Identifying Latent Structures in Restricted Latent Class Models

Gongjun Xu* and Zhuoran Shang†

Abstract

This paper focuses on a family of restricted latent structure models with wide applications in psychological and educational assessment, where the model parameters are restricted via a latent structure matrix to reflect pre-specified assumptions on the latent attributes. Such a latent matrix is often provided by experts and assumed to be correct upon construction, yet it may be subjective and misspecified. Recognizing this problem, researchers have been developing methods to estimate the matrix from data. However, the fundamental issue of the identifiability of the latent structure matrix has not been addressed until now. The first goal of this paper is to establish identifiability conditions that ensure the estimability of the structure matrix. With the theoretical development, the second part of the paper proposes a likelihood-based method to estimate the latent structure from the data. Simulation studies show that the proposed method outperforms the existing approaches. We further illustrate the method through a data set in educational assessment.

Keywords: Restricted latent class models; identifiability; cognitive diagnosis; Q-matrix.

1 Introduction

Restricted latent class models with diagnostic feature. Latent class models are popularly used in social sciences to model latent attributes that are not directly measurable, which assume that observed responses can be explained by a set of discrete latent attributes (Goodman, 1974; Agresti, 2013). This paper focuses on a family of *restricted* latent class models that have diagnostic feature. This class of models have wide applications in psychological and educational measurement, where a classification-based decision is made about an

*Department of Statistics, University of Michigan, 456 West Hall, 1085 South University, Ann Arbor, MI, 48109; Email: gongjun@umich.edu

†School of Statistics, University of Minnesota; Email: shang063@umn.edu

individual's latent attributes from his or her observed responses. In particular, a subject, such as an examinee or a patient, provides binary responses $\mathbf{R} = (R_1, ..., R_J)^\top$ to J diagnostic items, where $\top$ denotes the transpose. These responses are assumed to be explained by K unobserved binary latent attributes $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$. The binary value $\alpha_k \in \{0, 1\}$ indicates the absence or presence of the kth attribute, respectively. The vector $\boldsymbol{\alpha}$ specifies a latent class that is usually called an attribute profile or knowledge state. Such construction of $\boldsymbol{\alpha}$, which is different from the conventional latent class model setting, is assumed for the diagnosis purpose. For instance, teachers may want to know whether students have mastered certain skills; and psychiatrists want to know whether patients have certain mental disorders.

For these diagnostic models, another major difference from the conventional latent class models is that the model parameters are restricted by a binary latent structure matrix, called the Q-matrix. The Q-matrix reflects the pre-specified diagnostic relationships between the J items and the K latent attributes (see Section 2). The Q-restricted latent class models have the desirable diagnostic feature of providing informative cognitive profiles for every respondent, which allows for the design of more effective intervention strategies. These models have recently gained great popularity in educational proficiency assessment (e.g., Junker and Sijtsma, 2001; Hartz and Roussos, 2008; von Davier, 2008; Henson et al., 2009; de la Torre, 2011), psychiatric diagnosis (e.g., Templin and Henson, 2006; Chen et al., 2015), and many other disciplines (e.g., Tatsuoka, 2009; Rupp, Templin, and Henson, 2010). The models also provide the basis for computerized-adaptive diagnosis in online testing and learning (e.g., Wang, Lin, Chang, and Douglas, 2016; Xu, Wang, and Shang, 2016; Zhang and Chang, 2016).

Identifiability Issues and related literature. While the latent Q-matrix plays a key role for diagnosis assessment, identifiability of these restricted latent structure models has long been an issue, as noted in the literature (de la Torre and Douglas, 2004; Maris and Bechger, 2009; Tatsuoka, 2009; DeCarlo, 2011; von Davier, 2014; Xu and Zhang, 2016).

For unrestricted latent class models with binary responses, Gyllenberg, Koski, Reilink, and Verlaan (1994) showed that they are not identifiable in a strict sense. On the other hand, researchers have considered the generic identifiability of such models, which is defined following algebraic geometry terminology and implies that the set of parameters for which the identifiability does not hold has Lebesgue measure zero. Elmore, Hall, and Neeman (2005) and Allman, Matias, and Rhodes (2009) established generic identifiability results for a large set of latent structure models. Related identifiability results on finite mixture models have also been developed in Hall and Zhou (2003), Hall et al. (2005), Allman et al. (2011), Henry et al. (2014) and many others. However, the existing identifiability results for the unrestricted latent class models cannot be applied to the Q-restricted models due to the additional constraints that reduce the parameter space to a measure zero set. To address this issue, Xu (2017) recently proposed a marginal probability technique and established a set of sufficient conditions for the identifiability of these restricted models under the condition that the Q-matrix is correctly specified beforehand and known.

However, the latent Q-matrix, which is often provided by experts upon construction, is subjective and can be misspecified. The misspecification of the Q-matrix could lead to serious lack of fit and consequently inaccurate inferences on the latent attribute profiles. Moreover, in exploratory analysis of newly designed items, a large part or the whole Q-matrix may not be available. Recognizing these issues, researchers have been developing methods to estimate the Q-matrix from the response data (e.g., de la Torre, 2008; Barnes, 2010; DeCarlo, 2012; Liu, Xu, and Ying, 2012, 2013; Chiu, 2013; Chen, Liu, Xu, and Ying, 2015; de la Torre and Chiu, 2016). However, identifiability and related statistical properties of the Q-matrix have largely been an underexplored area in the literature and it is still not clear when the Q-matrix can be consistently estimated. Some special cases have been recently studied in Liu et al. (2013) and Chen et al. (2015); nevertheless, their theoretical techniques depend on some strong model assumptions and cannot be applied for the general cognitive diagnosis models in psychometrics assessment.

