PSYCHOMETRIKA—VOL. 85, NO. 1, 121–153
MARCH 2020
https://doi.org/10.1007/s11336-019-09693-2

A SPARSE LATENT CLASS MODEL FOR COGNITIVE DIAGNOSIS

YINYIN CHEN

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

STEVEN CULPEPPER

UNIVERSITY OF ILLINOIS AT URBANA–CHAMPAIGN

FENG LIANG

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Cognitive diagnostic models (CDMs) are latent variable models developed to infer latent skills, knowledge, or personalities that underlie responses to educational, psychological, and social science tests and measures. Recent research focused on theory and methods for using sparse latent class models (SLCMs) in an exploratory fashion to infer the latent processes and structure underlying responses. We report new theoretical results about sufficient conditions for generic identifiability of SLCM parameters. An important contribution for practice is that our new generic identifiability conditions are more likely to be satisfied in empirical applications than existing conditions that ensure strict identifiability. Learning the underlying latent structure can be formulated as a variable selection problem. We develop a new Bayesian variable selection algorithm that explicitly enforces generic identifiability conditions and monotonicity of item response functions to ensure valid posterior inference. We present Monte Carlo simulation results to support accurate inferences and discuss the implications of our findings for future SLCM research and educational testing.

Key words: sparse latent class models, Bayesian variable selection, identifiability.

1. Introduction

Cognitive diagnostic models (CDMs) are latent class models developed for the inference of educational, psychological, and social science tests. In CDMs, the latent variables are often defined as skills, knowledge, or personalities needed by a subject to solve a given test item. Consider a test that consists of J items and involves K skills. The observable response $\boldsymbol{Y} = (Y_1, \ldots, Y_J)$ for a subject is a binary random vector, indicating the correctness of the subject's answers to the J items. The latent class of a subject is indexed by a K-dimensional binary vector $\boldsymbol{\alpha}$, which are referred to as an *attribute profile*, which suggests the mastery of each skill. Given $\boldsymbol{\alpha}$, $\boldsymbol{Y}$ is modeled by a product of J independent Bernoulli random variables with parameter $\theta_{j,\boldsymbol{\alpha}} = \mathbb{P}(Y_j = 1|\boldsymbol{\alpha})$.

In many CDMs, $\theta_{j,\boldsymbol{\alpha}}$ depends on a K-dimensional binary vector $\boldsymbol{q}_j$, where element $q_{jk} = 1$ if attribute k is relevant to item j and zero otherwise. The relevant skills to all items are usually presented by a $J \times K$ matrix, $\boldsymbol{Q} = [\boldsymbol{q}_1 \ldots, \boldsymbol{q}_J]^{\mathrm{T}}$, called the $\boldsymbol{Q}$*-matrix*.

CDMs provide a statistical framework to identify relevant attributes of test items and to classify the mastery/non-mastery of test subjects on those attributes, which provide useful insights for researchers and educators. Various CDMs have been proposed in the literature and they differ from each other in their assumptions on how $\theta_{j,\boldsymbol{\alpha}}$ depends on the $\boldsymbol{Q}$-matrix. For example, the

Correspondence should be made to Steven Culpepper, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana–Champaign, 725 South Wright Street, Champaign, IL 61820, USA. Email: sculpepp@illinois.edu

DINA (Deterministic Input, Noisy 'And' gate) model (Haertel 1989; Junker and Sijtsma 2001), the generalized DINA model (de la Torre 2011) and the reduced reparameterized unified (rRUM) model (Hartz 2002; Rupp et al. 2010) are conjunctive models where all relevant skills are needed to have the highest positive response probability. On the other hand, under disjunctive models, such as the DINO (Deterministic Input, Noisy 'Or' gate) model (Templin and Henson 2006), at least one relevant skill is needed, whereas compensatory models (e.g., see Davier (2005), for a special case of the general diagnostic model) allow students to compensate for missing some skills by having others.

However, pre-specifying an appropriate CDM can be difficult in practice, especially when no prior knowledge of the test is available. Further, it is likely that different questions (items) in a single test need to be modeled by different CDMs. For instance, a mathematics test may include questions that can be solved using different skills, which implies a disjunctive model. The same test may also include questions that involve multiple steps and require students to master all relevant skills in order to get the final correct answer, which implies a conjunctive model. For such tests, the specification of a single CDM would fail to capture the real latent patterns.

To address this issue, a novel, model-free approach was proposed by Chen et al. (2015), which is based on an alternative representation of CDMs via a mixture of generalized linear models (GLMs). In particular, the jth Bernoulli parameter for latent class $\boldsymbol{\alpha}$ is modeled as

$$\mathbb{P}\left(Y_j = 1 \,|\boldsymbol{\alpha}, \boldsymbol{\beta}_j\right) = \Psi\left(\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k}\alpha_k + \sum_{k>k'}\sum \beta_{j,kk'}\alpha_k\alpha_{k'} + \cdots + \beta_{j,12\ldots K}\prod_{k=1}^{K}\alpha_k\right) \tag{1}$$

where $\Psi(\cdot)$ is an arbitrary cumulative distribution function (CDF) and $\boldsymbol{\beta}_j$ is a coefficient vector to be estimated. It can be shown that all of the aforementioned CDMs are special cases of (1) with particular sparsity patterns of $\boldsymbol{\beta}_j$. In addition, the sparsity pattern of $\boldsymbol{\beta}_j$ provides information about which attributes are relevant. Consequently, the estimation of $\boldsymbol{Q}$ can be reformulated as a variable selection problem involving GLMs.

A fundamental issue with latent mixture models is model identifiability. Throughout, identifiability is defined up to a permutation of the K attributes. That is, we do not discuss the trivial identifiability issue due to label switching, since it is a well-understood issue and we know how to handle it in practice. The first rigorous study on the identifiability of the $\boldsymbol{Q}$-matrix was given by Liu et al. (2013) with a focus on DINA models with known guessing parameters. Chen et al. (2015) extended the result of Liu et al. (2013) to DINA and DINO models with unknown model parameters. Xu (2017) established identifiability conditions for general CDMs when $\boldsymbol{Q}$ is known; Xu and Shang (2017) later provided identifiability conditions for general CDMs when $\boldsymbol{Q}$ is unknown. Although the results from Xu (2017) and Xu and Shang (2017) are applicable to the general model based on latent mixture of GLMs, their identifiability conditions, which require two identity matrices embedded in $\boldsymbol{Q}$, are too strong to be satisfied in practice. Recently, Fang et al. (2019) proposed identifiability conditions in terms of the distribution of $\boldsymbol{Y}$ rather than $\boldsymbol{Q}$, but their conditions are still stronger than the ones needed for our result.

In this paper, we provide a new set of *generic identifiability* conditions for general CDMs. Our conditions are weaker than the ones in the aforementioned papers since those papers studied conditions for *strict identifiability*. As stated by Allman et al. (2009), "…generic identifiability implies that the set of points for which identifiability does not hold has measure zero," which is enough for practical data analysis. For example, Bernoulli mixtures are not strictly identifiable (Goodman 1974; Gyllenberg et al. 1994a), but given Bernoulli mixtures are generically identifiable, they often lead to valid statistical inference in practice (Carreira-Perpiñán and Renals 2000; Allman et al. 2009). Results for generic identifiability are established in Allman et al. (2009) for

latent class models, which cannot be directly applied to CDMs, since CDMs are restricted latent class models. However, we can extend their proof technique, which is based on the tensor product framework in Kruskal (1976, 1977), to handle CDMs.

For a feasible implementation, we adopt a Bayesian approach to the estimation and variable selection of (1), and develop a Gibbs sampling algorithm for computation. Different from the vanilla Gibbs algorithm for Bayesian variable selection, our Gibbs algorithm is specially designed to ensure that each posterior draw of the sparse model parameters $\boldsymbol{\beta}_j$'s is from the identifiable space, while the algorithms from Chen et al. (2015), Xu and Shang (2017) and Fang et al. (2019) cannot ensure that their estimation of $\boldsymbol{Q}$ or other model parameters satisfies their identifiability conditions. In a recent work, Chen et al. (2018) also used MCMC method to stochastically search over identifiable parameter space. However, Chen et al. (2018) focus on DINA and DINO models with strict identifiability conditions, whereas we develop an MCMC sampling algorithm for general CDMs with weaker, generic identifiability conditions.

The remainder of this paper is organized as follows. Section 2 introduces the setup of the sparse latent class models. Section 3 addresses identifiability issues and discusses the generic identifiability conditions. Section 4 introduces the Gibbs sampler for posterior inference. We report results from a simulation study in Sect. 5 and results from two real applications in Sect. 6, and close with a discussion in Sect. 7. Proofs and other technical details are included in "Appendix".

2. Model and Applications

2.1. Model Setup

Consider a test consisting of J items and involving K latent skills. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_J)$ denote the J binary responses from a subject. Based on the mastery of the K skills, each subject has an *attribute profile* $\boldsymbol{\alpha} \in \{0, 1\}^K$, where $\alpha_k = 1$ indicates that the subject masters skill k and zero otherwise. In SLCM, given the *attribute profile* $\boldsymbol{\alpha}$, the J binary variables $Y_1, \ldots, Y_J$ are modeled as independent Bernoulli variables with Bernoulli parameter $\theta_{j,\boldsymbol{\alpha}} = \Psi\left(\boldsymbol{a}_{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\beta}_j\right)$ where

$$\boldsymbol{a}_{\boldsymbol{\alpha}} = \left(1, \alpha_1, \ldots, \alpha_K, \alpha_1\alpha_2, \ldots, \alpha_{K-1}\alpha_K, \ldots, \prod_{k=1}^{K}\alpha_k\right)^{\mathrm{T}} \tag{2}$$

is a 2^K-dimensional alternative representation of the binary vector $\boldsymbol{\alpha}$ with

$$\boldsymbol{a}_{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\beta}_j = \beta_{j,0} + \sum_{k=1}^{K}\beta_{j,k}\alpha_k + \sum\sum_{k>k'}\beta_{j,kk'}\alpha_k\alpha_{k'} + \cdots + \beta_{j,12\ldots K}\prod_{k=1}^{K}\alpha_k. \tag{3}$$

The regression coefficients $\boldsymbol{\beta}_j$ form a sparse vector, in which the nonzero elements represent the effects of skills or combinations of skills on the response of item j.

In particular, where coefficients in (3) are nonzero imply the dependence between item j and the K skills. To identify which main effects and/or interactions of the K skills are relevant to item j, we introduce a 2^K-dimensional binary *structure vector*

$$\boldsymbol{\delta}_j = (\delta_{j,0}, \delta_{j,1}, \ldots, \delta_{j,K}, \delta_{j,12}, \ldots, \delta_{j,(K-1)K}, \ldots, \delta_{j,1\ldots K})^{\mathrm{T}} \in \{0, 1\}^{2^K},$$

TABLE 1.
Sparse patterns of various CDMs

	$\delta_{j,0}$	$\delta_{j,1}$	$\delta_{j,2}$	$\delta_{j,3}$	$\delta_{j,12}$	$\delta_{j,13}$	$\delta_{j,23}$	$\delta_{j,123}$	Note
DINA	1				1				
DINO	1	1	1		1				$\beta_{j,1} = \beta_{j,2} = -\beta_{j,12}$
G-DINA	1	1	1		1				$\Psi(x) = x$
NC-RUM	1	1	1						$\Psi(x) = \exp(x)$
C-RUM	1	1	1						$\Psi(x) = \text{logit}^{-1}(x)$

with 1 indicating that the corresponding coefficient is active, i.e., its β value is nonzero, and 0 indicating that the coefficient is inactive, i.e., its β value is zero. The intercept $\beta_{j,0}$ is usually assumed to be active, so $\delta_{j,0}$ is fixed at 1.

From now on, we will use $\boldsymbol{B}_{J\times 2^K}$, referred to as the *coefficient matrix*, to denote the collection of $\boldsymbol{\beta}_j$'s, and $\boldsymbol{\Delta}_{J\times 2^K}$, referred to as the *sparsity matrix*, to denote the collection of $\boldsymbol{\delta}_j$'s.

2.2. Sparsity Patterns

Many popular CDMs can be reparameterized as special cases of SLCM with particular sparsity patterns, including the DINA model, the DINO model, the G-DINA (Generalized DINA) model, the NC-RUM (reduced noncompensatory reparameterized unified model) (DiBello et al. 1995; Rupp et al. 2010), and the C-RUM (Compensatory-RUM) (Hagenaars 1993; Maris 1999).

In Table 1, we present the sparsity patterns of these aforementioned CDMs, if being reparameterized as SLCMs. For simplicity, we assume there are $K = 3$ latent skills, and the first two skills are relevant to each item. The detailed derivation for reparameterization of CDMs is provided in "Appendix A".

The DINA model is a conjunctive model with $\theta_{j,\boldsymbol{\alpha}}$ taking only two possible values: one for students who have all the relevant skills, and the other for students who miss any of the relevant skills. If reparameterized as an SLCM, the DINA model is a sparse model with only one active coefficient, the highest interaction term that involves all the relevant skills, in addition to the intercept. In contrast, the DINO model is a disjunctive model, in which one value of $\theta_{j,\boldsymbol{\alpha}}$ is for students who have at least one of the relevant skills, and the other is for students who miss all of the relevant skills. If reparameterized as an SLCM, the DINO model turns out to be a dense model with all the main effects and interactions among the relevant skills being active. In addition, there are some constraints on the coefficients: the active coefficients in odd orders are all equal and positive, while the values of those in even orders are the additive inverse of the odd ones.

The G-DINA model is a generalization of the DINA model, which assumes $\theta_{j,\boldsymbol{\alpha}}$ can be decomposed into the sum of the effects from relevant skills and their interactions. So similar to the DINO model, the G-DINA model is also a dense model with all the main effects and interactions among the relevant skills being active. The NC-RUM model assumes that missing any of the relevant skills reduces the positive response probability by a multiplicative penalty term. Xu (2017) has shown that the NC-RUM model is equivalent to a log-link additive model with just the main effects. Similar to the NC-RUM model, the C-RUM model is also an additive model involving only the main effects, but with a logit link function.

An advantage of SLCM is that it contains all CDMs that admit a K-dimensional binary attribute profile, including the ones that do not fall into any of the aforementioned CDMs. Consider two sparsity patterns shown in Table 2 and assume the active coefficients are all positive. In the first case, mastering skill 1 alone increases the positive response probability, while mastering

TABLE 2.
Other sparse patterns of $\boldsymbol{q} = (1, 1, 0)$

$\delta_{j,0}$	$\delta_{j,1}$	$\delta_{j,2}$	$\delta_{j,3}$	$\delta_{j,12}$	$\delta_{j,13}$	$\delta_{j,23}$	$\delta_{j,123}$
1	1			1			
1		1		1			

skill 2 alone does not; mastering skill 2, however, increases the positive response probability conditioning on the mastery of skill 1. The second case has a similar interpretation. These two cases do not belong to any of the CDMs mentioned above, but can be modeled by SLCM.

2.3. *From $\boldsymbol{Q}$-matrix to $\boldsymbol{\Delta}$-matrix*

Although the $\boldsymbol{Q}$-matrix has been widely used for cognitive diagnostic modeling, it only provides partial information regarding the item–attributes relationship (Fang et al. 2019). Given $\boldsymbol{q}_j$, we know the relevant attributes, but it is not clear how they interact with each other to affect the positive response probability without specifying a particular CDM. In practice, it is challenging to pre-specify a CDM when no prior information of the test is available. Furthermore, it is possible that different items in a test follow different CDMs, so there is no single CDM that is appropriate for all the items in the test.

In SLCM, the sparsity matrix $\boldsymbol{\Delta}_{J\times 2^K}$, which does not require any pre-specified CDM, provides a more general and informative description of the item–attributes relationship. For example, all the CDMs listed in Tables 1 and 2 have the same $\boldsymbol{q}_j = (1, 1, 0)$ but their structure vectors $\boldsymbol{\delta}_j$'s could be different. In cases, the $\boldsymbol{Q}$-matrix is preferred as a summary of the relevant skills, we extract $\boldsymbol{q}_j$ based on $\boldsymbol{\delta}_j$ as follows: for any attribute k, if there exists a subset of the K attributes, $\{k_1, k_2, \ldots, k_l\} \subseteq \{1, \ldots, K\}$, such that, $k \in \{k_1, k_2, \ldots, k_l\}$ and $\delta_{j,k_1k_2\ldots,k_l} = 1$, then $q_{jk} = 1$, otherwise, $q_{jk} = 0$. That is, $q_{jk} = 1$ if any element of $\boldsymbol{\delta}_j$ that is relevant to skill k is nonzero.

