j,0} + \sum_{k=1}^{K_1} q^{(1)}_{j,k}\beta^{(1)}_{j,k}a^{(1)}_{i,k}\right)} \times \prod_{k=1}^{K_1} \frac{\exp\left(a^{(1)}_{i,k}\left(\beta^{(2)}_{k,0} + \sum_{m=1}^{K_2} q^{(2)}_{k,m}\beta^{(2)}_{k,m}a^{(2)}_{i,m}\right)\right)}{1+\exp\left(\beta^{(2)}_{k,0} + \sum_{m=1}^{K_2} q^{(2)}_{k,m}\beta^{(2)}_{k,m}a^{(2)}_{i,m}\right)} \right\} \\
&\times \prod_{i=1}^{N}\prod_{\ell=1}^{2^{K_2}} \pi_\ell^{\mathbb{1}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_\ell)} \times \prod_{\ell=1}^{2^{K_2}} \pi_\ell^{\delta-1} \times \prod_{j=1}^{J} \left\{ N(\beta^{(1)}_{j,0} \mid 0, \sigma^2_\beta) \prod_{k=0}^{K_1} N(\beta^{(1)}_{j,k} \mid 0, \sigma^2_\beta)\mathbb{1}(\beta^{(1)}_{j,k} > 0 \text{ if } q^{(1)}_{j,k} = 1) \right\} \\
&\times \prod_{k=1}^{K_1} \left\{ N(\beta^{(2)}_{k,0} \mid 0, \sigma^2_\beta) \prod_{m=0}^{K_2} N(\beta^{(2)}_{k,m} \mid 0, \sigma^2_\beta)\mathbb{1}(\beta^{(2)}_{k,m} > 0 \text{ if } q^{(2)}_{k,m} = 1) \right\} \\
&\times \prod_{i=1}^{N}\prod_{j=1}^{J} \mathrm{PG}(w^{(1)}_{i,j} \mid 1, 0) \cdot \prod_{i=1}^{N}\prod_{k=1}^{K_1} \mathrm{PG}(w^{(2)}_{i,k} \mid 1, 0).
\end{aligned}
$$

Our Gibbs sampler iteratively cycles through the following steps.

(1) Recall the notation $\mathcal{K}_j = \{k \in [K_1] :\ q^{(1)}_{j,k} = 1\}$. Define

$$\boldsymbol{\beta}^{(1)}_{j,\mathcal{K}_j} = (\beta^{(1)}_{j,0},\ \beta^{(1)}_{j,k};\ k \in \mathcal{K}_j),$$

which is a vector of length $|\mathcal{K}_j| + 1$. We introduce a notation $\mathbf{X}^{(1)}_j$, which is a $N \times |\mathcal{K}_j|$ matrix; the entries in this matrix are indexed by $a^{(1)}_{i,k}q^{(1)}_{j,k}$ where $i \in [N]$ and $k \in \{0\} \cup \mathcal{K}_j$. Sample $\boldsymbol{\beta}^{(1)}_{j,\mathcal{K}_j}$ from the (truncated) Multivariate Normal (MVN) distribution:

$$
\begin{aligned}
&\boldsymbol{\beta}^{(1)}_{j,\mathcal{K}_j} \sim \mathrm{MVN}(\boldsymbol{\mu}_{1j}, \boldsymbol{\Sigma}_{1j}), \quad \text{where} \\
&\boldsymbol{\Sigma}_{1j} = \left(\mathbf{X}^{(1)\top}_j \mathrm{diag}\left(\mathbf{W}^{(1)}_{:,j}\right)\mathbf{X}^{(1)}_j\right)^{-1}, \quad \boldsymbol{\mu}_{1j} = \boldsymbol{\Sigma}_{1j}\mathbf{X}^{(1)\top}_j\left(\mathbf{R}_{:,j} - 1/2\right).
\end{aligned}
$$

(2) Define a new notation

$$\mathcal{K}_{2,k} = \{m \in [K_2] :\ q^{(2)}_{k,m} = 1\}.$$

Define

$$\boldsymbol{\beta}^{(2)}_{k,\mathcal{K}_{2,k}} = (\beta^{(2)}_{k,0},\ \beta^{(2)}_{k,m};\ m \in \mathcal{K}_{2,k}),$$

which is a vector of length $|\mathcal{K}_{2,k}| + 1$. We introduce a notation $\mathbf{X}^{(2)}_k$, which is a $N \times |\mathcal{K}_{2,k}|$ matrix; the entries in this matrix are indexed by $a^{(2)}_{i,m} q^{(2)}_{k,m}$ where $i \in [N]$ and $m \in \{0\} \cup \mathcal{K}_{2,k}$. Sample $\boldsymbol{\beta}^{(2)}_{k,\mathcal{K}_{2,k}}$ from the (truncated) Multivariate Normal (MVN) distribution:

$$\begin{aligned} &\boldsymbol{\beta}^{(2)}_{k,\mathcal{K}_{2,k}} \sim \mathrm{MVN}(\boldsymbol{\mu}_{2k}, \boldsymbol{\Sigma}_{2k}), \quad \text{where} \\ &\boldsymbol{\Sigma}_{2k} = \left(\mathbf{X}^{(2)\top}_k \mathrm{diag}\left(\mathbf{W}^{(2)}_{:,k}\right) \mathbf{X}^{(2)}_k\right)^{-1}, \quad \boldsymbol{\mu}_{2k} = \boldsymbol{\Sigma}_{2k} \mathbf{X}^{(2)\top}_k \left(\mathbf{A}^{(1)}_{:,k} - 1/2\right). \end{aligned}$$