Main contributions. The first aim of this paper is to address the fundamental identifiability issue of the Q-matrix. Compared with the problem of identifying model parameters under a pre-specified structure matrix that was studied in Xu (2017), it is more challenging to establish the identifiability of the latent Q-matrix for several reasons. First, the current work focuses on a more complicated problem than that in Xu (2017). For the Q-restricted models, different Q-matrix corresponds to different set of model parameters and diagnostic constraints. The estimation of the Q-matrix therefore depends on the identification of unknown model parameters under each candidate Q-matrix, where the model parameters themselves may not always be identifiable under these candidates. Second, the Q-matrix of interest is a binary matrix; and the discreteness nature of the identifiability problem makes it different from Xu (2017) and the existing tools may not be directly applicable. We therefore develop new theoretical technique to establish the identifiability results. This paper focuses on a general setting that covers most of the popularly used diagnostic models and develops identifiability results for the Q-matrix, which provide not only theoretical justification for many of the existing estimation methods, but also useful information for related experimental designs, whereas in current applications the designs are usually experience-based and identifiability may not be ensured. Moreover, the proof techniques can be used to establish large sample theory of likelihood-based estimators.

The second aim of the paper is to develop a unified approach to estimate the latent Q-matrix under a general model setting. In particular, we consider two important cases in practice: when the whole Q-matrix is largely unknown and when a provisional Q-matrix is provided. Most existing estimation methods focus on specific diagnostic models with strong model assumptions and cannot be directly applied to the general diagnosis assessment, especially in the first case. Due to the discreteness nature of the Q-matrix, direct search of the maximum likelihood estimator is not practically feasible. We propose a computationally efficient likelihood-based method to estimate the latent structure. Asymptotic properties of the proposed estimator are established with the help of the developed identifiability theory.

Simulation results show the proposed method outperforms the existing methods.

The remainder of this paper is organized as follows. Section 2 introduces the class of restricted models of interest with some examples. Section 3 introduces the identifiability result. Section 4 proposes a likelihood-based estimation method and studies its theoretical properties. Sections 5 and 6 present simulation studies and real data analysis. A discussion is given in Section 7. The proofs and additional numerical results are presented in the supplementary Appendix.

2 Q-restricted latent class models

In this section, we first give an introduction of the considered restricted latent class models, followed by examples of several popularly used models. Assume that N subjects are randomly sampled from a target population and their attribute profiles $\boldsymbol{\alpha}_i, i = 1, ..., N$ independently follow a categorical distribution with probabilities $p_{\boldsymbol{\alpha}} := \mathrm{P}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha})$ for any $\boldsymbol{\alpha} \in \{0,1\}^K$, where $p_{\boldsymbol{\alpha}} \in (0,1)$ and $\sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = 1$. Given the ith subject's attribute profile $\boldsymbol{\alpha}_i$, the response R_{ij} to item j follows a Bernoulli distribution with positive response probability $\theta_{j,\boldsymbol{\alpha}_i} := \mathrm{P}(R_{ij} = 1 \mid \boldsymbol{\alpha}_i)$. In addition, the ith subject's responses $\mathbf{R}_i = \{R_{ij}, j = 1, \cdots, J\}$ are assumed conditionally independent given $\boldsymbol{\alpha}_i$. Such conditional independence assumption is commonly used in finite mixture literature, such as Hall and Zhou (2003) and Allman et al. (2009). We write $\Theta = (\theta_{j,\boldsymbol{\alpha}})$ as a $J \times 2^K$ matrix containing the θ parameters and $\mathbf{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)^\top$ as a 2^K dimensional vector. The unknown parameters of the latent class model include Θ and $\mathbf{p}$.

The cognitive diagnosis models (CDMs) are a class of restricted latent class models where the model parameters $\Theta = (\theta_{j,\boldsymbol{\alpha}})$ are constrained by pre-assumed relationships between the J items and the K latent attributes. Such relationships are specified through a $J \times K$ binary matrix, which is called Q-matrix in the literature. The entry $q_{jk} \in \{0,1\}$ of the Q-matrix indicates the absence or presence, respectively, of a link between the jth item and the kth latent attribute. For instance, the following self-explained Q-matrix corresponds to four items, three latent attributes, and $2^3 = 8$ latent classes.

$$Q = \begin{array}{c|ccc} & & \text{attribute} & \\ & \alpha_1 & \alpha_2 & \alpha_3 \\ \hline \text{item 1} & 1 & 0 & 0 \\ \text{item 2} & 0 & 1 & 0 \\ \text{item 3} & 1 & 0 & 1 \\ \text{item 4} & 0 & 1 & 1 \end{array} \tag{1}$$

Denote the jth row vector of Q by $Q_{j,\star}$, which gives the full attribute requirements of the jth item. For an attribute profile $\boldsymbol{\alpha}$, we write $\boldsymbol{\alpha} \succeq Q_{j,\star}$ if $\alpha_k \geq q_{jk}$ for any $k \in \{1, \ldots, K\}$, and $\boldsymbol{\alpha} \not\succeq Q_{j,\star}$ if there exists k such that $\alpha_k < q_{jk}$. We write $0_k = (0, ..., 0)^\top_{k\times 1}$ and $1_k = (1, ..., 1)^\top_{k\times 1}$, and omit the index of length when there is no ambiguity. Furthermore, let $\mathbf{e}_i$ be a standard basis vector, whose ith element is one and the rest are zeros.

The constraints on θ's are motivated as follows. For $\boldsymbol{\alpha} \succeq Q_{j,\star}$, a subject with $\boldsymbol{\alpha}$ has all the attributes for item j specified by the Q-matrix and would be "most capable" to provide a positive response; on the other hand, for $\boldsymbol{\alpha}' \not\succeq Q_{j,\star}$, a subject with $\boldsymbol{\alpha}'$ misses some related attribute and is expected to have a lower positive response probability than $\boldsymbol{\alpha} \succeq Q_{j,\star}$. In addition, a subject without mastery of any latent traits is expected to have the lowest positive response probability. These constraints on Θ are summarized as follows:

$$\max_{\boldsymbol{\alpha}:\, \boldsymbol{\alpha}\succeq Q_{j,\star}} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\, \boldsymbol{\alpha}\succeq Q_{j,\star}} \theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'} \geq \theta_{j,0}, \quad \text{for any } \boldsymbol{\alpha}' \not\succeq Q_{j,\star}. \tag{2}$$

Take item 1 in Equation (1) for an example. Under (2), subjects with $\alpha_1 = 1$ have a higher positive response probability than those with $\alpha_1 = 0$; on the other hand, $\boldsymbol{\alpha} = (1,0,0)^\top, (1,1,0)^\top, (1,0,1)^\top$ and $(1,1,1)^\top$ all have the same correct response probabilities.