3. Identifiability

Model identifiability is of great importance in the study of CDMs. In Statistics, a model is identifiable if it is theoretically possible to learn the true values of its underlying parameters after obtaining an infinite number of observations. Mathematically, this is equivalent to saying that different values of the parameters must correspond to different probability distributions of the observable variables. Identifiability conditions are technical restrictions, under which the model is identifiable. In this section, we establish a set of identifiability conditions of SLCM in terms of $\boldsymbol{\Delta}$. We first introduce the identifiability issue encountered in CDMs in Sect. 3.1 and review prior research in Sect. 3.2. Then, we introduce the generic identifiability in the context of SLCM in Sect. 3.3 and propose a set of generic identifiability conditions in Sect. 3.4. The proof sketch of generic identifiability is provided in Sect. 3.5, and detailed proofs are given in "Appendix B."

3.1. *Identifiability Issue*

In CDMs, the observable variables are $\boldsymbol{Y} = (Y_1, \ldots, Y_J)$,, and the parameters of interest are the latent class proportion vector $\boldsymbol{\pi}$ and the coefficient matrix $\boldsymbol{B}$. Denote the parameter space of $(\boldsymbol{\pi}, \boldsymbol{B})$ by

$$\Omega(\boldsymbol{\pi}, \boldsymbol{B}) = \{(\boldsymbol{\pi}, \boldsymbol{B}) : \boldsymbol{\pi} \in \Omega(\boldsymbol{\pi}), \boldsymbol{B} \in \Omega(\boldsymbol{B})\},$$

where $\Omega(\boldsymbol{\pi}) = \{x \in \mathbb{R}^{2^K} : x_1 + \cdots + x_{2^K} = 1, x_i > 0\}$ is a $(2^K - 1)$-dimensional simplex and $\Omega(\boldsymbol{B})$ is the parameter space of the coefficient matrix $\boldsymbol{B}$, which could be the whole space $\mathbb{R}^{J\times 2^K}$ or a subset of $\mathbb{R}^{J\times 2^K}$ if we constrain the $\boldsymbol{Q}$-matrix or the $\boldsymbol{\Delta}$-matrix.

Given an attribute profile/class $\boldsymbol{\alpha}$, the joint distribution of $\boldsymbol{Y} = \{Y_1, \ldots, Y_J\}$ is a product of Bernoulli distributions, which can be described by a J-dimensional $2 \times \cdots \times 2$ table

$$\mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{B}) = \bigotimes_{j=1}^{J} (\theta_{j,\boldsymbol{\alpha}}, 1 - \theta_{j,\boldsymbol{\alpha}})$$

where $\bigotimes$ denotes the Kronecker product. The *marginal distribution* of $\boldsymbol{Y}$ over different classes is given by

$$\mathbb{P}(\boldsymbol{\pi}, \boldsymbol{B}) = \sum_{\boldsymbol{\alpha}} \pi_{\boldsymbol{\alpha}} \mathbb{P}_{\boldsymbol{\alpha}}(\boldsymbol{B}).$$

Definition 1. (Identifiable). A parameter set $(\boldsymbol{\pi}, \boldsymbol{B}) \in \Omega(\boldsymbol{\pi}, \boldsymbol{B})$ is *identifiable*, if

$$\mathbb{P}(\boldsymbol{\pi}, \boldsymbol{B}) = \mathbb{P}(\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{B}}) \Leftrightarrow (\boldsymbol{\pi}, \boldsymbol{B}) \sim (\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{B}}),$$

where $(\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{B}})$ is another parameter set from $\Omega(\boldsymbol{\pi}, \boldsymbol{B})$ and "$\sim$" means the two sets of parameters are identical up to label switching.

3.2. Prior Work

CDMs are mixtures of Bernoulli distributions, which are known to be non-identifiable even if we ignore the non-identifiability due to label switching (Teicher et al. 1961; Yakowitz and Spragins,1968; Goodman 1974; Gyllenberg et al. 1994b). It is, however, possible for CDMs to be identifiable if certain restrictions on the $\boldsymbol{Q}$-matrix and/or $\{\theta_{j,\boldsymbol{\alpha}}\}$ are satisfied. For instance, as discussed in Chiu et al. (2009) and Liu et al. (2013), the completeness of $\boldsymbol{Q}$-matrix, a condition that requires the $\boldsymbol{Q}$-matrix to contain an identity matrix after row permutation, is believed to be a necessary condition for the identifiability of DINA models.

Different identifiability conditions have been proposed for different CDMs in the literature. For example, Liu et al. (2013) studied the DINA model with complete knowledge of guessing parameters, and Chen et al. (2015) studied the DINA/DINO models. For general CDMs, Xu (2017) and Xu and Shang (2017) proposed a set of identifiability conditions in terms of the $\boldsymbol{Q}$-matrix and the monotonicity constraints. But their conditions are too strong in practice, especially when K is large, which will be discussed in the next paragraph. Another set of identifiability conditions for general CDMs was proposed by Fang et al. (2019), which are still stronger than ours and also difficult to be incorporated in algorithms since they are expressed in terms of $\{\theta_{j,\boldsymbol{\alpha}}\}$.

The *monotonicity constraints* imposed by Xu (2017) and Xu and Shang (2017) are as follows:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}\in\mathbb{S}_0} \theta_{j,\boldsymbol{\alpha}} \geq \theta_{j,\mathbf{0}}, \\ &\max_{\boldsymbol{\alpha}\in\mathbb{S}_0} \theta_{j,\boldsymbol{\alpha}} < \min_{\boldsymbol{\alpha}\in\mathbb{S}_1} \theta_{j,\boldsymbol{\alpha}} = \max_{\boldsymbol{\alpha}\in\mathbb{S}_1} \theta_{j,\boldsymbol{\alpha}}, \end{aligned} \tag{4}$$

where $\mathbb{S}_0 = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j, \boldsymbol{\alpha} \succ \mathbf{0}\}$ is the set of classes with at least one skill but not mastering all the relevant skills and $\mathbb{S}_1 = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \succeq \boldsymbol{q}_j\}$ is the set of classes mastering all the relevant skills, where the

notation "$\boldsymbol{\alpha} \succeq \tilde{\boldsymbol{\alpha}}$" means $\alpha_k \geq \tilde{\alpha}_k$ for all $k = 1, \ldots, K$, and "$\alpha \succ \tilde{\alpha}$" means $\alpha_k \geq \tilde{\alpha}_k$ for all k and there exists at least one k_0, such that $\alpha_{k_0} > \tilde{\alpha}_{k_0}$. Constraints (4) imply that students without any skill (i.e., $\boldsymbol{\alpha} = \mathbf{0}$) should have the lowest probability to give a positive answer to item j, students with all the relevant skills ($\boldsymbol{\alpha} \in \mathbb{S}_1$) should have the highest probability, and additional skills beyond the relevant skills are not expected to increase the probability. Under such constraints, the model identifiability is guaranteed (Xu 2017; Xu and Shang 2017) when the $\boldsymbol{Q}$-matrix (after row swapping) takes the form of

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{I}_K \\ \boldsymbol{I}_K \\ \boldsymbol{Q}' \end{pmatrix}, \tag{5}$$

where $\boldsymbol{I}_K$ is an identity matrix of size K and there are additional constraints on $\boldsymbol{Q}'$. However, condition (5) implies that the first $2K$ items degenerate into items in a DINA/DINO model with only one relevant skill. When K is relatively large, e.g., close to $J/2$, such a $\boldsymbol{Q}$-matrix essentially forces the model to be a DINA/DINO model, which is contrary to the original intention of general CDMs.

3.3. Generic Identifiability

Generic identifiability, which is less stringent than Definition 1, is introduced in Allman et al. (2009). Generic identifiability allows some parameter values to be non-identifiable as long as these exceptional values are of measure zero with respect to the parameter space. As pointed out in Allman et al. (2009), generic identifiability of a model is generally sufficient in practice, since one is unlikely to face the non-identifiability problem when almost all parameters (except a measure zero set) are identifiable. Allman et al. (2009) has proved that a general mixture of Bernoulli products is generically identifiable, provided that the number of items is larger than twice the number of classes.

However, there is a gap between the result established in Allman et al. (2009) and SLCM studied in this paper. The parameter space of the former is fixed, corresponding to $\Omega(\boldsymbol{B}) = \mathbb{R}^{J \times 2^K}$ in the context of SLCM, whereas the parameter space $\Omega(\boldsymbol{B})$ of an SLCM varies with $\boldsymbol{\Delta}$ and is usually of dimension less than $J \times 2^K$. In other words, unless $\boldsymbol{\Delta} = \mathbf{1}_{J \times 2^K}$ (i.e., each coefficient in $\boldsymbol{B}$ is active), for any other $\boldsymbol{\Delta}$, the whole parameter space of $\boldsymbol{B}$ is a measure zero subspace of $\mathbb{R}^{J \times 2^K}$. So the result from Allman et al. (2009) is not applicable to SLCM, as it is meaningless to discuss measure zero subspace without a fixed parameter space.

To discuss generic identifiability for SLCM, we need to first define the parameter space by taking into consideration of their sparsity structures. For a given sparsity structure $\boldsymbol{\Delta}$, we denote its corresponding parameter space by

$$\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B}) = \{(\boldsymbol{\pi}, \boldsymbol{B}) : \boldsymbol{\pi} \in \Omega(\boldsymbol{\pi}), \boldsymbol{B} \in \Omega_{\boldsymbol{\Delta}}(\boldsymbol{B})\},$$

where $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{B})$ consists of the coefficient matrices that can only have nonzero entries at positions where the corresponding elements in $\boldsymbol{\Delta}$ is 1. So it suffices to think $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{B}) = \mathbb{R}^{|\boldsymbol{\Delta}|}$, where $|\boldsymbol{\Delta}|$ denotes the total sum of the entries in $\boldsymbol{\Delta}$.

Let $C_{\boldsymbol{\Delta}}$ denote the set of non-identifiable parameters from $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B})$:

$$\begin{aligned} C_{\boldsymbol{\Delta}} = \{(\boldsymbol{\pi}, \boldsymbol{B}) : \ & \mathbb{P}(\boldsymbol{\pi}, \boldsymbol{B}) = \mathbb{P}(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{B}}), \quad (\boldsymbol{\pi}, \boldsymbol{B}) \not\sim (\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{B}}), \\ & (\boldsymbol{\pi}, \boldsymbol{B}) \in \Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B}), \quad (\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{B}}) \in \Omega_{\tilde{\boldsymbol{\Delta}}}(\boldsymbol{\pi}, \boldsymbol{B})\}. \end{aligned}$$

Note that the non-identifiability of a parameter set $(\boldsymbol{\pi}, \boldsymbol{B}) \in C_{\boldsymbol{\Delta}}$ could be due to another parameter set $(\tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{B}})$ with a different sparsity structure $\tilde{\boldsymbol{\Delta}}$.

If the non-identifiable set $C_{\boldsymbol{\Delta}}$ is of measure zero with respect to $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B})$, we say $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B})$ is a generically identifiable parameter space.

Definition 2. (Generically identifiable). The parameter space $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B})$ is *generically identifiable*, if the Lebesgue measure of $C_{\boldsymbol{\Delta}}$ with respect to $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B})$ is zero.

To distinguish the two definitions of identifiability, in the following text, we refer to the identifiability defined in Definition 1 as *strict identifiability*.

3.4. Identifiability Conditions

In this section, we discuss two sets of conditions for generic identifiability and strict identifiability. We start with the following two conditions needed for generic identifiability.

(G1) The true sparsity matrix $\boldsymbol{\Delta}$ takes the form of $\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \boldsymbol{\Delta}' \end{pmatrix}$ after row swapping, where $\boldsymbol{\Delta}'$ is a $(J - 2K) \times 2^K$ binary matrix and $\boldsymbol{D}_1, \boldsymbol{D}_2 \in \mathbb{D}_g$ with

$$\mathbb{D}_g = \left\{ \boldsymbol{D} \in \{0, 1\}^{K \times 2^K} : \boldsymbol{D} = \begin{bmatrix} * \; 1 \; * \; \ldots \; * \; \ldots \; * \\ * \; * \; 1 \; \ldots \; * \; \ldots \; * \\ \vdots \; \vdots \quad \ddots \quad \vdots \\ * \; * \; * \; \ldots \; 1 \; \ldots \; * \end{bmatrix} \right\},$$

where $*$ can be either 0 or 1.

(G2) For any $k = 1, 2, \ldots, K$, there exists a $j_k > 2K$, such that $\delta_{j_k,k} = 1$.

Theorem 1 *The parameter space $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B})$ is generically identifiable, if conditions (G1) and (G2) are satisfied.*

Remark 1 Theorem 1 does not require the monotonicity constraints (4), but it remains valid if, in addition, the monotonicity constraints are imposed on $\boldsymbol{B}$.

Remark 2 Theorem 1 applies to any CDM that has a real analytic link function, including but not limited to the probit link, the logit link, the log link, and the identity link.

Remark 3 If the $*$ entries in condition (G1) are all 1's, i.e., $\boldsymbol{\Delta} = \mathbf{1}_{J \times 2^K}$, the corresponding model is a general mixture of Bernoulli products. Theorem 1 implies that a mixture of Bernoulli products is generically identifiable up to label switching, provided that $J \geq 2K + 1$, which is consistent with Corollary 5 in Allman et al. (2009).

If the $*$ entries in condition (G1) are all 0's except the intercept, the corresponding $\boldsymbol{Q}$-matrix is similar to (5), the $\boldsymbol{Q}$-matrix from Xu (2017) and Xu and Shang (2017) for strict identifiability. In fact, the technique we use to prove generic identifiability (Theorem 1) can be easily extended to show strict identifiability (Theorem 2) under the monotonicity constraints (4) and the following two conditions.

(S1) The true sparsity matrix $\boldsymbol{\Delta}$ takes the form of $\boldsymbol{\Delta} = \begin{pmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \\ \boldsymbol{\Delta}' \end{pmatrix}$ after row swapping, where $\boldsymbol{\Delta}'$ is a $(J - 2K) \times 2^K$ binary matrix and $\boldsymbol{D}_1 = \boldsymbol{D}_2 = \boldsymbol{D}_s$ with

$$\boldsymbol{D}_s = \begin{bmatrix} 1\,1\,0 \ldots 0 \ldots 0 \\ 1\,0\,1 \ldots 0 \ldots 0 \\ \vdots\ \vdots \quad \ddots \quad\quad \vdots \\ 1\,0\,0 \ldots 1 \ldots 0 \end{bmatrix}.$$

(S2) For any two classes of subjects, there exists at least one item in $\boldsymbol{\Delta}'$ such that they have different positive response probabilities.

Theorem 2 (Strict identifiability) *Any parameter from* $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{\pi}, \boldsymbol{B})$ *is strictly identifiable, if conditions (S1) and (S2) and the monotonicity constraints* (4) *hold.*

Theorem 2 is similar to the strict identifiability results in Xu (2017) and Xu and Shang (2017), although our proof technique, which is based on tensor product, is different from theirs. The proof of Theorem 2 is given in "Appendix B.2."

3.5. Proof of Generic Identifiability

In this section, we describe the core of our proof for Theorem 1, which is based on the results in Kruskal (1976, 1977) for the uniqueness of tensor decomposition and the tripartition approach in Allman et al. (2009).

First, we introduce some notation.

Definition 3 The *class-response matrix* $\boldsymbol{M}(\boldsymbol{\Delta}, \boldsymbol{B})$ is defined as a matrix of size $2^K \times 2^J$, where the entries are indexed by a row index $\boldsymbol{\alpha}$ and a column index $\boldsymbol{y}$. The $\boldsymbol{\alpha}$th row and $\boldsymbol{y}$th column element of $\boldsymbol{M}(\boldsymbol{\Delta}, \boldsymbol{B})$ is the probability that a subject with attribute profile $\boldsymbol{\alpha}$ gives response $\boldsymbol{y}$, i.e.,

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\alpha}, \boldsymbol{B}) = \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{y_j} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-y_j}.$$

Definition 4 For a class-response matrix $\boldsymbol{M}$, the *Kruskal rank* of $\boldsymbol{M}$ is the largest number I such that every I rows of $\boldsymbol{M}$ are independent.

Remark 4 If $\boldsymbol{M}$ is full row rank, the *Kruskal rank* of $\boldsymbol{M}$ is its row rank.

Next, we reformulate a result in Allman et al. (2009) as follows.