(3) Sample each $w^{(1)}_{i,j}$, $j \in [J]$ from the Polya-Gamma distribution:

$$w^{(1)}_{i,j} \sim \mathrm{PG}\Big(1, \beta^{(1)}_{j,0} + \sum_{k \in \mathcal{K}_j} \beta^{(1)}_{j,k} a^{(1)}_{i,k}\Big).$$

(4) Sample each $w^{(2)}_{i,k}$, $k \in [K_1]$ from the Polya-Gamma distribution:

$$w^{(2)}_{i,k} \sim \mathrm{PG}\Big(1, \beta^{(2)}_{k,0} + \sum_{m \in \mathcal{K}_{2,k}} \beta^{(2)}_{k,m} a^{(2)}_{i,m}\Big).$$

(5) Sample $\boldsymbol{\pi}^{\mathrm{deep}}$ from the Dirichlet distribution:

$$\boldsymbol{\pi}^{\mathrm{deep}} \sim \mathrm{Dirichlet}\left(\delta_1 + \sum\nolimits_{i=1}^{N} \mathbb{1}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_1), \ldots, \delta_{2^{K_2}} + \sum\nolimits_{i=1}^{N} \mathbb{1}(\boldsymbol{a}^{(2)}_i = \boldsymbol{\alpha}_{2^{K_2}})\right).$$

(6) Sample each entry $a^{(1)}_{i,k}$ from the Bernoulli distribution with the following probability:

$$\mathbb{P}(a^{(1)}_{i,k} = 1 \mid -) = \frac{\mathbb{P}(a^{(1)}_{i,k} = 1 \mid \boldsymbol{a}^{(2)}_i, \boldsymbol{\theta}^{(2)}) \mathbb{P}(\boldsymbol{r}_i \mid a^{(1)}_{i,k} = 1, \boldsymbol{a}^{(1)}_{i,-k}, \boldsymbol{\theta}^{(1)})}{\sum_{x=0,1} \mathbb{P}(a^{(1)}_{i,k} = x \mid \boldsymbol{a}^{(2)}_i, \boldsymbol{\theta}^{(2)}) \mathbb{P}(\boldsymbol{r}_i \mid a^{(1)}_{i,k} = x, \boldsymbol{a}^{(1)}_{i,-k}, \boldsymbol{\theta}^{(1)})},$$

where the conditional distributions $\mathbb{P}(a^{(1)}_{i,k} = x \mid \boldsymbol{a}^{(2)}_i, \boldsymbol{\theta}^{(2)})$ and $\mathbb{P}(\boldsymbol{r}_i \mid a^{(1)}_{i,k} = x, \boldsymbol{a}^{(1)}_{i,-k}, \boldsymbol{\theta}^{(1)})$

both follow the likelihood under the LLM.

(7) Sample each pattern $\boldsymbol{a}_i^{(2)}$ from the categorical distribution with $|\{0,1\}^{K_2}| = 2^{K_2}$ components with the following probabilities:

$$\begin{aligned}\mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell \mid -) &= \mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell \mid \boldsymbol{a}_i^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\pi}^{\text{deep}}); \\ &= \frac{\mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_\ell, \boldsymbol{\theta}^{(2)})}{\sum_{\ell'=1}^{2^{K_2}} \mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})},\end{aligned}$$

where the $\mathbb{P}(\boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'} \mid \boldsymbol{\pi}^{\text{deep}})$ and $\mathbb{P}(\boldsymbol{a}_i^{(1)} \mid \boldsymbol{a}_i^{(2)} = \boldsymbol{\alpha}_{\ell'}, \boldsymbol{\theta}^{(2)})$ also directly follow the definition of LLM, with the former being a Dirichlet distribution and the latter following a LLM model conditional distribution.

References

Chen, Y., Culpepper, S. A., and Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, 85(1):121–153.

Culpepper, S. A. (2019). An exploratory diagnostic model for ordinal responses with binary attributes: identifiability and estimation. *Psychometrika*, 84(4):921–940.

Gu, Y. and Dunson, D. B. (2021). Bayesian pyramids: Identifiable multilayer discrete latent structure models for discrete data. *arXiv preprint arXiv:2101.10373*.

Gu, Y. and Xu, G. (2020). Partial identifiability of restricted latent class models. *Annals of Statistics*, 48(4):2082–2107.

Gu, Y. and Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica*, 31:449–472.