The introduced models are important statistical tools developed in cognitive diagnosis to detect the presence or absence of multiple fine-grained skills or attributes. Many restricted latent class models have been proposed in the past decades for various application purposes (e.g., Junker and Sijtsma, 2001; Templin and Henson, 2006; DiBello et al., 1995; Hartz and Roussos, 2008; de la Torre and Douglas, 2004; von Davier, 2008; Henson et al., 2009; de la Torre, 2011). Below we introduce some of them as examples.

Example 1 (DINA model)**.** *The deterministic input noisy output "and" gate model (DINA; Junker and Sijtsma, 2001) assumes a conjunctive relationship among attributes, i.e., it is necessary to possess all the attributes indicated by the Q-matrix to be capable of providing a positive response. For an item j and a subject with $\boldsymbol{\alpha}$, the ideal response $\xi_{j,\boldsymbol{\alpha}}(Q) = I(\boldsymbol{\alpha} \succeq Q_{j,\star})$ indicates the capability of the subject answering the item positively. The uncertainty is further incorporated using two item-level parameters: the slipping parameter $s_j = P\{R_j = 0 \,|\, \xi_{j,\boldsymbol{\alpha}}(Q) = 1\}$ denotes the probability of making a negative response despite mastering all needed skills, and the guessing parameter $g_j = P\{R_j = 1 \,|\, \xi_{j,\boldsymbol{\alpha}}(Q) = 0\}$ denotes the probability of a positive response despite an incorrect ideal response. The response probability $\theta_{j,\boldsymbol{\alpha}}$ then takes the form $\theta_{j,\boldsymbol{\alpha}} = (1-s_j)^{\xi_{j,\boldsymbol{\alpha}}(Q)} g_j^{1-\xi_{j,\boldsymbol{\alpha}}(Q)}$. For the DINA model,* (2) *is satisfied if $1 - s_j > g_j$, which is usually assumed in practice.*

Example 2 (Reduced RUM)**.** *Under the reduced version of the reparameterized unified model (DiBello et al., 1995; Henson et al., 2009), $\theta_{j,\boldsymbol{\alpha}} = \pi_j \prod_{k=1}^{K} {\gamma_{jk}}^{q_{jk}(1-\alpha_k)}$, where π_j is the positive response probability for subjects who possess all required attributes and γ_{jk}, $0 < \gamma_{jk} < 1$, is the penalty parameter for not possessing the kth attribute. For the reduced RUM, assumptions* (2) *is satisfied.*

Example 3 (LCDM)**.** *The Loglinear-CDM (LCDM, Henson et al., 2009) is a restricted latent class model that models the relationships between categorical variables and attribute profiles as $\mathrm{logit}(\theta_{j,\boldsymbol{\alpha}}) = \boldsymbol{\beta}_j^\top h(\boldsymbol{\alpha}, Q_{j,\star})$, where the vector β_j represents a 2^K-dimensional vector of weights for the jth item and $h(\boldsymbol{\alpha}, Q_{j,\star})$ represents a set of linear combinations of the $\boldsymbol{\alpha}$ and $Q_{j,\star}$. In particular, the saturated model corresponds to*

$$\boldsymbol{\beta}_j^\top h(\boldsymbol{\alpha}, \mathbf{q}_j) = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_k + \sum_{k=1}^{K} \sum_{k'>k} \beta_{jkk'} q_{jk} q_{jk'} \alpha_k \alpha_{k'} + \cdots + \beta_{j12\cdots K} \prod_{k=1}^{K} q_{jk} \prod_{k=1}^{K} \alpha_k.$$

Note that for any $1 \leq h \leq K$ and any $1 \leq k_1 < \cdots < k_h \leq K$, if $\prod_{l=1}^{h} q_{j,k_l} = 0$, then $\beta_{j,k_1\cdots k_h}$ is not needed in the model and can be set as 0. The main effect model becomes the linear logistic model (LLM, see Hagenaars, 1993; Maris, 1999; de la Torre and Douglas, 2004) that $\text{logit}\,(\theta_{j,\boldsymbol{\alpha}}) = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_k$.

3 Identifiability results

We present the main identifiability results in this section, before which we first introduce some notations and formulate the definition of the identifiability of the Q-matrix.

The distribution of $\mathbf{R}$, conditional on the latent class $\boldsymbol{\alpha}$, is given by a J-way $2 \times \cdots \times 2$ table $\mathbb{P}_{\boldsymbol{\alpha}}(Q, \Theta) = \bigotimes_{j=1}^{J} (1 - \theta_{j,\boldsymbol{\alpha}}, \theta_{j,\boldsymbol{\alpha}})^\top$, where the $\mathbf{r} = (r_1, \cdots, r_J)^\top$th entry of the table is the probability of observing response vector $\mathbf{r}$ given Q-matrix, Θ, and latent class $\boldsymbol{\alpha}$, i.e., $\mathrm{P}(\mathbf{R} = \mathbf{r} \mid Q, \Theta, \boldsymbol{\alpha}) = \prod_{j=1}^{J} (\theta_{j,\boldsymbol{\alpha}})^{r_j} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-r_j}$. The marginal distribution of $\mathcal{R}$ is then given by $\mathbb{P}(Q, \Theta, \mathbf{p}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \mathbb{P}_{\boldsymbol{\alpha}}(Q, \Theta)\, p_{\boldsymbol{\alpha}}$, where the $\mathbf{r}$th entry is $\mathrm{P}(\mathbf{R} = \mathbf{r} \mid Q, \Theta, \mathbf{p}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \mathrm{P}(\mathbf{R} = \mathbf{r} \mid Q, \Theta, \boldsymbol{\alpha})\, p_{\boldsymbol{\alpha}}$.

The question of interest is when the Q-matrix is estimable from the response data $\mathbf{R}$. It is worthy to mention that the Q-matrix is expected to be identifiable only up to rearranging the orders of the columns. This is because when estimating the Q-matrix, the data do not contain information about the specific meaning of each attribute. For this reason, if Q and $\bar{Q}$ have an identical set of column vectors, we consider them as equivalent and write $Q \sim \bar{Q}$; otherwise, we write $Q \nsim \bar{Q}$. For example,

$Q =$

	α_1	α_2
item 1	1	0
item 2	1	1
item 3	1	1
item 4	1	1

$\sim \quad \bar{Q} =$

	α_2	α_1
item 1	0	1
item 2	1	1
item 3	1	1
item 4	1	1

Definition 1. *For the restricted models satisfying* (2)*, we say that the Q-matrix is identifi-*

able if for any $\bar{Q} \nsim Q$, there does not exist $(\bar{\Theta}, \bar{\mathbf{p}})$ such that $\mathbb{P}(Q, \Theta, \mathbf{p}) = \mathbb{P}(\bar{Q}, \bar{\Theta}, \bar{\mathbf{p}})$.