Theorem 3 (Allman et al. 2009) *Consider a general latent class model with r classes and J features, where* $J \geq 3$. *Suppose all entries of* $\boldsymbol{\pi}$ *are positive. If there exists a tripartition of the set* $\mathbb{J} = \{1, 2, \cdots, J\}$ *that divides* $\mathbb{J}$ *into three disjoint, nonempty subsets* $\mathbb{J}_1$, $\mathbb{J}_2$, $\mathbb{J}_3$ *such that the Kruskal ranks of the three class-response matrices* $\boldsymbol{M}_1$, $\boldsymbol{M}_2$, *and* $\boldsymbol{M}_3$ *satisfy*

$$I_1 + I_2 + I_3 \geq 2r + 2, \tag{6}$$

where I_s *denote the Kruskal rank of* $\boldsymbol{M}_s$ *for features in* $\mathbb{J}_s$, *then the parameters of the model are uniquely determined, up to label switching.*

We can view Theorem 3 from the perspective of tensor decomposition. The distribution of $\boldsymbol{Y}$ can be represented as a $2^{|\mathbb{J}_1|} \times 2^{|\mathbb{J}_2|} \times 2^{|\mathbb{J}_3|}$-dimensional three-way tensor $\boldsymbol{T}$ according to the tripartition $\mathbb{J}_1$, $\mathbb{J}_2$, and $\mathbb{J}_3$ defined in Theorem 3. The $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3)$th element is the probability of observing $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3)$, namely,

$$\begin{aligned}\boldsymbol{T}_{(\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3)} &= \mathbb{P}(\boldsymbol{Y}_{\mathbb{J}_1} = \boldsymbol{y}_1, \boldsymbol{Y}_{\mathbb{J}_2} = \boldsymbol{y}_2, \boldsymbol{Y}_{\mathbb{J}_3} = \boldsymbol{y}_3 | \boldsymbol{\pi}, \boldsymbol{B}) \\ &= \sum_{\boldsymbol{\alpha}} \pi_{\boldsymbol{\alpha}} \mathbb{P}(\boldsymbol{Y}_{\mathbb{J}_1} = \boldsymbol{y}_1, \boldsymbol{Y}_{\mathbb{J}_2} = \boldsymbol{y}_2, \boldsymbol{Y}_{\mathbb{J}_3} = \boldsymbol{y}_3 | \boldsymbol{B}, \boldsymbol{\alpha}) \\ &= \sum_{\boldsymbol{\alpha}} \pi_{\boldsymbol{\alpha}} \mathbb{P}(\boldsymbol{Y}_{\mathbb{J}_1} = \boldsymbol{y}_1 | \boldsymbol{B}, \boldsymbol{\alpha}) \mathbb{P}(\boldsymbol{Y}_{\mathbb{J}_2} = \boldsymbol{y}_2 | \boldsymbol{B}, \boldsymbol{\alpha}) \mathbb{P}(\boldsymbol{Y}_{\mathbb{J}_3} = \boldsymbol{y}_3 | \boldsymbol{B}, \boldsymbol{\alpha}),\end{aligned}$$

where the last equation is due to the fact that given attribute profile $\boldsymbol{\alpha}$, components of $\boldsymbol{Y}$ are independent. Therefore, identifiability is equivalent to the uniqueness of the following tensor decomposition:

$$\begin{aligned}\boldsymbol{T} &= \sum_{\boldsymbol{\alpha}} \pi_{\boldsymbol{\alpha}} \boldsymbol{M}_{1,\boldsymbol{\alpha}} \otimes \boldsymbol{M}_{2,\boldsymbol{\alpha}} \otimes \boldsymbol{M}_{3,\boldsymbol{\alpha}} \\ &= \sum_{\boldsymbol{\alpha}} \tilde{\boldsymbol{M}}_{1,\boldsymbol{\alpha}} \otimes \boldsymbol{M}_{2,\boldsymbol{\alpha}} \otimes \boldsymbol{M}_{3,\boldsymbol{\alpha}},\end{aligned}$$

where $\boldsymbol{M}_{s,\boldsymbol{\alpha}}$ is the $\boldsymbol{\alpha}$th row of $\boldsymbol{M}_s$, and $\tilde{\boldsymbol{M}}_{1,\boldsymbol{\alpha}} = \pi_{\boldsymbol{\alpha}} \boldsymbol{M}_{1,\boldsymbol{\alpha}}$. Results from Kruskal (1976, 1977) state that if the sum of the Kruskal ranks of $\tilde{\boldsymbol{M}}_1, \boldsymbol{M}_2, \boldsymbol{M}_3$ is larger than or equal to $2r + 2$, then the tensor decomposition is unique up to simultaneous permutation and rescaling of the rows. Since $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3$ are all class-response matrices, of which every row sums to 1, the uniqueness of tensor decomposition implies model identifiability.

Now, we are ready to give our proof. We divide $\boldsymbol{\Delta}$ into $\boldsymbol{D}_1$, $\boldsymbol{D}_2$, $\boldsymbol{\Delta}'$, which correspond to $\mathbb{J}_1$, $\mathbb{J}_2$, $\mathbb{J}_3$, respectively. For this tripartition, both $\mathbb{J}_1$ and $\mathbb{J}_2$ contain K items and their sparsity matrices are from $\mathbb{D}_g$, respectively; the remaining $J - 2K$ items are included in $\mathbb{J}_3$ corresponding to $\boldsymbol{\Delta}'$ satisfying condition (G2). Accordingly, we decompose the parameter space into three parts, $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{B}) = \Omega_{\boldsymbol{D}_1} \otimes \Omega_{\boldsymbol{D}_2} \otimes \Omega_{\boldsymbol{\Delta}'}$. To check inequality (6), we show that under conditions (G1) and (G2), we have $I_1 = 2^K$, $I_2 = 2^K$ and $I_3 \geq 2$ hold almost everywhere in $\Omega_{\boldsymbol{D}_1}, \Omega_{\boldsymbol{D}_2}, \Omega_{\boldsymbol{\Delta}'}$, respectively. Consequently, identifiability holds almost everywhere in $\Omega_{\boldsymbol{\Delta}}(\boldsymbol{B})$.

In Theorem 4, we show that for any $\boldsymbol{D} \in \mathbb{D}_g$, the class-response matrix $\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B})$ with size $2^K \times 2^K$, is of rank 2^K (full rank) almost everywhere in $\Omega_{\boldsymbol{D}}$. Therefore, with condition (G1), we have $I_1 + I_2 = 2 \cdot 2^K$ hold (almost everywhere in $\Omega_{\boldsymbol{D}_1} \otimes \Omega_{\boldsymbol{D}_2}$). To prove Theorem 1, it then suffices to show that with condition (G2), we have $I_3 \geq 2$ (almost everywhere in $\Omega_{\boldsymbol{\Delta}'}$). In fact, under condition (G2), the following statement holds almost everywhere in $\Omega_{\boldsymbol{\Delta}'}$: for any two different classes, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$, there must exist one $j_0 > 2K$, such that $\theta_{j_0,\boldsymbol{\alpha}_1} \neq \theta_{j_0,\boldsymbol{\alpha}_2}$. The exceptional case is that when $\beta_{j_k,k} = 0$ holds for some k, which is of Lebesgue measure zero with respect to $\Omega_{\boldsymbol{\Delta}'}$. Note that the $\boldsymbol{\alpha}$th row of $\boldsymbol{M}_3$ is given by $\boldsymbol{M}_{3,\boldsymbol{\alpha}} = \bigotimes_{j=2K+1}^{J} (\theta_{j,\boldsymbol{\alpha}}, 1 - \theta_{j,\boldsymbol{\alpha}})$. So $\theta_{j_0,\boldsymbol{\alpha}_1} \neq \theta_{j_0,\boldsymbol{\alpha}_2}$ implies $\boldsymbol{M}_{3,\boldsymbol{\alpha}_1} \neq \boldsymbol{M}_{3,\boldsymbol{\alpha}_2}$. Therefore, rows of $\boldsymbol{M}_3$ are unique, which implies that the Kruskal rank of $\boldsymbol{M}_3$ is at least 2, i.e., $I_3 \geq 2$, (almost everywhere in $\Omega_{\boldsymbol{\Delta}'}$).

Theorem 4 *Given $\boldsymbol{D} \in \mathbb{D}_g$, the corresponding class-response matrix, $\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B})$, is of full rank except some values of $\boldsymbol{B}$ from a measure zero set with respect to $\Omega_{\boldsymbol{D}}$, i.e.,*

$$\lambda_{\Omega_{\boldsymbol{D}}}\{\boldsymbol{B} \in \Omega_{\boldsymbol{D}} : \det[\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B})] = 0\} = 0, \tag{7}$$

where $\lambda_{\Omega_{\boldsymbol{D}}}\{A\}$ denotes the Lebesgue measure of set A with respect to $\Omega_{\boldsymbol{D}}$.

The proof of Theorem 4 for general $\boldsymbol{D} \in \mathbb{D}_g$ is given in "Appendix B.1." Next, we look at two special cases.

Example 1 $\boldsymbol{D}_a = \mathbf{1}_{K\times 2^K}$.

Proof $\boldsymbol{D}_a$ implies a general mixture of K-dimensional Bernoulli products with 2^K classes. There exists a one-to-one mapping between the probability matrix $\boldsymbol{\theta} = \{\theta_{j,\boldsymbol{\alpha}}\} \in [0,1]^{K\times 2^K}$ and the coefficient matrix $\boldsymbol{B} \in \Omega_{\boldsymbol{D}_a}(\boldsymbol{B}) = \mathbb{R}^{K\times 2^K}$. Equation (7) holds if the solution set $\det[\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B})] = \det[\boldsymbol{M}(\boldsymbol{\theta})] = 0$ in terms of $\boldsymbol{\theta}$ is of measure zero with respect to $[0,1]^{K\times 2^K}$, i.e.,

$$\lambda_{[0,1]^{K\times 2^K}}\{\boldsymbol{\theta} \in [0,1]^{K\times 2^K} : \det[\boldsymbol{M}(\boldsymbol{\theta})] = 0\} = 0.$$

Note that $\det[\boldsymbol{M}(\boldsymbol{\theta})]$ is a polynomial function of $\boldsymbol{\theta}$ with finite degrees and is not constantly zero for any $\boldsymbol{\theta} \in [0,1]^{K\times 2^K}$, the solution set forms a proper subvariety which must be of dimension less than $K \times 2^K$ (Cox et al. 1994; Allman et al. 2009), hence, of Lebesgue measure zero. □

Example 2. $\boldsymbol{D}_s = (\mathbf{1}_K, \boldsymbol{I}_K, \mathbf{0})$.

Proof. $\boldsymbol{D}_s$ implies a DINA model with K items, K skills and $\boldsymbol{Q} = \boldsymbol{I}_K$. For any item, $\boldsymbol{\beta}_j$ can be reparameterized by the guessing parameter $g_j = \Psi(\beta_{j,0})$ and the slipping parameter $s_j = 1 - \Psi(\beta_{j,0} + \beta_{j,j})$. We can rewrite the class-response matrix as the Kronecker product of K of 2×2 sub-matrices,

$$\boldsymbol{M}(\boldsymbol{D}_s, \boldsymbol{B}) = \bigotimes_{j=1}^{K} \begin{bmatrix} g_j & 1-g_j \\ 1-s_j & s_j \end{bmatrix} := \bigotimes_{j=1}^{K} \boldsymbol{M}^j$$

By the property of Kronecker products, we have $rank(\boldsymbol{M}(\boldsymbol{D}_s, \boldsymbol{B})) = \prod_{j=1}^{K} rank(\boldsymbol{M}^j)$. Therefore, $\boldsymbol{M}(\boldsymbol{D}_s, \boldsymbol{B})$ is of full rank unless there exists at least one j_0, such that $g_{j_0} = 1 - s_{j_0}$, or equivalently $\beta_{j_0,j_0} = 0$. Therefore, the dimension of the exceptional set is less than the dimension of $\Omega_{\boldsymbol{D}_s}$, hence of Lebesgue measure zero. □

Remark 5. In the above two special cases, after re-parameterizing $\boldsymbol{B}$, we simply compare the dimension of the solution set and that of the parameter space. This technique, however, cannot be applied to general settings where the equation $\det[\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B})] = 0$ is not easy to solve.

4. MCMC for Model Estimation

4.1. Bayesian Formulation

Consider a SLCM with N subjects, J items and K skills. Let $\boldsymbol{\alpha}_i$ denote the attribute profile of subject i, and Y_{ij} denote the response of subject i to item j. Symbols like $\boldsymbol{\beta}_j$ and $\boldsymbol{\delta}_j$ are defined in Sect. 2. Throughout, we use subscript $i = 1, \ldots, N$ to index subjects, $j = 1, \ldots, J$ to index items, and $p = 1, \ldots, 2^K$ to index elements of the coefficients with $p = 1$ corresponding to the intercept.

We formulate the Bayesian model as follows.

- The probit model on Y_{ij}:

$$Y_{ij} \mid \boldsymbol{\alpha}_i \sim \text{Bernoulli}\left(\Phi\left(\boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{a}_{\boldsymbol{\alpha}_i}\right)\right),$$

  where $\Phi(\cdot)$ is the probit link function.
- The spike-and-slab prior on $\boldsymbol{B}$:

$$\begin{aligned} \beta_{jp} \mid \delta_{jp} &\sim \begin{cases} \mathcal{N}(0, \sigma_\beta^2) & \delta_{jp} = 1 \\ \mathcal{I}(\beta_{jp} = 0) & \delta_{jp} = 0 \end{cases}, \\ \delta_{jp} \mid \omega &\sim \text{Bernoulli}\,(\omega)\,, \\ \omega &\sim \text{Beta}(w_0, w_1), \end{aligned}$$

  where $\mathcal{I}$ is an indicator function. The intercept is always set active with $\delta_{j1} = 1$. In addition, the prior distribution on $\boldsymbol{B}$ is restricted on $\mathcal{B}(\boldsymbol{B})$, a subset of $\mathbb{R}^{J \times 2^K}$ where the monotonicity constraints (4) and strict/generic identifiability conditions are satisfied.
- The prior on latent attributes $\boldsymbol{\alpha}_i$:

$$\boldsymbol{\alpha}_i \mid \boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\pi}), \quad \boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{d}_0).$$

Here, $(\sigma_\beta^2, w_0, w_1, \boldsymbol{d}_0)$ are user-specified hyper-parameters. A similar sparsity inducing prior specification was used by Culpepper (2019) for general CDMs.

4.2. The Gibbs Sampling Algorithm

Following the data augmentation approach in Albert and Chib (1993), we introduce an augmented variable $Z_{ij} \sim \mathcal{N}(\boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{a}_{\boldsymbol{\alpha}_i}, 1)$ and $Y_{ij} = \mathcal{I}(Z_{ij} > 0)$. In particular, let $\boldsymbol{Z}_j$ denote the N augmented variables for item j, and $\boldsymbol{Z}_j$ can be written as the following Gaussian regression model,

$$\boldsymbol{Z}_j \left| \boldsymbol{\alpha}_{1:N}, \boldsymbol{\beta}_j \right. \sim \mathcal{N}_N\left(\boldsymbol{X}\boldsymbol{\beta}_j, \mathbf{I}_N\right),$$

where $\boldsymbol{X} = \left(\boldsymbol{a}_{\boldsymbol{\alpha}_1}, \ldots, \boldsymbol{a}_{\boldsymbol{\alpha}_N}\right)^{\mathrm{T}}$ is an $N \times 2^K$ design matrix shared for all $j = 1, \ldots, J$.

Although vanilla MCMC algorithms are available for Bayesian variable selection with Gaussian regression models, they are not applicable here. We derive a tailored MCMC scheme for our model, which ensures that each posterior draw of $\boldsymbol{B}$ is from the valid parameter space $\mathcal{B}(\boldsymbol{B})$.

A key step in our algorithm is to sample the indicator variable δ_{jp}. First, we need to decide whether δ_{jp} is allowed to be changed, i.e., whether the identifiability conditions and the monotonicity constraints still hold when δ_{jp} is replaced by $1 - \delta_{jp}$. For strict identifiability or generic identifiability, the explicit conditions of $\boldsymbol{\Delta}$ are given in Sect. 3. Checking whether a $\boldsymbol{\Delta}$ matrix satisfies these conditions is straightforward. If these conditions are satisfied either with $\delta_{jp} = 0$ or 1, then δ_{jp} is allowed to be updated, otherwise leave δ_{jp} unchanged in this iteration.

Next, we derive the sampling distribution for δ_{jp} if it is allowed to be updated. Since δ_{jp} is deterministic if β_{jp} is given, we need to derive the conditional distribution of δ_{jp} given other variables except β_{jp}, namely,

$$\delta_{jp} \left| \boldsymbol{Z}_j, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\beta}_{j(p)}, \omega, \sigma_\beta^2 \right. \sim \text{Bernoulli}\left(\tilde{\omega}_{jp}\right), \tag{8}$$

where $\boldsymbol{\beta}_{j(p)}$ is the coefficient vector $\boldsymbol{\beta}_j$ without the pth element and $\tilde{\omega}_{jp}$ is the Bernoulli parameter we need to determine.