We next illustrate which type of Q-matrix structure is required for the identifiability results. An important and basic structure that have been studied in the literature is the completeness of the Q-matrix, where we say a Q-matrix is complete if $\{\mathbf{e}_j^\top : j = 1, ..., K\} \subset \{Q_{j,\star} : j = 1, \cdots, J\}$; see, e.g., Chiu et al. (2009). In other words, a Q-matrix is complete if there exist K rows of Q that can be ordered to form the K-dimensional identity matrix $\mathcal{I}_K$. A simple example of a complete Q-matrix is the $K \times K$ identity matrix $\mathcal{I}_K$.

We start with a simple and ideal case. We consider the model introduced in Example 1 and the ideal case where the jth response $R_j = \xi_{j,\boldsymbol{\alpha}}(Q)$, where $\xi_{j,\boldsymbol{\alpha}}(Q) = I(\boldsymbol{\alpha} \succeq Q_{j,\star})$; that is, the capable subjects always provide positive responses and incapable subjects always give negative responses. In this ideal case, $\theta_{j,\boldsymbol{\alpha}} = \xi_{j,\boldsymbol{\alpha}}(Q)$ and $\mathbf{p}$ is unspecified. The completeness of a Q-matrix is sufficient and necessary for the identifiability of $\mathbf{p}$ in the considered ideal case when Q is known (Chiu et al., 2009; Xu and Zhang, 2016). Liu et al. (2013) further showed that for this ideal case, a sufficient condition for the identifiability of the Q-matrix is that the Q-matrix is complete and each attribute is required by at least two items.

Example 4. *Consider Q in Equation (3) as an example. It is not complete and we show it is not identifiable. In particular, for $\bar{Q}$ in (3), where all elements of $\bar{Q}$ are same as Q except $\bar{q}_{31} = 0$, we show Q and $\bar{Q}$ are not distinguishable based on responses generated under Q.*

$Q =$

	α_1	α_2
item 1	1	0
item 2	1	1
item 3	1	1
item 4	1	1

;

$\bar{Q} =$

	α_1	α_2
item 1	1	0
item 2	1	1
item 3	0	1
item 4	1	1

(3)

Consider the ideal case with $\theta_{j,\boldsymbol{\alpha}} = \xi_{j,\boldsymbol{\alpha}}(Q)$ and $\bar{\theta}_{j,\boldsymbol{\alpha}} = \xi_{j,\boldsymbol{\alpha}}(\bar{Q})$. Let the true model parameter associated with Q be $\mathbf{p}$. We now construct a different $\bar{\mathbf{p}}$ by setting $\bar{p}_{(0,1)} = 0$ and $\bar{p}_{(0,0)} = p_{(0,0)} + p_{(0,1)}$ while the other elements same as $\mathbf{p}$. For such $\bar{\mathbf{p}}$ and the $\bar{Q}$ in (3), $P(\mathbf{R} = \mathbf{r} \mid \bar{Q}, \bar{\Theta}, \boldsymbol{\alpha})\bar{p}_{\boldsymbol{\alpha}} = P(\mathbf{R} = \mathbf{r} \mid Q, \Theta, \boldsymbol{\alpha})p_{\boldsymbol{\alpha}}$ for any $\mathbf{r}$ and $\boldsymbol{\alpha} \in \{(1,0)^\top, (1,1)^\top\}$. In addition,

$P\{\mathbf{R}=\mathbf{r} \mid \bar{Q}, \bar{\Theta}, \boldsymbol{\alpha}=(0,0)^\top\} \cdot \bar{p}_{(0,0)} + P\{\mathbf{R}=\mathbf{r} \mid \bar{Q}, \bar{\Theta}, \boldsymbol{\alpha}=(0,1)^\top\} \cdot \bar{p}_{(0,1)} = P\{\mathbf{R}=\mathbf{r} \mid Q, \Theta, \boldsymbol{\alpha}=(0,0)^\top\} \cdot p_{(0,0)} + P\{\mathbf{R}=\mathbf{r} \mid Q, \Theta, \boldsymbol{\alpha}=(0,1)^\top\} \cdot p_{(0,1)}$. *Therefore,* $P(\mathbf{R}=\mathbf{r} \mid \bar{Q}, \bar{\Theta}, \bar{\mathbf{p}}) = P(\mathbf{R}=\mathbf{r} \mid Q, \Theta, \mathbf{p})$ *for any* $\mathbf{r}$*. From Definition 1,* Q *in* (3) *is not identifiable.*

For more general restricted latent class models satisfying constraints (2), we provide in the following a unified sufficient condition that ensures the identifiability of the Q-matrix. Although the above ideal model is a very special case of the considered models, it shows the necessity to require that the true Q-matrix is complete. Moreover, for application purpose, we also need to ensure the identifiability of the model parameters under the true Q-matrix; such identifiability conditions have been studied in Xu (2017). We assume the following identifiability conditions.

C1 The true Q-matrix takes the form of $Q^\top = \{\mathcal{I}_K; \mathcal{I}_K; (Q')^\top\}^\top$ after row swapping, where Q' is a $(J-2K) \times K$ binary matrix.

C2 Given Q arranged as in C1, for any attribute profiles $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ and $\boldsymbol{\alpha} \succeq \boldsymbol{\alpha}'$, $(\theta_{j,\boldsymbol{\alpha}}; j > 2K)^\top \neq (\theta_{j,\boldsymbol{\alpha}'}; j > 2K)^\top$.

Remark 1. *Condition C1 implies that Q is complete and each attribute is required by at least two items. The completeness of the Q-matrix is a necessary condition for the identifiability of the population proportion parameters* $p_{\boldsymbol{\alpha}}$ *under the simple DINA model. For instance, for the Q-matrix in Example 4, it is not complete and we can see subjects with* $\boldsymbol{\alpha}=(0,0)$ *and* $\boldsymbol{\alpha}=(0,1)$ *are not distinguishable from their responses. Without completeness, we can easily construct nonidentifiable Q-matrix as illustrated in Example 4. Condition C1 requires two complete matrices. This follows from the previous study of the DINA model in Example 1 (Liu et al., 2013; Chen et al., 2015). Beyond the literature on cognitive diagnosis, the completeness type structure has been used in confirmatory analysis of multidimensional item response theory, where the attributes are modeled as continuous latent variables (e.g., Reckase, 2009). The developed theoretical results in this paper could also be extended to other latent structure models in social science such as the mixed membership model, where* *it has been shown that the mixed membership model can be equivalently represented as a restricted latent class models with similar completeness requirement (e.g., Erosheva et al., 2007). Condition C2 implies that for attribute profiles $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ and $\boldsymbol{\alpha} \succeq \boldsymbol{\alpha}'$, there exists at least one item in Q' such that subjects with $\boldsymbol{\alpha}$ have different positive response probabilities from subjects with $\boldsymbol{\alpha}'$. Both C1 and C2 hold if there are three identity submatrices in the Q-matrix. From Theorem 1 in Xu (2017), C1 and C2 ensure the identifiability of the model parameters $(\Theta, \mathbf{p})$ under the true Q-matrix while C1 itself cannot ensure that.*

Theorem 1. *Consider the restricted models satisfying* (2). *Under conditions C1 and C2, the Q-matrix is identifiable.*

Theorem 1 specifies conditions under which the Q-matrix is identifiable from the response data. The result is under a general setting satisfying assumption (2) and it covers many existing models as special cases. More importantly, the result allows different items to follow different underlying diagnostic assumptions. In addition, together with Theorem 1 in Xu (2017), we have both Q and the model parameters $(\Theta, \mathbf{p})$ are identifiable under C1 and C2.

Corollary 1. *Consider the restricted models satisfying* (2). *The Q-matrix and model parameters* $(\Theta, \mathbf{p})$ *are identifiable under conditions C1 and C2.*

Remark 2. *The identifiability result would provide a guideline of how to design the diagnostic items and how to calibrate the new designed items from response data. It is recommended to have at least two complete matrices in the test; moreover, each attribute is recommended to be required by at least three items. The identifiability result would also help to improve existing diagnostic tests. For instance, when researchers find that the estimation results are problematic and the Q-matrix does not satisfy the identifiability conditions, it is recommended to design new items such that the identifiability is ensured. Moreover, with a subset of items carefully designed by experts to satisfy the identifiability conditions, we can use the responses to estimate the Q-matrix of new items and to detect possible misspecifications of existing items. We propose a likelihood-based estimation method in Section 4.*

Remark 3. *The identifiability results generalize the existing results in two ways. First, the current work provides a unified identifiability result that is applicable to many diagnostic models. For the identifiability of the Q-matrix, there are few studies in the literature which only focus on some special cases. For instance, Chen et al. (2015) focused on the DINA model and showed that identifying Q under the DINA model requires two copies of* $\mathcal{I}_K$ *and a third item measuring each attribute. The first requirement is the Condition C1 and the second one is related to C2 of this paper. Second, the identifiability results do not require test items to follow the same diagnostic model. For instance, some items can follow the DINA while others can follow the Reduced-RUM or LCDM. More flexible diagnostic tests therefore can be designed following the identifiability results.*

Remark 4. *The generic identifiability results in Allman et al. (2009) can not be directly applied in the current model setting. This is because under the same Q-matrix, there may be several cognitive diagnosis models of interest. For instance, the DINA model can be taken as a submodel of the LCDM under the same Q-matrix. In this case, the parameters under the DINA model lie in a subspace of the parameter space under the LCDM, and generic identifiability results for the more general LCDM may not ensure the identifiability of the DINA model. When the identifiability conditions are not satisfied, such as the Q-matrix is not complete, then we may expect to obtain partial identification results as recently studied in Henry et al. (2014) and identify the Q-matrix up to certain equivalent class. For instance, the incomplete Q-matrix in Example 4 would be in the same identification class as* $\bar{Q}$ *in the example. In analysis with a provisional Q-matrix, such partial identifiability result may lead to "locally identifiability" near the provisional Q-matrix due to the discreteness of the Q-matrix. On the other hand, the problem in this work takes a different setting from existing studies such as Henry et al. (2014), which assumes the existence of an additional variable that provides a source of variation in the mixture weights while leaves component distributions unchanged, and their results cannot be directly applied.*

4 Estimation of the Q-matrix

4.1 Likelihood-based estimation of the Q-matrix

In Sections 4.1, we consider the estimation of the Q-matrix in a full exploratory analysis setting, where no information on the Q-matrix is provided. In Section 4.2, we study the case where a provisional Q-matrix is available. When there is no confusion, in the following, we use $(Q, \Theta, \mathbf{p})$ to denote a general candidate set of the Q-matrix and model parameters, and use $(Q_0, \Theta_0, \mathbf{p}_0)$ to denote the true values.

We consider an information-based approach to estimate the Q-matrix. Note that under the general CDM setting, a Q-matrix may correspond to a set of different submodels of the Q-restricted latent class model. For instance, the DINA model can be considered as a submodel of the LCDM under the same Q-matrix. In order to account for the model complexity, a natural choice is to use the information criterion, and we choose the Q-matrix estimator (up to column permutation) such that it minimizes the following objective function

$$\hat{Q} \sim \underset{Q,\Theta,\mathbf{p}}{\arg\min} -l_N(Q, \Theta, \mathbf{p}; \mathcal{R}) + \lambda \times \#\{\Theta^Q\}, \tag{4}$$

where $l_N(Q, \Theta, \mathbf{p}; \mathcal{R})$ is the marginal log-likelihood of $(Q, \Theta, \mathbf{p})$, $\mathcal{R} = \{\mathbf{R}_i, i = 1, \cdots, N\}$ is the observed response data, $\#\{\Theta^Q\}$ denotes the number of free item parameters in matrix Θ under the Q-introduced constraints, and $\lambda > 0$ is a regularization parameter that indicates the penalty level on the model complexity. For instance, when $\lambda = 1$ this is equivalent to is Akaike's information criterion (AIC) and when $\lambda = \log N/2$, this is similar to the Bayesian information criterion (BIC).