To compute $\tilde{\omega}_{jp}$, we need to evaluate the integrated likelihood of $\boldsymbol{Z}_j$ with respect to β_{jp}, as the evidence for $\delta_{jp} = 1$. Express the density function of $\boldsymbol{Z}_j$ as

$$p\left(\boldsymbol{Z}_j \left| \boldsymbol{\alpha}_{1:N}, \boldsymbol{\beta}_j \right.\right) = (2\pi)^{-\frac{N}{2}} \exp\left[-\frac{1}{2}\left(\widetilde{\boldsymbol{Z}}_j - \boldsymbol{X}_p\beta_{jp}\right)^{\mathrm{T}}\left(\widetilde{\boldsymbol{Z}}_j - \boldsymbol{X}_p\beta_{jp}\right)\right] \tag{9}$$

where $\widetilde{\boldsymbol{Z}}_j = \boldsymbol{Z} - \boldsymbol{X}_{(p)}\boldsymbol{\beta}_{j(p)}$, and $\boldsymbol{X}_{(p)}$ is an $N \times \left(2^K - 1\right)$ matrix that excludes the pth column. If there is no constraint on β_{jp}, it is straightforward to compute the integrated likelihood of (9) with respect to $\beta_{jp} \sim \mathcal{N}(0, \sigma_\beta^2)$. However, with the monotonicity constraints on $\boldsymbol{B}$, it is not clear what kind of values β_{jp} is allowed to take.

Recall the monotonicity constraints on $\boldsymbol{B}$: for $j = 1, \ldots, J$,

$$\min_{\boldsymbol{\alpha}} \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{a}_{\boldsymbol{\alpha}} = \beta_{j1},$$
$$\max_{\boldsymbol{\alpha}:\boldsymbol{a}_{\boldsymbol{\alpha}} \not\succeq \boldsymbol{\delta}_j} \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{a}_{\boldsymbol{\alpha}} < \min_{\boldsymbol{\alpha}:\boldsymbol{a}_{\boldsymbol{\alpha}} \succeq \boldsymbol{\delta}_j} \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{a}_{\boldsymbol{\alpha}}.$$

Next, we show that if β_{jp} is lower bounded, the constraints above are satisfied.

Proposition 1. *Suppose the coefficient matrix in iteration t satisfies identifiability conditions and the monotonicity constraints, $\boldsymbol{B}^{(t)} \in \mathcal{B}(\boldsymbol{B})$, and only β_{jp} ($p > 1$) is sampled in iteration $t + 1$. If $\beta_{jp}^{(t+1)} > L$, then $\boldsymbol{B}^{(t+1)} \in \mathcal{B}(\boldsymbol{B})$. The lower bound L is given by*

$$L = \max\left(\max_{\boldsymbol{\alpha} \in \mathbb{L}_1}(-\gamma'_{j,\boldsymbol{\alpha}}), \max_{\boldsymbol{\alpha} \in \mathbb{L}_0} \gamma'_{j,\boldsymbol{\alpha}} - \gamma'_{j,\boldsymbol{q}_j^{(t)}}\right),$$

where $\gamma'_{j,\boldsymbol{\alpha}} = \boldsymbol{\beta}_j^{(t)T}\boldsymbol{a}_{\boldsymbol{\alpha}} - \beta_{j1}^{(t)} - \beta_{jp}^{(t)}$, $\mathbb{L}_1 = \{\boldsymbol{\alpha} | a_{\boldsymbol{\alpha},p} = 1\}$, $\mathbb{L}_0 = \{\boldsymbol{\alpha} | a_{\boldsymbol{\alpha},p} = 0, \boldsymbol{\alpha} \succ \mathbf{0}\}$ and $a_{\boldsymbol{\alpha},p}$ denotes the pth element of $\boldsymbol{a}_{\boldsymbol{\alpha}}$.

The proof of Proposition 1 is given in "Appendix E." Using Proposition 1, we can show that the Bernoulli parameter in (8) is given by

$$\tilde{\omega}_{jp} = \frac{\Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \omega\left(\frac{\tilde{\sigma}_p}{\sigma_\beta}\right) \Phi\left(\frac{\tilde{\mu}_{jp}-L}{\tilde{\sigma}_p}\right) \exp\left(\frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\tilde{\sigma}_p^2}\right)}{\Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \omega\left(\frac{\tilde{\sigma}_p}{\sigma_\beta}\right) \Phi\left(\frac{\tilde{\mu}_{jp}-L}{\tilde{\sigma}_p}\right) \exp\left(\frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\tilde{\sigma}_p^2}\right) + 1 - \omega},$$

where $\boldsymbol{X}_p$ denotes the pth column of matrix $\boldsymbol{X}$, $\tilde{\sigma}_p^2 = \left(\boldsymbol{X}'_p\boldsymbol{X}_p + \sigma_\beta^{-2}\right)^{-1}$ and $\tilde{\mu}_{jp} = \boldsymbol{X}'_p\widetilde{\boldsymbol{Z}}_j\left(\boldsymbol{X}'_p\boldsymbol{X}_p + \sigma_\beta^{-2}\right)^{-1}$. See "Appendix D" for the detailed derivation.

After updating δ_{jp}, we update β_{jp} based on the full conditional distribution below:

$$\beta_{jp} \left| \boldsymbol{Z}_j, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\beta}_{j(p)}, \sigma_\beta^2, \delta_{jp} \right. \sim \mathcal{N}\left(\tilde{\mu}_{jp}, \tilde{\sigma}_p^2\right)\left[\mathcal{I}\left(\beta_{jp} > L\right)\right]^{\delta_{jp}}\left[\mathcal{I}\left(\beta_{jp} = 0\right)\right]^{1-\delta_{jp}}. \tag{10}$$

We summarize the sampling algorithm of $\boldsymbol{B}$ and $\boldsymbol{\Delta}$ in Algorithm 1, and the full sampling algorithm in Algorithm 2.

Algorithm 1: Sample $\boldsymbol{B}$ and $\boldsymbol{\Delta}$

for $j \leftarrow 1$ *to* J **do**
  for $p \leftarrow 1$ *to* 2^K **do**
    for $\boldsymbol{\alpha} \in \{0,1\}^{2^K}$ **do**
      $\gamma'_{\boldsymbol{\alpha}} \leftarrow \boldsymbol{\beta}_j^{\mathrm{T}} \boldsymbol{a}_{\boldsymbol{\alpha}} - \beta_{j0} - \beta_{jp}$
    end
    $\tilde{\sigma}_p^2 \leftarrow \left(\boldsymbol{X}'_p \boldsymbol{X}_p + \sigma_\beta^{-2}\right)^{-1}$, $\widetilde{\mathbf{Z}}_j = \boldsymbol{Z} - \boldsymbol{X}_{(p)} \boldsymbol{\beta}_{j(p)}$, $\tilde{\mu}_{jp} \leftarrow \boldsymbol{X}'_p \widetilde{\mathbf{Z}}_j \left(\boldsymbol{X}'_p \boldsymbol{X}_p + \sigma_\beta^{-2}\right)^{-1}$.
    $L \leftarrow \max\{\max_{\boldsymbol{\alpha} \in \mathbb{L}_1} -\gamma'_{\boldsymbol{\alpha}}, \max_{\boldsymbol{\alpha} \in \mathbb{L}_0} \gamma'_{\boldsymbol{\alpha}} - \gamma'_{\boldsymbol{q}_j}\}$,
    if *($L \le 0$) and ($\delta'_{jp} = 1 - \delta_{jp}$ satisfies the identifiability conditions)* **then**

$$\tilde{\omega}_{jp} \leftarrow \frac{\Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \omega \left(\frac{\tilde{\sigma}_p^2}{\sigma_\beta^2}\right)^{\frac{1}{2}} \Phi\left(\frac{\tilde{\mu}_{jp} - L}{\tilde{\sigma}_p}\right) \exp\left(\frac{1}{2} \frac{\tilde{\mu}_{jp}^2}{\tilde{\sigma}_p^2}\right)}{\Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \omega \left(\frac{\tilde{\sigma}_p^2}{\sigma_\beta^2}\right)^{\frac{1}{2}} \Phi\left(\frac{\tilde{\mu}_{jp} - L}{\tilde{\sigma}_p}\right) \exp\left(\frac{1}{2} \frac{\tilde{\mu}_{jp}^2}{\tilde{\sigma}_p^2}\right) + 1 - \omega}.$$

      Sample δ_{jp} from Bernoulli $(\tilde{\omega}_{jp})$.
    end
    if $\delta_{jp} = 1$ **then**
      Sample β_{jp} from truncated normal distribution,
      $\mathcal{N}(\tilde{\mu}_{jp}, \tilde{\sigma}_p^2)\mathcal{I}(\beta_{jp} > L)$.
    end
    if $\delta_{jp} = 0$ **then**
      $\beta_{jp} = 0$.
    end
  end
end
return $\boldsymbol{B}$, $\boldsymbol{\Delta}$

4.3. Bayesian Penalization

In this subsection, we show that the spike-and-slab prior, a Bayesian variable selection technique, is equivalent to a mixture of L_0 and L_2 penalty terms on $\boldsymbol{B}$.

Proposition 2. *If σ_β^2, ω are fixed such that $\sigma_\beta^2 > \dfrac{(1-\omega)^2}{2\pi\omega^2}$, and the prior distribution of $\boldsymbol{\pi}$ is Dirichlet($\mathbf{1}$), then the Bayesian MAP estimate of $(\boldsymbol{\pi}, \boldsymbol{B})$ is equivalent to the optimum of the objective function with a mixture of L_0 and L_2 penalty terms on $\boldsymbol{B}$, i.e.,*

$$\arg\max_{\boldsymbol{\pi}, \boldsymbol{B}} p(\boldsymbol{\pi}, \boldsymbol{B} | \boldsymbol{Y}_N) = \arg\max_{\boldsymbol{\pi}, \boldsymbol{B}} \Big[-\ell(\boldsymbol{\pi}, \boldsymbol{B} | \boldsymbol{Y}_N) + \lambda_1 \|\boldsymbol{B}\|_0 + \lambda_2 \|\boldsymbol{B}\|_2 \Big]$$

where $\boldsymbol{Y}_N$ is the observed responses matrix of N test subjects, $\ell(\boldsymbol{\pi}, \boldsymbol{B} | \boldsymbol{Y}_N)$ is the log-likelihood function, and $\lambda_1, \lambda_2 > 0$ are constants.

Proof. The marginal posterior distribution of $(\boldsymbol{\pi}, \boldsymbol{B})$ is written as

$$p(\boldsymbol{\pi}, \boldsymbol{B} | \boldsymbol{Y}_N) \propto f(\boldsymbol{Y}_N | \boldsymbol{\pi}, \boldsymbol{B}) p(\boldsymbol{\pi}) p(\boldsymbol{B} | \sigma_\beta^2, \boldsymbol{\Delta}) p(\boldsymbol{\Delta} | \omega)$$

Algorithm 2: Full Gibbs sampling algorithm

Input: $\boldsymbol{Y}_{N\times J}$, initial values of $\boldsymbol{\pi}$, $\boldsymbol{\alpha}_{1:N}$, $\boldsymbol{B}$, total chain length T, burn-in period b, and hyper-parameters σ_β^2, w_0, w_1, $\boldsymbol{d}_0$.

for $t \leftarrow 1$ *to* T **do**
  for $j \leftarrow 1$ *to* J **do**
    for $\boldsymbol{\alpha} \in \{0,1\}^K$ **do**
      $\theta_{j,\boldsymbol{\alpha}} \leftarrow \Phi(\boldsymbol{a}_{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\beta}_j)$.
    end
  end
  for $i \leftarrow 1$ *to* N **do**
    Sample $\boldsymbol{\alpha}_i$ from the multinomial distribution,
    $\mathbb{P}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}|\boldsymbol{\pi}, \boldsymbol{y}_i) \propto \pi_\alpha \prod_{j=1}^J \theta_{j,\boldsymbol{\alpha}}^{y_{ij}}(1-\theta_{j,\boldsymbol{\alpha}})^{1-y_{ij}}$
  end
  for $\boldsymbol{\alpha} \in \{0,1\}^K$ **do**
    $\tilde{N}_{\boldsymbol{\alpha}} \leftarrow \sum_{i=1}^N \mathbf{1}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha})$
    Sample $\boldsymbol{\pi}$ from Dirichlet$_{2^K}(\tilde{\boldsymbol{N}} + \boldsymbol{d}_0)$
  end
  for $j \leftarrow 1$ *to* J **do**
    Sample $\boldsymbol{Z}_j$ from the truncated normal distribution,
    $\mathcal{N}_N(\boldsymbol{X}\boldsymbol{\beta}_j, \boldsymbol{I}_N)\prod_{i=1}^N[\mathcal{I}(Z_{ij}>0)]^{y_{ij}}[\mathcal{I}(Z_{ij}<0)]^{1-y_{ij}}$.
  end
  Sample $\boldsymbol{B}$ and $\boldsymbol{\Delta}$ using Algorithm 1,
  Sample ω from Beta $\left(\sum_{j,p}(1-\delta_{jp})+\omega_0, \sum_{j,p}\delta_{jp}+\omega_1\right)$.
  if $t > b$ **then**
    $\boldsymbol{B}^{(t-b)} \leftarrow \boldsymbol{B}$, $\boldsymbol{\Delta}^{(t-b)} \leftarrow \boldsymbol{\Delta}$
  end
end
return $\boldsymbol{B}^{(1)}\cdots, \boldsymbol{B}^{(T-b)}$, $\boldsymbol{\Delta}^{(1)}\cdots, \boldsymbol{\Delta}^{(T-b)}$.

Taking natural logarithm on both sides and ignoring the constant, we have

$$\begin{aligned}
\log p(\boldsymbol{B}, \boldsymbol{\pi}|\boldsymbol{Y}_N) =& \ell(\boldsymbol{Y}_N|\boldsymbol{\pi}, \boldsymbol{B}) + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{B}|\sigma_\beta^2, \boldsymbol{\Delta}) + \log p(\boldsymbol{\Delta}|\omega) \\
=& \ell(\boldsymbol{\pi}, \boldsymbol{B}|\boldsymbol{Y}_N) - \|\boldsymbol{B}\|_0 \log\sqrt{2\pi\sigma_\beta^2} - \frac{1}{2\sigma_\beta^2}\|\boldsymbol{B}\|_2 - \|\boldsymbol{B}\|_0 \log\frac{\omega}{1-\omega} \\
=& \ell(\boldsymbol{\pi}, \boldsymbol{B}|\boldsymbol{Y}_N) - \|\boldsymbol{B}\|_0 \log\frac{\omega\sqrt{2\pi\sigma_\beta^2}}{1-\omega} - \frac{1}{2\sigma_\beta^2}\|\boldsymbol{B}\|_2 \\
=& \ell(\boldsymbol{\pi}, \boldsymbol{B}|\boldsymbol{Y}_N) - \lambda_1\|\boldsymbol{B}\|_0 - \lambda_2\|\boldsymbol{B}\|_2,
\end{aligned}$$

where $\lambda_1 = \log\frac{\omega\sqrt{2\pi\sigma_\beta^2}}{1-\omega}$ and $\lambda_2 = \frac{1}{2\sigma_\beta^2}$. □

Remark 6. Although L_0 penalty is the natural choice to account for the model complexity, it is computationally inefficient. For tractable computation, Xu and Shang (2017) proposed to use truncated lasso penalty (TLP) as an approximation, while we use the spike-and-slab prior.