Due to the discreteness nature of the latent structure matrix, direct estimation of maximum likelihood estimator is computational demanding. The key idea of the proposed method is to reformulate the problem of estimating the Q-matrix as a problem of variable selection. For computational convenience, we consider the general LCDM framework in Example 3 where the monotonicity assumption can be easily incorporated. The proposed approaches can be easily applied to other link functions. For any $j \in \{1, \cdots, J\}$, define a 2^K-dimensional parameter vector $\boldsymbol{\beta}_j = \big(\beta_{j,0},\ \beta_{j,k_1\cdots k_h},\ \text{for any } 1 \le h \le K \text{ and any } 1 \le k_1 < \cdots < k_h \le K\big)^\top$. We reparametrize the $\theta_{j,\boldsymbol{\alpha}}$ parameters under a matrix Q by

$$\text{logit}(\theta_{j,\boldsymbol{\alpha}}) = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk}\alpha_k + \sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} \beta_{jkk'}\alpha_k\alpha_{k'} + \cdots + \beta_{j12\cdots K}\prod_{k=1}^{K}\alpha_k, \tag{5}$$

where for any $1 \le h \le K$ and any $1 \le k_1 < \cdots < k_h \le K$, $\beta_{j,k_1\cdots k_h} = 0$ if $\prod_{l=1}^h q_{j,k_l} = 0$. Note that when $\prod_{l=1}^h q_{j,k_l} \neq 0$, $\beta_{j,k_1\cdots k_h}$ may be or not be 0, which depends on the cognitive diagnosis model assumption on the jth item. For instance, for $Q_{j,\star} = 1_K^\top$, $\prod_{l=1}^h q_{j,k_l} \neq 0$ always holds, but under the DINA model, we have $\boldsymbol{\beta}_j = (\beta_{j,0}, 0, \cdots, 0, \beta_{j,1\cdots K})^\top$ while under the saturated LCDM, $\boldsymbol{\beta}_j = (\beta_{j,0}, \beta_{j,1}, \cdots, \beta_{j,1\cdots K})^\top$.

From the above construction, for any item j, the item vector $Q_{j,\star}$ is uniquely determined by the sparsity structure of the vector $\boldsymbol{\beta}_j$. On the other hand, the sparsity structure of $\boldsymbol{\beta}_j$ is not uniquely determined by $Q_{j,\star}$, as illustrated by the example in the last paragraph. As a consequence, the estimation of the Q-matrix in equation (4) is equivalent to the estimation of the sparsity structure of B, i.e.,

$$\hat{B} \sim \underset{B,\mathbf{p}}{\arg\min} -l_N(B, \mathbf{p}; \mathcal{R}) + \lambda \sum_{j=1}^{J} \sum_{\substack{1\le h\le K \\ 1\le k_1<\cdots<k_h\le K}} I(\beta_{j,k_1\cdots k_h} \neq 0), \tag{6}$$

where $B = \{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_J\}$ is a set of candidate model parameters, $l(B, \mathbf{p}; \mathcal{R})$ is the log-likelihood evaluated at $(B, \mathbf{p})$ under the model (5) with $Q = \mathbf{1}_{J\times K}$. Let $\hat{S}$ be the index set of the nonzero β's in $\hat{B}$. Then based on $\hat{S}$, we can uniquely obtain an estimate $\hat{Q}$ (up to column permutation).

Directly solving (6) is still computationally challenging due to the L_0 penalty terms, i.e., $I(\beta_{j,k_1\cdots k_h} \neq 0)$. Motivated by the work of Shen et al. (2012), which studied constrained L_0 likelihood and its computational surrogate, we replace the L_0 function $I(\beta_{j,k_1\cdots k_h} \neq 0)$, by its surrogate $J_\tau(\beta_{j,k_1\cdots k_h}) := \min(|\beta_{j,k_1\cdots k_h}|/\tau, 1)$ to construct an approximation. The $J_\tau(\cdot)$ is a truncated L_1 penalty (TLP) function and the parameter τ decides the size of coefficients to be shrunk toward zero. We then estimate Q by

$$\hat{B} \sim \arg\min_{B,\mathbf{p}} -l_N(B, \mathbf{p}; \mathcal{R}) + \lambda \sum_{j=1}^{J} \sum_{\substack{1\le h\le K \\ 1\le k_1<\cdots<k_h\le K}} J_\tau(\beta_{j,k_1\cdots k_h}). \tag{7}$$

The constrained counterpart problem of (7) can be written as

$$\hat{B} \sim \arg\min_{B,\mathbf{p}} -l_N(B, \mathbf{p}; \mathcal{R}) \text{ subject to } \sum_{j=1}^{J} \sum_{\substack{1\le h\le K \\ 1\le k_1<\cdots<k_h\le K}} J_\tau(\beta_{j,k_1\cdots k_h}) \le M, \tag{8}$$

for some positive constant M.

Let B_0 be the $J \times 2^K$ vector of true model parameters corresponding to Θ_0 under Q_0. Note that when $\tau < \min\{|\beta| \neq 0, \beta \in B_0\}$, the surrogate $J_\tau(\cdot)$ becomes exactly the L_0 penalty, and therefore via tuning τ, we expect the selection method in (7) performs similarly to the information-based selection in (6). Theoretically, thanks to the identifiability result in Section 3, we have the following results on its consistency and asymptotic behaviors.

We need some notations to state the theoretical properties. Let S_0 be the index set of nonzero β's in B_0 and B_{0,S_0} be the vector of these nonzero β's. Denote the cardinality of S_0 by M_0. Let $\hat{B}_0$ be the oracle maximum likelihood estimator provided that the true Q_0 and the specific diagnostic model assumption were known a priori, i.e., the index set S_0 was known, and $\hat{B}_{0,S_0}$ be the estimated $\hat{\beta}$'s indexed by S_0. Similarly, for any candidate B and index set S, we let B_S be the vector of β's indexed by S. We further write $\eta = (B, \mathbf{p})$, $\eta_0 = (B_0, \mathbf{p}_0)$, $\eta_{0,S} = (B_{0,S}, \mathbf{p}_0)$, $\hat{\eta} = (\hat{B}, \hat{\mathbf{p}})$, $\hat{\eta}_0 = (\hat{B}_0, \hat{\mathbf{p}})$, and $\hat{\eta}_{0,S} = (\hat{B}_{0,S}, \hat{\mathbf{p}})$. In addition, we assume the following condition:

- C3 The true parameters B_{0,S_0} are bounded and the Fisher information matrix evaluated at η_{0,S_0}, denoted by $\mathbf{I}_{S_0}$, is nonsingular.