5. Monte Carlo Simulation

5.1. Overview

To test the performance of the proposed algorithm, we employ Monte Carlo simulation, using different attribute sizes ($K = 3, 4$), different sample sizes ($N = 500, 1000, 2000$), and different

correlations among attributes ($\rho = 0, 0.15, 0.25$). The unknown true $\boldsymbol{Q}$-matrices, with $J = 20$ items, satisfying the strict identifiability conditions, are as follows:

$$\boldsymbol{Q} = \begin{pmatrix} 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&1&0\\ 1&0&1\\ 0&1&1\\ 1&1&0\\ 1&0&1\\ 0&1&1\\ 1&1&0\\ 1&0&1\\ 1&1&1\\ 1&1&1\\ 1&1&1 \end{pmatrix} \qquad \boldsymbol{Q} = \begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&1&0&0\\ 0&1&1&0\\ 0&0&1&1\\ 1&0&0&1\\ 1&0&1&0\\ 0&1&0&1\\ 1&1&0&0\\ 0&0&1&1\\ 0&1&1&1\\ 1&0&1&1\\ 1&1&0&1\\ 1&1&1&0 \end{pmatrix}.$$

For the $\rho = 0$ cases, the attribute profile $\boldsymbol{\alpha}$ is generated uniformly from the 2^K classes. In other words, different attributes are independent. For the $\rho > 0$ cases, dependence among attributes is introduced using the method of Chiu et al. (2009). Specifically, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$ is generated from the multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with variance 1 and correlation ρ such that $\boldsymbol{\Sigma} = (1-\rho)\boldsymbol{I}_K + \rho \mathbf{1}_{K\times K}$. The attribute profile $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is given by $\alpha_k = \mathcal{I}(\xi_k \geq 0)$, $k = 1, \ldots, K$.

The positive response probability of each item is set between 0.2 and 0.8 using the method of Xu and Shang (2017). In particular, we set the probability of attribute profiles with K'_j out of the K_j relevant attributes to be $0.2 + (0.8 - 0.2) \times K'_j / K_j$.

Our model does not explicitly include a $\boldsymbol{Q}$-matrix, but in order to compare with other methods, we recover the $\boldsymbol{Q}$-matrix by aggregating $\hat{\boldsymbol{B}}$, the posterior mean of $\boldsymbol{B}$, in the following way:

1. sum up the values of the relevant coefficients

$$\tilde{q}^{(t)}_{jk} = \hat{\beta}^{(t)}_{j,k} + \sum_{k_1} \hat{\beta}^{(t)}_{j,k_1 k} + \sum_{k1<k2} \hat{\beta}^{(t)}_{j,k_1 k_2 k} + \ldots, \quad t = 1, \ldots, T$$

2. standardize

$$\bar{q}_{jk} = \frac{\sum_{t=1}^{\mathrm{T}} \tilde{q}^{(t)}_{jk}}{\max_{k\in\{1,\ldots,K\}} \sum_{t=1}^{\mathrm{T}} \tilde{q}^{(t)}_{jk}}$$

3. identify $\hat{q}_{jk}$ as 1 if $\bar{q}_{jk}$ exceeds a fixed threshold. Here, we set the cutoff to 0.5 in all configurations

$$\hat{q}_{jk} = 1\{\bar{q}_{jk} > \text{cutoff}\} \qquad .$$

TABLE 3.
Recovery of $\mathbf{Q}$ for $K = 3$

ρ	N	Strictly identifiable space				Generically identifiable space			
		Matrix	Item	TPR	FPR	Matrix	Item	TPR	FPR
0	500	0.582	0.974	0.988	0.005	0.586	0.975	0.989	0.006
	1000	0.936	0.997	0.999	0.001	0.946	0.997	0.999	0.001
	2000	1.000	1.000	1.000	0.000	0.998	0.999	1.000	0.000
0.15	500	0.530	0.970	0.988	0.008	0.564	0.971	0.990	0.010
	1000	0.934	0.997	0.999	0.001	0.928	0.996	0.999	0.001
	2000	0.998	1.000	1.000	0.000	0.996	1.000	1.000	0.000
0.25	500	0.540	0.971	0.989	0.009	0.508	0.966	0.990	0.014
	1000	0.960	0.998	0.999	0.001	0.898	0.995	0.999	0.003
	2000	0.994	1.000	1.000	0.000	0.990	1.000	1.000	0.000

TABLE 4.
Recovery of $\boldsymbol{B}$ for $K = 3$ and $K = 4$

ρ	N	Strictly identifiable space				Generically identifiable space			
		$K = 3$		$K = 4$		$K = 3$		$K = 4$	
		RMSE	aBias	RMSE	aBias	RMSE	aBias	RMSE	aBias
0	1000	0.112	0.065	0.114	0.057	0.115	0.071	0.128	0.086
	2000	0.074	0.044	0.091	0.047	0.075	0.046	0.088	0.059
0.15	1000	0.108	0.064	0.108	0.057	0.109	0.069	0.136	0.096
	2000	0.075	0.045	0.091	0.049	0.077	0.049	0.112	0.074
0.25	1000	0.106	0.064	0.107	0.058	0.110	0.071	0.145	0.102
	2000	0.077	0.047	0.095	0.052	0.080	0.052	0.129	0.080

5.2. *Results*

We generated 500 independent replications for each configuration. Table 3 summarizes the results of the recovery of the $\boldsymbol{Q}$-matrix for $K = 3$. The metrics we use here are the same as in Xu and Shang (2017). The column "Matrix" records the matrix-level $\boldsymbol{Q}$-matrix recovery rates of the 500 replications. The column "Item" gives the item-level recovery rates. "TPR" and "FPR" are two entry-level rates. The column "TPR" is the true positive rate, i.e., the proportion of 1's in the true $\boldsymbol{Q}$-matrix correctly estimated; "FPR" is the false positive rate, i.e., the proportion of 0's in the true $\boldsymbol{Q}$-matrix incorrectly estimated as 1's. Table 4 summarizes the results of the recovery of $\boldsymbol{B}$ for $K = 3$ and $K = 4$. RMSE is the averaged root-mean-square error, and aBias is the averaged absolute values of the estimated biases.

For each configuration, we restricted the parameter search to the strictly identifiable space (under strict identifiability conditions) and the generically identifiable space (under generic identifiability conditions), respectively; we report the results in the columns "Strictly Identifiable Space" and "Generically Identifiable Space." The recovery rates of the former are generally higher. Note that the true parameters live in both spaces, and the strictly identifiable space is much smaller than the generic one. The posterior draws might be likely to be closer to the true parameters if we were to restrict the parameter search to a smaller space. However, we observe that such differences of accuracy become smaller when the sample size becomes larger.

TABLE 5.
Computation time (minutes) per replication

	$N = 500$	$N = 1000$	$N = 2000$
$K = 3$	2.94	4.91	8.48
$K = 4$	3.82	6.23	11.38

It is worth mentioning that our MCMC algorithm is quite efficient. It can provide estimates in several minutes. Table 5 gives average computation times for a Markov chain with 30,000 burn-in samples and 10,000 post-burn-in samples in a MacBook Pro (Retina, Early 2015) with 2.9 GHz Intel Core i5 processor.

6. Real Data Analysis

In this section, we apply our algorithm to two real applications. The first dataset is Tatsuoka's Fraction Subtraction dataset (Tatsuoka 1984, 2002). The second dataset is from an experimental IQ test offered on the Open Psychometrics website.[1]

6.1. Fraction Subtraction Data

This dataset contains responses to a set of fraction subtraction items collected from $N = 536$ middle school students. It has been widely analyzed in the literature, e.g., Xu and Shang (2017), Chen et al. (2018), Chen et al. (2015), de la Torre and Douglas (2004).

Table 6 presents the estimated coefficient matrix by our Gibbs sampling algorithm with K set at 3. The estimated $\boldsymbol{B}$ is one of the posterior draws in the Markov chain such that its corresponding $\boldsymbol{\Delta}$ matrix is closest to the average of $\boldsymbol{\Delta}$ among all the posterior draws in the chain. It is not hard to see that the estimated $\boldsymbol{B}$ follows the generic identifiability conditions and the monotonicity constraints.

The three attributes can be roughly interpreted as (i) applying subtraction to the integer and the fraction separately, (ii) determining common denominator, (iii) converting to improper fraction. The detected skills are consistent with the findings of Chen et al. (2015, 2018).

The estimated coefficients illustrate one benefit of focusing on $\boldsymbol{\Delta}$ versus $\boldsymbol{Q}$. That is, it provides insight into how underlying attributes function together. For example, Item 4, $3\frac{1}{2} - 2\frac{2}{3}$, requires improper fraction conversion, followed by common denominator determination and subtraction. In our results, Item 4 has three positive coefficients, β_3, β_{13} and β_{23}. It suggests that mastering improper fraction conversion, skill (iii), helps to solve the item. Conditioning on the mastery of skill (iii), knowing skill (i) and (ii) increases the success rate. But for those who do not master skill (iii), mastery of the other two skills does not compensate. In most previous analyses (e.g., Xu and Shang 2017; Chen et al. 2015, 2018), the estimated $\boldsymbol{Q}$-matrices suggest only the crucial requirement of skill (iii); they do not indicate that skills (i) and (ii) are also relevant.

Another observation is that the signs of interaction terms imply the logic gates of the involved skills. For example, the estimated β_{13}'s of Items 5, 6, and 7 are positive, while those of Items 9 and 10 are negative. In fact, the former three items all have two distinct solution paths: (a) convert the first number to an improper fraction (skill (iii)), and then apply subtraction; or (b) apply subtraction to the integer and the fraction separately (skill (i)). The negative β_{13}'s of those items indicate that

[1] https://openpsychometrics.org/_rawdata/.

TABLE 6.
Estimated $\boldsymbol{B}$ for fraction subtraction data

Item	Content	β_0	β_1	β_2	β_3	β_{12}	β_{13}	β_{23}	β_{123}
1	$\frac{5}{3}-\frac{3}{4}$	−1.74		2.80					0.29
2	$\frac{3}{4}-\frac{3}{8}$	−1.54		2.18		1.46		1.23	
3	$\frac{5}{6}-\frac{1}{9}$	−2.65		3.72		0.03			
4	$3\frac{1}{2}-2\frac{2}{3}$	−0.81			1.12		0.42	0.66	
5	$1\frac{1}{8}-\frac{1}{8}$	−0.99	1.56	0.55	2.24		−0.97		
6	$3\frac{4}{5}-3\frac{2}{5}$	−1.21	2.06	0.72	2.27		−1.55		
7	$4\frac{5}{7}-1\frac{4}{7}$	−1.44	1.99	0.65	1.89		−1.01		
8	$4\frac{3}{5}-3\frac{4}{10}$	−0.68	0.09	0.42		0.64		0.89	
9	$3-2\frac{1}{5}$	−2.48	0.70	1.78	0.83		1.81		
10	$2-\frac{1}{3}$	−2.24	1.02	1.66	1.30		1.49	0.15	
11	$4\frac{4}{12}-2\frac{7}{12}$	−2.14		0.85	1.94				0.71
12	$4\frac{1}{3}-2\frac{3}{4}$	−1.55			2.19			0.66	
13	$7\frac{3}{5}-\frac{4}{5}$	−2.12			3.06	0.80			2.11
14	$4\frac{1}{10}-2\frac{8}{10}$	−1.82	0.65	0.65	1.97		−0.59		
15	$4-1\frac{4}{3}$	−3.11		1.48	1.24		2.09		
16	$4\frac{1}{3}-2\frac{5}{3}$	−2.39			2.90				0.82
17	$3\frac{3}{8}-2\frac{5}{6}$	−2.34	0.18	0.40	0.43		1.15	2.08	−0.09

students who have mastered either skill (i) or (iii) will have high success probabilities; mastery of both is not expected to increase success rates, implying an "or" gate. In contrast, Items 9 and 10 can be solved only by a two-step solution path, i.e., improper fraction conversion (skill(iii)) followed by separate subtractions (skill(i)). The positive β_{13}'s of these items indicate that mastering the two skills together would largely increase the success probability, which implies that these two skills function through an "and" gate.

6.2. Experimental Matrix Reasoning Data

This dataset contains responses of $N = 400$ subjects to a set of IQ test questions. In each question, a matrix with one tile missing is given, as well as eight options for that missing tile. Participants are required to choose the most appropriate tile among the eight options. An example question is given in Fig. 1. We study the $J = 20$ questions for which the corresponding correct rates are larger than 25%. We list the question matrices in Fig. 2. For more details (the options and the correct answers), please refer to the documentation on Open Psychometrics.

We observe that the patterns of Q1, Q2, Q6, Q8, Q10, Q11, Q14-Q20, and Q23-Q25 are row-wise; as a result, for each question, for the participants to figure out the correct answer to each such question, they must learn the pattern from the first two rows and apply it to the last row. For example, in Q2, the three tiles in each complete row are the same, so the last tile in the incomplete row must be the same as the ones to its left. In Q6, the tiles from left to right on each row are sequentially rotated 90 degrees in the clockwise direction, so we recover the missing tile by 90 degrees clockwise rotation of the tile to its left. On the other hand, for Q1, Q3, Q4, Q5, Q12,

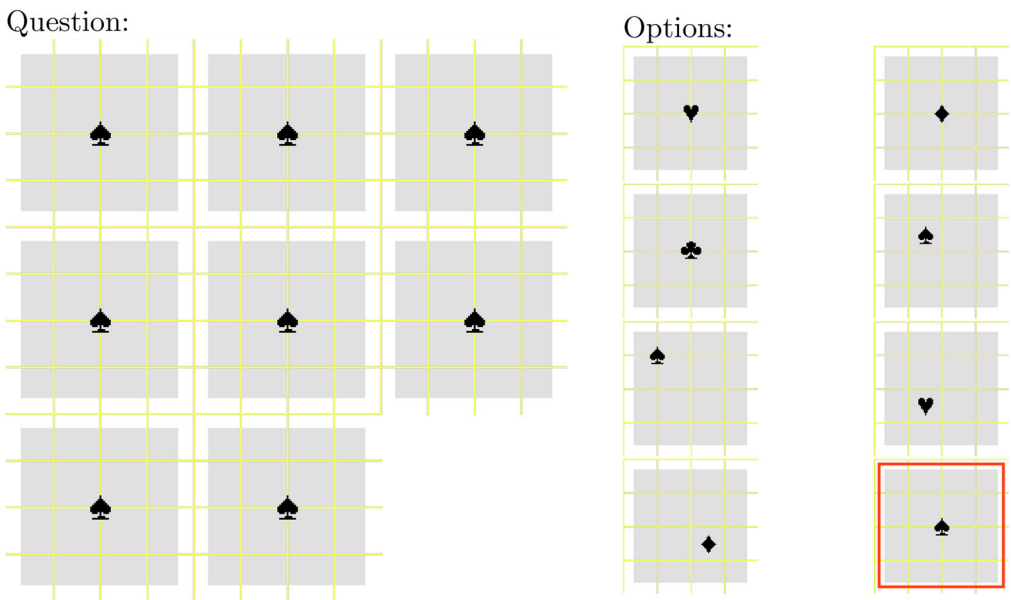

FIGURE 1.
An example question (Q1): the left is the matrix missing one tile, and the right is the eight possible tiles for participants to choose, and the red marked one is the correct answer

and Q13, the participants are required to infer the whole picture of the matrix from the provided tiles and find the one tile that can complete the best overall pattern. For example, in Q3, if the participants can recognize that the whole picture contains one small diamond hollow square and one large spade hollow square, then they are able to choose the proper tile, a tile with a diamond on top left corner and spades on the bottom and the right, to perfectly complete a symmetrical pattern. Similarly, the matrix of Q4 is shaped like a number "2", so we can infer that the last tile should include a horizontal line in the middle. At the same time, for Q6, Q8, Q10, Q11, Q14, Q15, Q19, Q20, Q23, and Q24, the missing tiles are similar to some tiles on the same matrices but with little change, including suites changes (Q19, Q20, Q23), rotation (Q6, Q14), and stretching (Q8). However, for Q1, Q2, Q4, Q5, Q9, Q16, and Q25, the missing tiles are exact the same as one of their neighboring tiles.

From the above analysis, we summarize the four skills to solve these questions as (i) learning the row-wise pattern, (ii) learning the pattern from the whole picture, (iii) changing from neighbors, and (iv) copying from neighbors.

Table 7 presents the estimated coefficients provided by our algorithm with K set at 4; the four estimated attributes are roughly consistent with the four skills analyzed above.

7. Discussion

This work focuses primarily on the study of model identifiability and parameter estimation of SLCM, in which most of the challenges are caused by the restrictions introduced from the context of CDMs. We prove the generic identifiability conditions for SLCM, which relaxes the constraints in the strict identifiability conditions in Xu (2017) and Xu and Shang (2017). We develop a Gibbs sampling algorithm for SLCM, which enforces identifiability conditions and the monotonicity constraints for valid posterior inference. The simulation results demonstrate that our algorithm

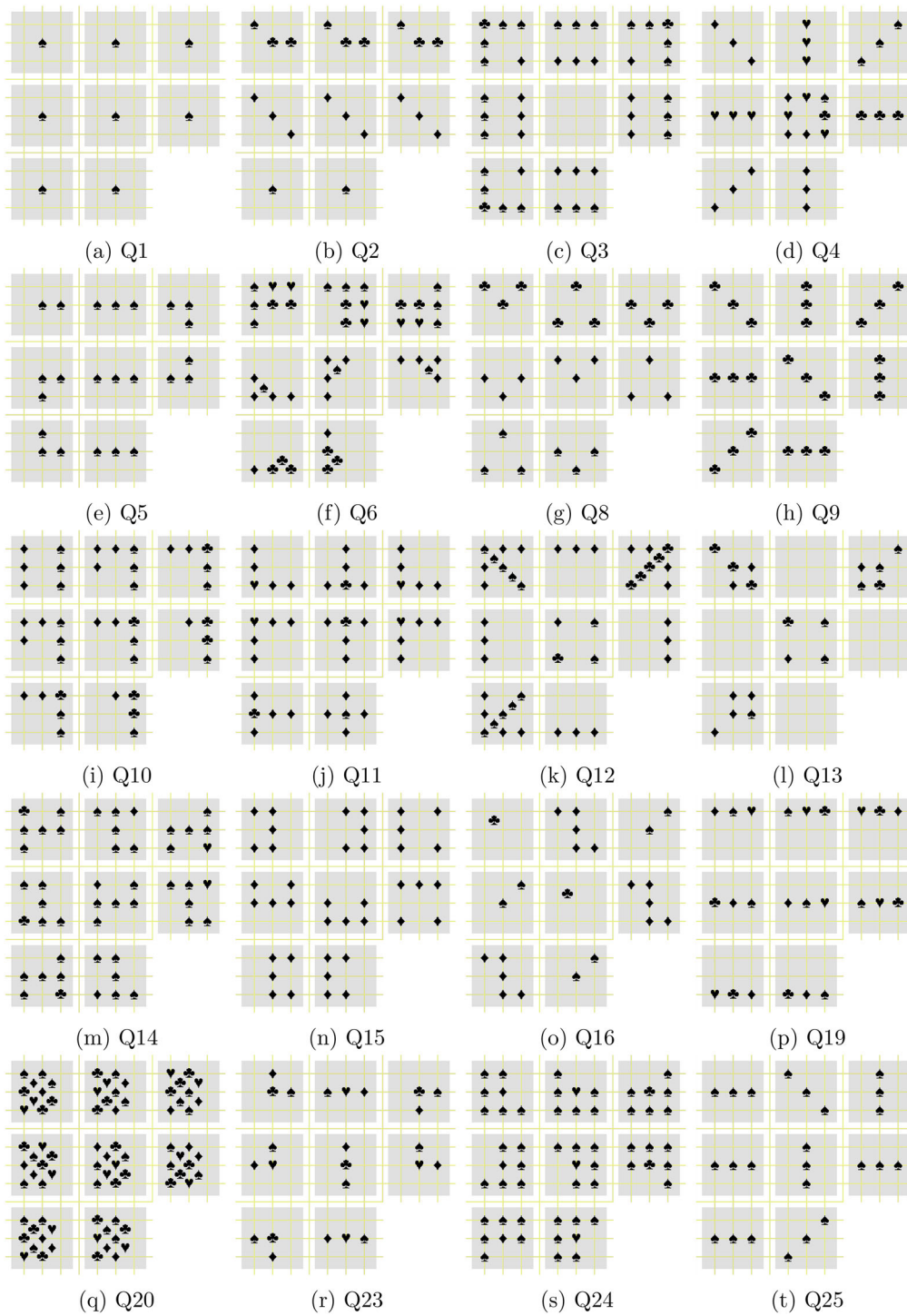

FIGURE 2.
Twenty questions of IQ test

TABLE 7.
Estimated $\boldsymbol{B}$ for IQ test Data: the boldfaced coefficients are the ones consistent with the analysis

	β_0	β_1	β_2	β_3	β_4	β_{12}	β_{13}	β_{14}	β_{23}	β_{24}	β_{34}	β_{123}	β_{124}	β_{134}	β_{234}	β_{1234}
Q1	0.0	**1.3**	**1.0**		**2.0**							0.3				
Q2	0.3	**0.7**	0.7		**1.6**								0.4			
Q3	0.0		**1.6**		0.7											
Q4	−0.6		**1.1**													
Q5	0.1		**1.0**								**0.5**				**0.9**	
Q6	−0.2		0.8	**0.5**												
Q8	−0.6			**0.9**		0.6							0.5			
Q9	0.0	**0.3**			**1.1**											
Q10	−0.7	**0.5**		**1.3**												
Q11	−0.2	**0.4**		**0.6**												
Q12	−1.0		**1.6**			0.2										
Q13	−0.9		**0.9**													
Q14	−0.8			**0.7**		1.2	**0.9**	0.2								
Q15	−0.8	**1.6**		**0.9**												
Q16	−0.2	**1.5**		0.9												
Q19	−0.8	**0.4**		**0.8**			**0.7**									
Q20	−1.0						**0.4**							0.4		
Q23	−1.1	**1.2**														
Q24	−0.9	**1.0**	0.2			0.2	**0.5**					0.1				
Q25	−0.3	**1.0**														

efficiently estimates the model in different configurations, with accuracy comparable to that of alternative models.

Our proof for generic identifiability is based upon the technique of generic identifiability in Allman et al. (2009) and the sufficient conditions of Kruskal (1976, 1977) for the uniqueness of three-way tensor decomposition. We note that the conditions we provide are sufficient but may not be necessary. We note further that Xu (2017) has developed a different technique, working directly on the class-response matrix. An interesting direction for future research is to draw connections between the two sets of techniques, perhaps shedding light on the identification of sufficient and necessary conditions for model identifiability for SLCM.

In this study, the number of attributes, K, is assumed to be known and fixed, a limitation in practice, because it is usually difficult to specify K beforehand, especially when no prior information is available. It is always of interest to study the model with unknown K, and to develop algorithms to select K in model fitting. One promising method is nonparametric Bayesian, which is able to infer from the data an adequate model size. Nevertheless, the model identifiability may continue to present challenges.

A Connections of SLCM to Popular CDMs

In this section, we discuss the connections between SLCM and popular CDMs. To simplify the expression, we assume the relevant skills of item j are $k_1, \ldots, k_R$, i.e., $q_{jk_1} = \cdots = q_{jk_R} = 1$, $q_{jk} = 0$, otherwise.

Example 3. (DINA model). The deterministic input noisy output "and" gate model (Haertel 1989; Junker and Sijtsma 2001) is a conjunctive model. It assumes that a student is most capable of answering question j positively only if he/she masters all of its relevant skills. The item response function takes the following form,

$$\mathbb{P}(Y_j = 1|\boldsymbol{\alpha}, \boldsymbol{q}_j) = (1 - s_j)^{\mathbf{1}(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j)} g_j{}^{\mathbf{1}(\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j)},$$

where $s_j = \mathbb{P}(Y_j = 1|\boldsymbol{\alpha} \succeq \boldsymbol{q}_j)$ is the slipping parameter, which is the probability that a student capable for item j but response negatively and $g_j = \mathbb{P}(Y_j = 1|\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j)$ is the guessing parameter, which is the probability that a non-master answers positively. It is assumed that $g_j < 1 - s_j$ in most applications. The DINA model can be written as

$$\mathbb{P}(Y_j = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}_j) = \Psi\left(\beta_{j,0} + \beta_{j,k_1 \ldots k_R} \alpha_{k_1} \ldots \alpha_{k_R}\right)$$

where only one coefficient, besides the intercept, in $\boldsymbol{\beta}_j$ is active,

$$\delta_{j,0} = \delta_{j,k_1 \ldots k_R} = 1, \quad \delta_{j,p} = 0 \ \text{ otherwise.}$$

The guessing parameter g_j and slipping parameter s_j is given by,

$$g_j = \Psi(\beta_{j,0}), \quad s_j = 1 - \Psi(\beta_{j,0} + \beta_{j,k_1 \ldots k_R}).$$

Example 4. (DINO model). The deterministic input noisy output "or" gate model (Templin and Henson 2006) is a disjunctive model, which assumes that a student is capable to answer question j positively if at least one of the relevant skills is mastered. The item response function is

$$\mathbb{P}(Y_j = 1|\boldsymbol{\alpha}, \boldsymbol{q}_j) = (1 - s_j)^{\mathbf{1}(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{q}_j > 0)} g_j{}^{\mathbf{1}(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{q}_j = 0)}$$

where s_j and g_j are defined the same as in DINA, and $g_j < 1 - s_j$ is assumed. The DINO model can be reparameterized as

$$\mathbb{P}(Y_j = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}_j) = \Psi\left(\beta_{j,0} + \sum_{r=1}^{R} \beta_{j,k_r} \alpha_{k_r} + \sum_{k_r > k_r'} \sum \beta_{j,k_r k_r'} \alpha_{k_r} \alpha_{k_r'} + \cdots + \beta_{j,k_1 \ldots k_R} \prod_{r=1}^{R} \alpha_{k_r}\right)$$

where the coefficients containing only the relevant skills are active,

$$\delta_{j,0} = \delta_{j,k_1} = \cdots = \delta_{j,k_R} = \delta_{j,k_1 k_2} = \cdots = \delta_{j,k_{R-1} k_R} = \cdots = \delta_{j,k_1 \ldots k_R} = 1, \quad \delta_{j,p} = 0 \ \text{ otherwise}$$

The coefficients with odd orders are all equal and positive. The coefficients with even orders are the additive inverse of those with odd orders.

$$\begin{aligned} &\beta_{j,k_1} = \beta_{j,k_2} = \cdots = \beta_{j,k_R} = \beta_{j,k_1 k_2 k_3} = \cdots = \beta_{j,k_{R-2} k_{R-1} k_R} = \cdots \\ &= -\beta_{j,k_1 k_2} = \cdots = -\beta_{j,k_{R-1} k_R} = -\beta_{j,k_1 k_2 k_3 k_4} = \cdots = -\beta_{j,k_{R-3} k_{R-2} k_{R-1} k_R} = \cdots . \end{aligned}$$

The guessing parameter g_j is in the same form of the one in DINA model and slipping parameter s_j is given by $1-\Psi(\boldsymbol{a}_{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\beta}_j)$, with $\boldsymbol{\alpha}$ satisfying $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{q}_j > 0$, which is equivalent to $1-\Psi(\beta_{j,0}+\beta_{j,k_r})$, $r=1,\ldots,R$,

$$g_j = \Psi(\beta_{j,0}), \quad s_j = 1-\Psi(\beta_{j,0}+\beta_{j,k_r}), r=1,\ldots,R.$$

Example 5. (G-DINA model). The DINA model is generalized to the G-DINA model by de la Torre (2011), which takes the form of

$$\mathbb{P}(Y_j=1|\boldsymbol{\alpha},\boldsymbol{q}_j) = \beta_{j,0}+\sum_{k=1}^{K}\beta_{j,k}q_{jk}\alpha_k + \sum\sum_{k>k'}\beta_{j,kk'}q_{jk}\alpha_k q_{jk'}\alpha_{k'} + \cdots + \beta_{j,12\ldots K}\prod_{k=1}^{K}q_{jk}\alpha_k.$$

By using the identity link in Eq. (1), it can be written as

$$\mathbb{P}(Y_j=1|\boldsymbol{\alpha},\boldsymbol{\beta}_j) = \beta_{j,0}+\sum_{r=1}^{R}\beta_{j,k_r}\alpha_{k_r} + \sum\sum_{k_r>k_r'}\beta_{j,k_rk_r'}\alpha_{k_r}\alpha_{k_r'} + \cdots + \beta_{j,k_1\ldots k_R}\prod_{r=1}^{R}\alpha_{k_r}$$

where the coefficients containing only the relevant skills are active,

$$\delta_{j,0}=\delta_{j,k_1}=\cdots=\delta_{j,k_R}=\delta_{j,k_1k_2}=\cdots=\delta_{j,k_{R-1}k_R}=\cdots=\delta_{j,k_1\ldots k_R}=1, \quad \delta_{j,p}=0 \text{ otherwise.}$$

Example 6. (NC-RUM model). Under the reduced noncompensatory reparameterized unified model (DiBello et al. 1995; Rupp et al., 2010), attributes have a noncompensatory relationship with observed response. It assumes missing any relevant skill would inflict a penalty on the positive response probability.

$$\mathbb{P}(Y_j=1|\boldsymbol{\alpha},\boldsymbol{q}_j) = b_j\prod_{k=1}^{K}r_{j,k}^{q_{jk}(1-\alpha_k)}$$

b_j is the positive response probability for students who possess all relevant skills and $r_{j,k}$, $0<r_{j,k}<1$, is the penalty for not mastering kth attribute. As pointed by Xu (2017), by using the exponential link function, NC-RUM can be equivalently written as

$$\mathbb{P}(Y_j=1|\boldsymbol{\alpha},\boldsymbol{\beta}_j) = \exp\left(\beta_{j,0}+\sum_{r=1}^{R}\beta_{j,k_r}\alpha_{k_r}\right),$$

where the main effects of relevant attributes are active,

$$\delta_{j,0}=\delta_{j,k_1}=\cdots\delta_{j,k_R}=1, \quad \delta_{j,p}=0 \text{ otherwise.}$$

The parameters are given by

$$b_j = \exp\left(\beta_{j,0}+\sum_{r=1}^{R}\beta_{j,k_r}\right), \quad r_{j,k} = \begin{cases}\exp(-\beta_{j,k_r}), & \text{if } k\in\{k_1,\ldots,k_R\}\\ 1, & \text{otherwise}\end{cases}$$

Example 7. (C-RUM model). Compensatory-RUM (Hagenaars 1993; Maris 1999) is given by,

$$\mathbb{P}(Y_j = 1|\boldsymbol{\alpha}, \boldsymbol{q}_j) = \frac{\exp\left(\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k} q_{jk} \alpha_k\right)}{\exp\left(\beta_{j0} + \sum_{k=1}^{K} \beta_{j,k} q_{jk} \alpha_k\right) + 1}.$$

Equivalently,

$$\mathbb{P}(Y_j = 1|\boldsymbol{\alpha}, \boldsymbol{q}_j) = \text{logit}^{-1}\left(\beta_{j,0} + \sum_{r=1}^{R} \beta_{j,k_r} \alpha_{k_r}\right)$$

where $\Psi(\cdot)$ is the inverse of the logit function and the main effects of relevant attributes are active,

$$\delta_{j,0} = \delta_{j,k_1} = \cdots \delta_{j,k_R} = 1, \quad \delta_{j,p} = 0 \text{ otherwise.}$$

B Proof of Theorems

In this section, we provide the proof of Theorems 4 and 2.

B.1 Proof of Theorem 4

We first introduce Lemma 5 (Mityagin 2015; Dang 2015) which shows that the solution set of a real analytic function is of Lebesgue measure zero if the function is not constantly 0. Then in Proposition 3, we show that $G_{\boldsymbol{D}}(\boldsymbol{B}) := \det[\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B})]$ is a real analytic function, and in Proposition 4, we show that $G_{\boldsymbol{D}}(\boldsymbol{B})$ is not constantly zero for any $\boldsymbol{B} \in \Omega_{\boldsymbol{D}}(\boldsymbol{B})$ if $\boldsymbol{D} \in \mathbb{D}_g$, so that Lemma 5 applies and Theorem 4 is proved.

Lemma 5. (Mityagin 2015; Dang 2015) *If $f : \mathbb{R}^n \to \mathbb{R}$ is a real analytic function which is not identically zero, then the set $\{\boldsymbol{x} : f(\boldsymbol{x}) = 0\}$ has Lebesgue measure zero.*

Proposition 3. *$G_{\boldsymbol{D}}(\boldsymbol{B}) = \det[\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B})] : \Omega_{\boldsymbol{D}} \to \mathbb{R}$ is a real analytic function of $\boldsymbol{B}$.*

Proof. $G_{\boldsymbol{D}}(\boldsymbol{B})$ is a composition function:

$$G_{\boldsymbol{D}}(\boldsymbol{B}) = \det[\boldsymbol{M}] = h(\boldsymbol{\theta}_{\boldsymbol{\alpha}_0}, \ldots, \boldsymbol{\theta}_{\boldsymbol{\alpha}_{2^K-1}}) = h\left(\Psi(\boldsymbol{B}_{\boldsymbol{D}}\boldsymbol{a}_{\boldsymbol{\alpha}_0}), \ldots, \Psi(\boldsymbol{B}_{\boldsymbol{D}}\boldsymbol{a}_{\boldsymbol{\alpha}_{2^K-1}})\right)$$

where $h(\boldsymbol{\theta}) : [0, 1]^{K \times 2^K} \to \mathbb{R}$ is a polynomial function and $\Psi(\cdot)$ is a CDF.
$\Psi(\cdot)$ is a real analytic function because a CDF is an integral of a real analytic function, and $h(\boldsymbol{\theta})$ is also a real analytic function since it is a polynomial. Therefore, the composition function $G_D(\boldsymbol{B})$ is a real analytic function due to the fact that the composition of real analytic functions is a real analytic function. □

Proposition 4. *If $\boldsymbol{D} \in \mathbb{D}_g$, there exists some $\boldsymbol{B} \in \Omega_{\boldsymbol{D}}(\boldsymbol{B})$, s.t., $G_{\boldsymbol{D}}(\boldsymbol{B}) \neq 0$.*

Proof. Let $\boldsymbol{B}^1 = (\boldsymbol{1}_K, \boldsymbol{I}_K, \boldsymbol{0}) \in \Omega_{\boldsymbol{D}}(\boldsymbol{B}), \forall \boldsymbol{D} \in \mathbb{D}_g$. As shown in Example 2, $\boldsymbol{M}(\boldsymbol{D}, \boldsymbol{B}^1)$ is of full rank, so that $G_{\boldsymbol{D}}(\boldsymbol{B}^1) \neq 0$. □

Remark 7. $G_{\boldsymbol{D}}(\boldsymbol{B}) \not\equiv 0$ is not a trivial conclusion holds for all kinds of $\boldsymbol{D}$. $\boldsymbol{D} \in \mathbb{D}$ is a sufficient condition for $G_{\boldsymbol{D}}(\boldsymbol{B}) \not\equiv 0$. If $\boldsymbol{D} \notin \mathbb{D}$, it is possible $G_{\boldsymbol{D}}(\boldsymbol{B}) \equiv 0$. See the following example.