Proposition 1. *Under the conditions in Theorem 1 and condition C3, if $M = M_0$ and*

$\tau < \delta$ for some small constant δ, then for the optimization problem in (8)*, there exist positive constants c_1 and c_2 such that for any N, $P(\hat{B} \nsim \hat{B}_0) \leq \exp\{-c_1 N + c_2\}$ and $P(\hat{Q} \nsim Q_0) \leq \exp\{-c_1 N + c_2\}$. Furthermore, $\sqrt{N}(\hat{\eta}_{S_0} - \eta_{0,S_0})$ and $\sqrt{N}(\hat{\eta}_{0,S_0} - \eta_{0,S_0})$ have the same limiting Gaussian distribution with mean zero and covariance $\mathbf{I}_{S_0}^{-1}$.*

Proposition 1 shows the consistency of the estimated Q matrix and the convergence rate is of exponential order $\exp\{-c_1 N + c_2\}$. It also implies that $P(\hat{S} \nsim S_0) \to 0$ and the estimated model parameters $\hat{\eta}$ achieve the oracle limiting distribution. We also obtain the consistency result for the primary optimization problem in (7).

Proposition 2. *Assume the conditions in Theorem 1 and C3. Further suppose that λ and τ depend on N such that $N^{-1/2}\lambda \to 0$, $N^{1/2}\tau \to \infty$, and $N^{-1/2}\lambda\tau^{-1} \to \infty$. Then for the optimization problem in* (7)*, $P(\hat{S} \nsim S_0) \to 0$ and $P(\hat{Q} \nsim Q_0) \to 0$. Furthermore, $\sqrt{N}(\hat{\eta}_{S_0} - \eta_{0,S_0})$ weakly converges to the Gaussian distribution with mean zero and covariance $\mathbf{I}_{S_0}^{-1}$.*

Remark 5. *Propositions 1 and 2 theoretically justify the proposed estimation procedure and also provide the asymptotic distributions for statistical inference on the model parameters. To compute standard errors of the estimated model parameters, we need a consistent estimator of $\mathbf{I}_{S_0}$, which can be obtained from the restricted latent class model under the estimated Q-matrix. Thanks to Propositions 1 and 2, such $\hat{\mathbf{I}}_{S_0}$ is consistent under conditions C1-C3.*

The selection of λ and τ is crucial to the successful detection of latent structure. Proposition 2 gives an asymptotic guideline to choose λ and τ. Note that the conditions imply that $\lambda \to \infty$, $\tau \to 0$, $N^{-1/2}\lambda \to 0$, $N^{1/2}\tau \to \infty$, and $N^{-1/2}\lambda\tau^{-1} \to \infty$. A sufficient condition is that $\lambda = N^{1/2-\epsilon_1}$ and $\tau = N^{-\epsilon_2}$ for small positive constants $\epsilon_2 > \epsilon_1 > 0$.

For data analysis, we propose to use information criteria such as the BIC to select the tuning parameters. In particular, for each candidate pair of tuning parameters (λ, τ), we obtain the estimated vector $\hat{B}_{(\lambda,\tau)}$, the index set of its nonzero elements $\hat{S}_{(\lambda,\tau)}$, and the implied Q-matrix $\hat{Q}_{(\lambda,\tau)}$. Then we estimate the constrained maximum likelihood estimator of η with the β's indexed by $\hat{S}^c_{(\lambda,\tau)}$, the complement set of $\hat{S}_{(\lambda,\tau)}$, being constrained to be 0.

The maximum likelihood estimator depends on (λ, τ) only via the estimated $\hat{S}_{(\lambda,\tau)}$ and we denote it by $\hat{\eta}^*_{\hat{S}(\lambda,\tau)}$. We further define $IC(\hat{S}_{(\lambda,\tau)}, c_N) = -2l_N(\hat{\eta}^*_{\hat{S}(\lambda,\tau)}; \mathcal{R}) + c_N \times \#\{\hat{\eta}^*_{\hat{S}(\lambda,\tau)}\}$, where c_N is some constant depending on N. When $c_N = \log N$, the IC becomes the BIC. Among a candidate set of (λ, τ)'s, we choose the one that minimizes the IC value and take the corresponding $\hat{Q}$ to be the final estimator of the Q-matrix. The following proposition gives conditions that ensure the selection consistency of this procedure.

Proposition 3. *Assume the conditions in Theorem 1 and C3. Further assume that $c_N \to \infty$, $c_N = o(N)$, and there exists (λ_N, τ_N) in the candidate set of tuning parameters such that the limiting conditions in Proposition 2 are satisfied. Then the probability of the above IC procedure selecting the true Q-matrix converges to 1 as $N \to \infty$.*

Proposition 3 ensures the consistency of the BIC, which is further supported by the simulation studies in Section 5. Alternatively we can use other information criteria satisfying conditions in Proposition 3 to select the final Q-matrix, such as those proposed in Chen and Chen (2008), Zhang and Shen (2010), Fan and Tang (2013) and many others.

Remark 6. *Directly solving the optimization problem in* (7) *could be computationally inefficient due to the latent structure setting. Instead, it is solved via an EM algorithm. We also propose a fast pre-screening method to get reasonable starting points by solving a regularized likelihood of the main effect LCDM model. Please refer to the Appendix for more details.*

4.2 Stepwise estimation with a provisional Q-matrix

In this section we adapt the estimation method in the previous section to the case when there is an initial yet maybe misspecified Q-matrix given by practitioners. The provisional $Q^{(0)}$ is often believed by practitioners to be close to the true Q_0 with only a few possible misspecifications. Although the method in Section 4.1 can be directly applied by using the $Q^{(0)}$ as a starting matrix of the estimation algorithm, in data analysis with limited sample size, this method often tends to find a "global optimal" Q-matrix that has a low information criterion value such as BIC but may differ $Q^{(0)}$ with many items. Such an estimated Q-matrix, though statistically fits the data better, may be difficult to interpret for the purpose of cognitive diagnosis. To incorporate such practical need into the estimation procedure, we adapt the method in Section 4.1 to be a stepwise estimation procedure with each step focusing on updating one item.