Example 8. Assume $K = 3$ and the main effect of the first skill is inactive for all items, i.e., $\delta_{j,1} = 0,\ \forall 1 \le j \le K$, then $\boldsymbol{D}_{3\times 8}$ takes the form

$$\begin{bmatrix} * & 0 & * & * & * & * & * & * \\ * & 0 & 1 & * & * & * & * & * \\ * & 0 & * & 1 & * & * & * & * \end{bmatrix}.$$

For any $\boldsymbol{B} \in \Omega_{\boldsymbol{D}}(\boldsymbol{B})$ and any response $\boldsymbol{y} \in \{0,1\}^3$,

$$\boldsymbol{M}_{\boldsymbol{\alpha}=(0,0,0),\boldsymbol{y}}(\boldsymbol{D},\boldsymbol{B}) = \boldsymbol{M}_{\boldsymbol{\alpha}=(1,0,0),\boldsymbol{y}}(\boldsymbol{D},\boldsymbol{B}).$$

So the two rows of $\boldsymbol{M}(\boldsymbol{D},\boldsymbol{B})$ are identical, and $\boldsymbol{M}(\boldsymbol{D},\boldsymbol{B})$ is not full row rank, i.e., $\det[\boldsymbol{M}(\boldsymbol{D},\boldsymbol{B})] \equiv 0$.

By Lemma 5 and Propositions 3 and 4, Theorem 4 is proved.

B.2 Proof of Theorem 2

Proof. As shown in Example 2, for any $\boldsymbol{B} \in \Omega_{\boldsymbol{D}_s}(\boldsymbol{B})$, the corresponding class-response matrix is of full rank, $rank(\boldsymbol{M}(\boldsymbol{D}_s,\boldsymbol{B})) = 2^K$, holds if and only if, for each item, the success probabilities for students with the relevant skill and those without the relevant skill are different. In fact, if the two probabilities are the same, the monotonicity constraints would be violated. Then, using notation from Sect. 3.5, we conclude that under condition (S1) and the monotonicity constraints, $rank(\boldsymbol{M}_1) = rank(\boldsymbol{M}_2) = 2^K$.
For $\boldsymbol{M}_3$, as each element is nonnegative and each row sums to 1. Under condition (S2), there must exist one item j, such that $\theta_{j,\boldsymbol{\alpha}_s} \ne \theta_{j,\boldsymbol{\alpha}_t}$, so $rank(\boldsymbol{M}_3) \ge 2$.

□

C Initialization from the Identifiable Space

Initialization of the sparsity matrix $\boldsymbol{\Delta}^{(0)}_{J\times 2^K}$:

1. activate the intercepts.
   Fix the entries in the first column of $\boldsymbol{\Delta}^{(0)}$ (i.e., $\boldsymbol{\Delta}^{(0)}_{\cdot 1}$) as 1. Denote the remaining $J \times (2^K - 1)$ sub-matrix as $\tilde{\boldsymbol{\Delta}}^{(0)}$.
2. construct $\boldsymbol{D}^{(0)}_1$ and $\boldsymbol{D}^{(0)}_2$.
   Fix the first $2K$ rows of $\tilde{\boldsymbol{\Delta}}^{(0)}$ to be

$$\begin{pmatrix} \boldsymbol{I}_K & \boldsymbol{0} \\ \boldsymbol{I}_K & \boldsymbol{0} \end{pmatrix}.$$

3. construct $\tilde{\boldsymbol{\Delta}}^{\prime(0)}$.
   (a) Randomly select K indexes, $j_1,\ldots,j_K$, from the set $\{2K+1,\ldots,J\}$ with replacement and set $\tilde{\boldsymbol{\Delta}}^{(0)}_{j_k,k} = 1$.

(b) Sample the remaining entries in $\tilde{\boldsymbol{\Delta}}^{'(0)}$ by

$$\delta_{jp}^{(0)}|w^{(0)} \sim \text{Bernoulli}(w^{(0)}), \quad j > 2K, \ (j,p) \notin \{(j_k,k)\}_{k=1}^{K}$$

where $w^{(0)} \sim \text{Beta}(w_0, w_1)$ and w_0, w_1 are the parameters of the prior distribution and are treated as fixed.

(c) Check the row sum. If any row of $\boldsymbol{\Delta}^{'(0)}$ sums to 0, then we randomly pick up an entry on this row and set it at 1.

4. shuffle the rows.
Draw a $J \times J$ permutation matrix $\boldsymbol{P} = (\boldsymbol{e}_{j_1}, \ldots, \boldsymbol{e}_{j_J})$ where $(j_1, \ldots, j_J)$ is a permutation of $(1, \ldots, J)$, and let $\tilde{\boldsymbol{\Delta}}^{(0)} \leftarrow \boldsymbol{P}\tilde{\boldsymbol{\Delta}}^{(0)}$.

The above initialization is designed for strict identifiability conditions. To generate a $\boldsymbol{\Delta}$ under the generic identifiability conditions, we just need to enlarge the range of entries sampling from the prior distribution in step 3b. Specifically, we change step 3b to

Sample the remaining entries in $\tilde{\boldsymbol{\Delta}}^{'(0)}$ by

$$\delta_{jp}^{(0)}|w^{(0)} \sim \text{Bernoulli}(w^{(0)}), \quad (j,p) \notin \{(j_k,k),(k,k),(K+k,k)\}_{k=1}^{K}.$$

Initialization of the coefficients matrix $\boldsymbol{\beta}_{J\times 2^K}^{(0)} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)^{\mathrm{T}}$:

$$\beta_{jp}^{(0)} = 0, \text{ if } \delta_{jp}^{(0)} = 0,$$

$$\beta_{jp}^{(0)}|\delta_{jp}^{(0)} = 1 \propto \mathcal{N}(0, \sigma_\beta^2) I(\beta_{jp}^{(0)} > 0).$$

D Derivation of $\tilde{\omega}_{jp}$

$$\delta_{jp} \left| \boldsymbol{Z}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}_{j(p)}, \omega, \sigma_\beta^2 \right. \sim \text{Bernoulli}\left(\tilde{\omega}_{jp}\right)$$

$$\tilde{\omega}_{jp} = \frac{\omega \int_L^\infty p\left(\boldsymbol{Z}_j \left|\boldsymbol{\alpha}, \boldsymbol{\beta}_j\right.\right) p\left(\beta_{jp} \left|\sigma_\beta^2\right.\right) \mathrm{d}\beta_{jp}}{\omega \int_L^\infty p\left(\boldsymbol{Z}_j \left|\boldsymbol{\alpha}, \boldsymbol{\beta}_j\right.\right) p\left(\beta_{jp} \left|\sigma_\beta^2\right.\right) \mathrm{d}\beta_{jp} + (1-\omega)\, p\left(\boldsymbol{Z}_j \left|\boldsymbol{\alpha}, \boldsymbol{\beta}_{j(p)}, \beta_{jp} = 0\right.\right)}$$

The numerator is

$$
\begin{aligned}
&\omega \int_L^{\infty} p\left(\boldsymbol{Z}_j \left| \boldsymbol{\alpha}, \boldsymbol{\beta}_j\right.\right) p\left(\beta_{jp} \left| \sigma_\beta^2\right.\right) \mathrm{d}\beta_{jp} \\
&= \omega \int_L^{\infty} (2\pi)^{-\frac{N}{2}} \exp\left[-\frac{1}{2}\left(\widetilde{\boldsymbol{Z}}_j - \boldsymbol{A}_p\beta_{jp}\right)'\left(\widetilde{\boldsymbol{Z}}_j - \boldsymbol{A}_p\beta_{jp}\right)\right] \\
&\quad \cdot \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} (2\pi)^{-\frac{1}{2}} \frac{1}{\sigma_\beta} \exp\left(-\frac{\beta_{jp}^2}{2\sigma_\beta^2}\right) \mathrm{d}\beta_{jp} \\
&= (2\pi)^{-\frac{N}{2}} \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \frac{\omega}{\sigma_\beta} \\
&\quad \times \int_L^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{A}_p'\boldsymbol{A}_p + \frac{1}{\sigma_\beta^2}\right)\beta_{jp}^2 - 2\boldsymbol{A}_p'\widetilde{\boldsymbol{Z}}_j\beta_{jp} + \widetilde{\boldsymbol{Z}}_j'\widetilde{\boldsymbol{Z}}_j\right]\right\} \mathrm{d}\beta_{jp} \\
&= (2\pi)^{-\frac{N}{2}} \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \frac{\omega\tilde{\sigma}_p}{\sigma_\beta} \exp\left[-\frac{1}{2}\widetilde{\boldsymbol{Z}}_j'\widetilde{\boldsymbol{Z}}_j + \frac{1}{2}\tilde{\sigma}_p^2\left(\boldsymbol{A}_p'\widetilde{\boldsymbol{Z}}_j\right)^2\right] \\
&\quad \times \int_L^{\infty} (2\pi)^{-\frac{1}{2}} \frac{1}{\tilde{\sigma}_p} \exp\left[-\frac{1}{2\tilde{\sigma}_p^2}\left(\beta_{jp} - \tilde{\sigma}_p^2\boldsymbol{A}_p'\widetilde{\boldsymbol{Z}}_j\right)^2\right] \mathrm{d}\beta_{jp} \\
&= (2\pi)^{-\frac{N}{2}} \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \frac{\omega\tilde{\sigma}_p}{\sigma_\beta} \exp\left(-\frac{1}{2}\widetilde{\boldsymbol{Z}}_j'\widetilde{\boldsymbol{Z}}_j + \frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\tilde{\sigma}_p^2}\right) \int_{-\frac{\tilde{\mu}_{jp}-L}{\tilde{\sigma}_p}}^{\infty} (2\pi)^{-\frac{1}{2}} \\
&\quad \times \exp\left[-\frac{1}{2}\left(\frac{\beta_{jp} - \tilde{\mu}_{jp}}{\tilde{\sigma}_p}\right)^2\right] \mathrm{d}\left(\frac{\beta_{jp} - \tilde{\mu}_{jp}}{\tilde{\sigma}_p}\right) \\
&= (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\widetilde{\boldsymbol{Z}}_j'\widetilde{\boldsymbol{Z}}_j\right) \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \frac{\omega\tilde{\sigma}_p}{\sigma_\beta} \exp\left(\frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\sigma_p^2}\right) \Phi\left(\frac{\tilde{\mu}_{jp} - L}{\tilde{\sigma}_p}\right)
\end{aligned}
$$

where $\tilde{\sigma}_p^2 = \left(\boldsymbol{A}_p'\boldsymbol{A}_p + \sigma_\beta^{-2}\right)^{-1}$ and $\tilde{\mu}_{jp} = \boldsymbol{A}_p'\widetilde{\boldsymbol{Z}}_j\left(\boldsymbol{A}_p'\boldsymbol{A}_p + \sigma_\beta^{-2}\right)^{-1}$. Accordingly, $\tilde{\omega}_{jp}$ is,

$$
\begin{aligned}
\tilde{\omega}_{jp} &= \frac{(2\pi)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\widetilde{\boldsymbol{Z}}_j'\widetilde{\boldsymbol{Z}}_j\right) \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \frac{\omega\tilde{\sigma}_p}{\sigma_\beta} \exp\left(\frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\sigma_p^2}\right) \Phi\left(\frac{\tilde{\mu}_{jp}-L}{\tilde{\sigma}_p}\right)}{(2\pi)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\widetilde{\boldsymbol{Z}}_j'\widetilde{\boldsymbol{Z}}_j\right) \Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \frac{\omega\tilde{\sigma}_p}{\sigma_\beta} \exp\left(\frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\sigma_p^2}\right) \Phi\left(\frac{\tilde{\mu}_{jp}-L}{\tilde{\sigma}_p}\right) + (1-\omega)(2\pi)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\widetilde{\boldsymbol{Z}}_j'\widetilde{\boldsymbol{Z}}_j\right)} \\
&= \frac{\Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \omega\left(\frac{\tilde{\sigma}_p}{\sigma_\beta}\right) \Phi\left(\frac{\tilde{\mu}_{jp}-L}{\tilde{\sigma}_p}\right) \exp\left(\frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\tilde{\sigma}_p^2}\right)}{\Phi\left(\frac{-L}{\sigma_\beta}\right)^{-1} \omega\left(\frac{\tilde{\sigma}_p}{\sigma_\beta}\right) \Phi\left(\frac{\tilde{\mu}_{jp}-L}{\tilde{\sigma}_p}\right) \exp\left(\frac{1}{2}\frac{\tilde{\mu}_{jp}^2}{\tilde{\sigma}_p^2}\right) + 1 - \omega}
\end{aligned}
$$

E Proof of Lower Bound L

In this section, we derive the lower bound L of β_{jp} (Proposition 1).

Suppose at time t, we have a $\boldsymbol{B}^{(t)} \in \mathcal{B}(\boldsymbol{B})$ satisfying the monotonocity constraints (4), and we only sample β_{jp} at time $t+1$ and leave any other coefficient the same as the one at time t, i.e., $\beta_{js}^{(t+1)} = \beta_{js}^{(t)}, \quad \forall s \neq p.$

In what follows, we introduce some notations.
We denote β_{jp} and δ_{jp} as β_p and δ_p, respectively. That is, we omit the subscript of item, j, as the lower bound of coefficient β_{jp} does not depend on any other coefficient of other items.
Let $\gamma_{\boldsymbol{\alpha}} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{a}_{\boldsymbol{\alpha}} - \beta_0 = \Psi^{-1}(\theta_{\boldsymbol{\alpha}}) - \beta_0$ be the sum of the linear component excluding the intercept for class $\boldsymbol{\alpha}$. Further, let $\gamma_{\boldsymbol{\alpha},-p} = \begin{cases} \gamma_{\boldsymbol{\alpha}} - \beta_p & \boldsymbol{\alpha} \in \mathbb{L}_1^p \\ \gamma_{\boldsymbol{\alpha}} & \boldsymbol{\alpha} \in \mathbb{L}_0^p, \end{cases}$ denote the sum of the linear component excluding the intercept and the pth coefficients for class $\boldsymbol{\alpha}$, where $\mathbb{L}_1^p = \{\boldsymbol{\alpha} | \boldsymbol{a}_{\boldsymbol{\alpha},p} = 1\}$ and $\mathbb{L}_0^p = \{\boldsymbol{\alpha} | \boldsymbol{a}_{\boldsymbol{\alpha},p} = 0, \boldsymbol{\alpha} \succ \boldsymbol{\alpha}_0\}$.
We rewrite the monotonicity constraints (4) as follows,

$$\begin{aligned} &\min_{\boldsymbol{\alpha} \succ \boldsymbol{\alpha}_0} \theta_{\boldsymbol{\alpha}} \geq \theta_{\boldsymbol{\alpha}_0} \\ &\max_{\boldsymbol{\alpha} \in \mathbb{S}_0} \theta_{\boldsymbol{\alpha}} < \min_{\boldsymbol{\alpha} \in \mathbb{S}_1} \theta_{\boldsymbol{\alpha}} = \max_{\boldsymbol{\alpha} \in \mathbb{S}_1} \theta_{\boldsymbol{\alpha}} \end{aligned} \tag{$\star$}$$

where $\mathbb{S}_0 = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \not\succeq \boldsymbol{q}, \boldsymbol{\alpha} \succ \boldsymbol{\alpha}_0 = \mathbf{0}\}$ is the set the classes that not mastering all the relevant skills, and $\mathbb{S}_1 = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \succeq \boldsymbol{q}\}$ is the set of classes mastering all the relevant skills.
Note that $\Psi(\cdot)$ is a strictly increasing function, we have the following equivalent form of the monotonicity constraints ($\star$):

$$\min_{\boldsymbol{\alpha} \succ \boldsymbol{\alpha}_0} \gamma_{\boldsymbol{\alpha}} \geq \gamma_{\boldsymbol{\alpha}_0} = 0, \tag{11}$$

$$\gamma_{\boldsymbol{q}} = \max_{\boldsymbol{\alpha} \in \mathbb{S}_1} \gamma_{\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha} \in \mathbb{S}_1} \gamma_{\boldsymbol{\alpha}} > \max_{\boldsymbol{\alpha} \in \mathbb{S}_0} \gamma_{\boldsymbol{\alpha}}. \tag{12}$$

In SLCM, $\boldsymbol{q}$ is uniquely determined by the structure vector $\boldsymbol{\delta}$. Mathematically, $\boldsymbol{q} = \arg\min_{\boldsymbol{\alpha}: \boldsymbol{a}_{\boldsymbol{\alpha} \succeq \boldsymbol{\delta}}} |\boldsymbol{\alpha}|$, where $|\cdot|$ is the cardinality. By such definition, $\gamma_{\boldsymbol{q}} = \max_{\boldsymbol{\alpha} \in \mathbb{S}_1} \gamma_{\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha} \in \mathbb{S}_1} \gamma_{\boldsymbol{\alpha}}$ always holds, therefore, to verify (12), we only need to check,

$$\gamma_{\boldsymbol{q}} > \max_{\boldsymbol{\alpha} \in \mathbb{S}_0} \gamma_{\boldsymbol{\alpha}}. \tag{13}$$

In the following two remarks, we list some observations that are useful in the proof.

Remark 8.
1. $\mathbb{L}_0^p \bigcup \mathbb{L}_1^p = \mathbb{S}_0 \bigcup \mathbb{S}_1 = \{\boldsymbol{\alpha} | \boldsymbol{\alpha} \succ \boldsymbol{\alpha}_0\}$
2. $\boldsymbol{a}_{\boldsymbol{q},p} = 1 \Rightarrow \mathbb{S}_1 \subseteq \mathbb{L}_1^p, \mathbb{S}_0 \supseteq \mathbb{L}_0^p$.

Remark 9.
1. $\forall \boldsymbol{\alpha}, \quad \gamma_{\boldsymbol{\alpha},-p}^{(t)} = \gamma_{\boldsymbol{\alpha},-p}^{(t+1)} := \gamma_{\boldsymbol{\alpha},-p}$
2. $\forall \boldsymbol{\alpha} \in \mathbb{L}_0^p, \quad \gamma_{\boldsymbol{\alpha}}^{(t+1)} = \gamma_{\boldsymbol{\alpha}}^{(t)}$
3. $\forall \boldsymbol{\alpha} \in \mathbb{L}_1^p, \quad \gamma_{\boldsymbol{\alpha}}^{(t+1)} = \gamma_{\boldsymbol{\alpha}}^{(t)} - \boldsymbol{\beta}_p^{(t)} + \boldsymbol{\beta}_p^{(t+1)}$
4. $\forall \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{L}_1^p, \quad \boldsymbol{\gamma}_{\boldsymbol{\alpha}_1}^{(t)} > \boldsymbol{\gamma}_{\boldsymbol{\alpha}_2}^{(t)} \Rightarrow \boldsymbol{\gamma}_{\boldsymbol{\alpha}_1}^{(t+1)} > \boldsymbol{\gamma}_{\boldsymbol{\alpha}_2}^{(t+1)}$
5. $\forall \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{L}_0^p, \quad \boldsymbol{\gamma}_{\boldsymbol{\alpha}_1}^{(t)} > \boldsymbol{\gamma}_{\boldsymbol{\alpha}_2}^{(t)} \Rightarrow \boldsymbol{\gamma}_{\boldsymbol{\alpha}_1}^{(t+1)} > \boldsymbol{\gamma}_{\boldsymbol{\alpha}_2}^{(t+1)}$

In the following lemma, we give the sufficient and necessary condition for (11).

Lemma 6. (Lower bound 1)

$$\min_{\boldsymbol{\alpha} \succ \boldsymbol{\alpha}_0} \gamma_{\boldsymbol{\alpha}}^{(t+1)} \geq \gamma_{\boldsymbol{\alpha}_0}^{(t+1)} = 0$$

holds if and only if

$$\beta_p^{(t+1)} \geq \max_{\boldsymbol{\alpha} \in \mathbb{L}_1^p} (-\gamma_{\boldsymbol{\alpha},-p}). \tag{14}$$

Proof. Since $\boldsymbol{B}^{(t)} \in \mathcal{B}(\boldsymbol{B})$, we have $\min_{\boldsymbol{\alpha} \in \mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t+1)} = \min_{\boldsymbol{\alpha} \in \mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t)} \geq 0$. So we only need to consider $\boldsymbol{\alpha} \in \mathbb{L}_1^p$, such that

$$\min_{\boldsymbol{\alpha} \in \mathbb{L}_1^p} \gamma_{\boldsymbol{\alpha}}^{(t+1)} = \min_{\boldsymbol{\alpha} \in \mathbb{L}_1^p} (\gamma_{\boldsymbol{\alpha},-p}^{(t)} + \beta_p^{(t+1)}) \geq 0,$$

which holds if and only if (14) holds. □

We show the relationship between $\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)}$ and $\gamma_{\boldsymbol{q}^{(t)}}^{(t)}$ in the following lemma.

Lemma 7.

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t)},-p} + \beta_p^{(t+1)}.$$

Proof.

- If $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$, $\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t)},-p} + \beta_p^{(t+1)}$.
- If $\boldsymbol{q}^{(t+1)} \succ \boldsymbol{q}^{(t)}$, it implies $\delta_p^{(t)} = 0$ and $\delta_p^{(t+1)} = 1$. Therefore, $\boldsymbol{q}^{(t)}, \boldsymbol{q}^{(t+1)} \in \mathbb{S}_1^{(t)}$, so that,

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t)} = \gamma_{\boldsymbol{q}^{(t)}}^{(t)} = \gamma_{\boldsymbol{q}^{(t)},-p}.$$

  Then,

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t+1)},-p} + \beta_p^{(t+1)} = \gamma_{\boldsymbol{q}^{(t)},-p} + \beta_p^{(t+1)}.$$

- If $\boldsymbol{q}^{(t+1)} \prec \boldsymbol{q}^{(t)}$, it means $\delta_p^{(t)} = 1$, $\beta_p^{(t+1)} = \delta_p^{(t+1)} = 0$, and $\boldsymbol{a}_{\boldsymbol{q}^{(t+1)}} + \boldsymbol{e}_p = \boldsymbol{a}_{\boldsymbol{q}^{(t)}}$. Then,

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t+1)},-p} = \gamma_{\boldsymbol{q}^{(t)},-p} = \gamma_{\boldsymbol{q}^{(t)},-p} + \beta_p^{(t+1)}.$$

□

Next, we give the sufficient and necessary condition for (13) in the following lemma.

Lemma 8. (Lower bound 2) *Suppose* $\delta_p^{(t+1)} = 1$,

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} > \max_{\boldsymbol{\alpha} \in \mathbb{S}_0^{(t+1)}} \gamma_{\boldsymbol{\alpha}}^{(t+1)},$$

if and only if,

$$\beta_p^{(t+1)} > \max_{\boldsymbol{\alpha} \in \mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha},-p} - \gamma_{\boldsymbol{q}^{(t)},-p}. \tag{15}$$

Proof. Since $\delta_p^{(t+1)} = 1$, by Remark 2, we have $\mathbb{S}_0^{(t+1)} \supseteq \mathbb{L}_0^p$.
It is easy to see that if (12) holds at time $t+1$, then (15) holds, because

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t)},-p} + \beta_p^{(t+1)} > \max_{\boldsymbol{\alpha}\in\mathbb{S}_0^{(t+1)}} \gamma_{\boldsymbol{\alpha}}^{(t+1)} \geq \max_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t+1)} = \max_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha},-p}.$$

Next we show that if (15) holds, then (13) holds at time $t+1$.
Because (12) holds at time t, we have,

$$\gamma_{\boldsymbol{q}^{(t)}}^{(t)} > \max_{\boldsymbol{\alpha}\in\mathbb{S}_0^{(t)}} \gamma_{\boldsymbol{\alpha}}^{(t)} \geq \max_{\boldsymbol{\alpha}\in\mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t)}} \gamma_{\boldsymbol{\alpha}}^{(t)}. \tag{16}$$

Next, we check (13) in two different scenarios.

- If $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(t+1)}$, $\mathbb{S}_0^{(t)} = \mathbb{S}_0^{(t+1)}$, then by (16) and Remark 9.4, we obtain

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t)}}^{(t+1)} > \max_{\boldsymbol{\alpha}\in\mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t+1)}} \gamma_{\boldsymbol{\alpha}}^{(t+1)}.$$

- If $\boldsymbol{q}^{(t)} \prec \boldsymbol{q}^{(t+1)}$, then since $\boldsymbol{q}^{(t+1)}, \in \mathbb{S}_1^{(t)}$, we have $\gamma_{\boldsymbol{q}^{(t+1)}}^{(t)} > \max_{\boldsymbol{\alpha}\in\mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t)}} \gamma_{\boldsymbol{\alpha}}^{(t)}$. By Remark 9.4, we have

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} > \max_{\boldsymbol{\alpha}\in\mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t)}} \gamma_{\boldsymbol{\alpha}}^{(t+1)}.$$

  On the other hand, since $\boldsymbol{\delta}^{(t+1)} = \boldsymbol{\delta}^{(t)} + \boldsymbol{e}_p$,

$$\{\boldsymbol{\alpha}|\boldsymbol{\alpha} \in \mathbb{S}_0^{(t+1)}, \boldsymbol{\alpha} \notin \mathbb{S}_0^{(t)}\} = \{\boldsymbol{\alpha}|\boldsymbol{\alpha} \succeq \boldsymbol{\delta}^{(t)}, \boldsymbol{\alpha} \not\succeq \boldsymbol{\delta}^{(t+1)}\} \subseteq \mathbb{L}_0^p = (\mathbb{L}_1^p)^c,$$

  leading to,

$$\mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t+1)} = \mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t)}.$$

□

Proof of Proposition 1. Suppose $\delta_p^{(t+1)} = 1$, by Lemmas 6 and 8, the monotonicity constraints hold at time $t+1$, if

$$\begin{aligned}\beta_p^{(t+1)} &> \max\left\{\max_{\boldsymbol{\alpha}\in\mathbb{L}_1^p}(-\gamma_{\boldsymbol{\alpha},-p}), \max_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha},-p} - \gamma_{\boldsymbol{q}^{(t)},-p}\right\} \\ &:= \max(L_1, L_2) = L.\end{aligned}$$

In the following two lemmas, we discuss the flipping rule of δ_p.

Lemma 9. (Flipping rule 1) *If $\delta_p^{(t)} = 0$, $\delta_p^{(t+1)} = 0$, the monotonicity constraints hold at time $t+1$ and $L = 0$.*

Proof. The monotonicity constraints hold at time $t+1$, because $\boldsymbol{B}^{(t)} = \boldsymbol{B}^{(t+1)}$ and $\boldsymbol{B}^{(t)}$ satisfy the constraints.

- $L_1 = -\min_{\boldsymbol{\alpha}\in\mathbb{L}_1^p} \gamma_{\boldsymbol{\alpha}}^{(t)} \leq 0$ because (11) holds at t.
- $L_2 = \max_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t)} - \gamma_{\boldsymbol{q}^{(t)}}^{(t)} = 0$ because

$$\gamma_{\boldsymbol{q}^{(t)}}^{(t)} = \min_{\boldsymbol{\alpha}\in\mathbb{S}_1^{(t)}} \gamma_{\boldsymbol{\alpha}}^{(t)} \leq \max_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t)} \leq \max_{\boldsymbol{\alpha}} \gamma_{\boldsymbol{\alpha}}^{(t)} = \gamma_{\boldsymbol{q}^{(t)}}^{(t)},$$

since $\mathbb{L}_0^p \bigcap \mathbb{S}_1^{(t)}$ is not empty.

Therefore, $L = \max(L_1, L_2) = 0$. □

Lemma 10. (Flipping rule 2) *Suppose* $\delta_p^{(t)} = 1, \delta_p^{(t+1)} = 0$. *The monotonicity constraints hold at time* $t+1$ *if* $L \leq 0$.

Proof. If $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(t+1)}$, the statement can be proved easily by Lemma 6 and Lemma 8. We check (11) and (13) in for the case that $\boldsymbol{q}^{(t)} \succ \boldsymbol{q}^{(t+1)}$.

- Since $L_1 = -\min_{\boldsymbol{\alpha}\in\mathbb{L}_1^p} \gamma_{\boldsymbol{\alpha}}^{(t+1)} \leq 0$ and $\min_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t+1)} = \min_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t)} \geq 0$, (11) holds at $t+1$.
- By Remark 9.4, for any $\boldsymbol{\alpha} \in \mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t+1)}$,

$$\gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} = \gamma_{\boldsymbol{q}^{(t)}}^{(t+1)} > \gamma_{\boldsymbol{\alpha}}^{(t+1)} \tag{17}$$

Together with $L_2 = \max_{\boldsymbol{\alpha}\in\mathbb{L}_0^p} \gamma_{\boldsymbol{\alpha}}^{(t+1)} - \gamma_{\boldsymbol{q}^{(t+1)}}^{(t+1)} \leq 0$, (17) holds for any $\boldsymbol{\alpha} \in \mathbb{L}_0^p \bigcup (\mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t+1)})$.

Further, as shown in the proof of Lemma 8, we have $\{\boldsymbol{\alpha} | \boldsymbol{\alpha} \in \mathbb{S}_0^{(t+1)}, \boldsymbol{\alpha} \notin \mathbb{S}_0^{(t)}\} \subseteq \mathbb{L}_0^p$, such that $\mathbb{S}_0^{(t+1)} \subseteq \mathbb{L}_0^p \bigcup (\mathbb{L}_1^p \bigcap \mathbb{S}_0^{(t+1)})$.

□

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679.

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, *37*, 3099–3132.

Carreira-Perpiñán, M., & Renals, S. (2000). Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, *12*, 141–152.

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q-matrix. *Psychometrika*, *83*, 89–108.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633–665.

Cox, D., Little, J., O'Shea, D., & Sweedler, M. (1994). Ideals, varieties, and algorithms. *American Mathematical Monthly*, *101*(6), 582–586.

Culpepper, S. A. (2019). Estimating the cognitive diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika*, *84*, 333–357.

Dang, N. V. (2015). Complex powers of analytic functions and meromorphic renormalization in QFT. arXiv preprint arXiv:1503.00995.

Davier, M. (2005). *A general diagnostic model applied to language testing data*. ETS Research Report Series, *2005*(2)

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353.
DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment, Chapter 15* (pp. 361–389). New York: Routledge.
Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika*, *84*(1), 19–40.
Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.
Gyllenberg, M., Koski, T., Reilink, E., & Verlaan, M. (1994a). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, *31*(2), 542–548.
Gyllenberg, M., Koski, T., Reilink, E., & Verlaan, M. (1994b). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, *31*(2), 542–548.
Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301–321.
Hagenaars, J. A. (1993). *Loglinear models with latent variables* (Vol. 94). Newbury Park, CA: Sage Publications Inc.
Hartz, S.M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Ph.D. thesis, University of Illinois at Urbana-Champaign.
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.
Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, *41*(3), 281–293.
Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Its Applications*, *18*(2), 95–138.
Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning q-matrix. *Bernoulli*, *19*(5A), 1790–1817.
Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.
Mityagin, B. (2015). The zero set of a real analytic function. arXiv preprint arXiv:1512.07276.
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.
Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *51*(3), 337–350.
Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Final report, Technical report.
Teicher, H., et al. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, *32*(1), 244–248.
Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287.
Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, *45*(2), 675–707.
Xu, G., & Shang, Z. (2017). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*, 1284–1295.
Yakowitz, S. J., & Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, *39*, 209–214.

Manuscript Received: 10 AUG 2018
Final Version Received: 13 DEC 2019
Published Online Date: 11 JAN 2020