The stepwise procedure starts the EM algorithm in Section 4.1 using the provisional $Q^{(0)}$ and the estimated model parameters under $Q^{(0)}$ as initial values. We denote the BIC under the $Q^{(0)}$-restricted general CDM as $BIC^{(0)}$. In the M-step, we estimate $\hat{\boldsymbol{\beta}}_j^{(0)}$, $j = 1, \cdots, J$. Instead of update all $\boldsymbol{\beta}$'s as in the exploratory estimation in Section 4.1, for each item j, we introduce a matrix $Q^{(0,j)}$ that updates $Q^{(0)}$ with only the jth row, denoted by $Q_{j,\star}^{(0,j)}$, according to the estimated $\hat{\boldsymbol{\beta}}_j^{(0)}$. Note that $Q_{j,\star}^{(0,j)}$ is uniquely determined by $\hat{\boldsymbol{\beta}}_j^{(0)}$ and $Q^{(0,j)}$ may be the same as $Q^{(0)}$. Let $BIC^{(0,j)}$ be the BIC under the matrix $Q^{(0,j)}$. If there is an item $j^{(1)}$ such that $j^{(1)} = \arg\min_{j:\ Q^{(0,j)} \neq Q^{(0)},\ BIC^{(0,j)} < BIC^{(0)}} BIC^{(0,j)}$, then we update the Q-matrix as $Q^{(1)} = Q^{(0,j^{(1)})}$. Note that there may exists an item h with $BIC^{(0,h)} < BIC^{(0,j^{(1)})}$ but $Q^{(0,h)} = Q^{(0)}$, that is, for the hth item, there is a submodel having a lower BIC than the general CDM under the same Q-matrix. To account for such submodel effects during estimation, for any item h such that $BIC^{(0,h)} \leq BIC^{(0,j^{(1)})}$, we update the item response model $\theta_{h,\boldsymbol{\alpha}}$ according to the nonzero structure of $\hat{\boldsymbol{\beta}}_h^{(0)}$, while for other items we still use the general CDM. This ends the first step of the stepwise estimation method. We repeat the preceding procedure until the BIC starts to increase. Theoretically, Proposition 3 ensures the estimation procedure to find at least a local optimal Q-matrix.

Remark 7. *When the sample size is not large enough, the stepwise detection procedure may overestimate the number of the misspecified items. In order to control the number of false positive detections, we propose to use a bagging method to reduce the estimation variance. Specifically, we resample N individuals' response with replacement from the original data set and perform the stepwise estimation procedure. We repeat this M times with M a relatively large number and denote the estimated Q-matrices by $Q_m^*, m = 1, \cdots, M$. Then we calculate* *the average estimator* $\bar{Q}^* = (\bar{q}^*_{jk})_{J\times K} := \frac{1}{M}\sum_{m=1}^{M} Q^*_m$ *and the final detected entries are those with* $\bar{q}^*_{jk} > s$ *if the initial* $q^{(0)}_{jk} = 0$ *and* $\bar{q}^*_{jk} < s$ *if* $q^{(0)}_{jk} = 1$*. Here* s *is a threshold value to classify* $\bar{q}^*_{jk}$ *as 0 or 1, and a natural choice is* 0.5.

5 Simulation Results

We illustrate the performance of the proposed estimation procedures via two simulation studies. For the first study in Section 5.1, we assume no prior information on the Q-matrix. For the second study in Section 5.2, a Q-matrix is given yet misspecified with a few items.

We introduce the simulation setting that will be used in both studies. We consider latent attributes with dimension $K = 3, 4$ and 5, and the test length $J = 20$. The true Q-matrices, shown as following, are chosen such that our identifiability conditions are satisfied.

$$Q_3 = \begin{pmatrix} 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&1&0\\ 1&0&1\\ 0&1&1\\ 1&1&0\\ 1&0&1\\ 0&1&1\\ 1&1&0\\ 1&0&1\\ 1&1&1\\ 1&1&1\\ 1&1&1 \end{pmatrix} \quad Q_4 = \begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&1&0&0\\ 0&1&1&0\\ 0&0&1&1\\ 1&0&0&1\\ 1&0&1&0\\ 0&1&0&1\\ 1&1&0&0\\ 0&0&1&1\\ 0&1&1&1\\ 1&0&1&1\\ 1&1&0&1\\ 1&1&1&0 \end{pmatrix} \quad Q_5 = \begin{pmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 1&1&0&0&0\\ 0&1&1&0&0\\ 0&0&1&1&0\\ 0&0&0&1&1\\ 1&0&0&0&1\\ 1&1&1&0&0\\ 0&1&1&1&0\\ 0&0&1&1&1\\ 1&0&0&1&1\\ 1&1&0&0&1 \end{pmatrix}$$

In both studies we use simulated data from two types of latent class models: the DINA and the saturated LCDM. Both are designed such that the correct response probabilities for all items are between 0.2 and 0.8. For the DINA, the slipping and guessing parameters of all items are set to be 0.2. For the LCDM and any item j requiring K_j attributes, we set the correct response probabilities of attribute profiles with K_j' out of the K_j required attributes to be $0.2 + (0.8 - 0.2) \times K_j'/K_j$. Note that the DINA model has $2J$ item parameters and the LCDM has $\sum_{j=1}^{J} 2^{K_j}$ item parameters under the true Q-matrix.

It is natural that one subject's latent attributes are correlated. To consider the dependence, we use the following two steps to simulate the true latent profiles (Chen et al., 2015). First generate $\boldsymbol{x}_i = (x_{i1}, ..., x_{iK}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, for $i = 1, ..., N$ where $\Sigma = (1-\rho)I_K + \rho \mathbf{1}_K \mathbf{1}_K^T$; then the attribute profile α_{ik} is set to be 1 if $x_{ik} \geq 0$ and 0 otherwise. In both studies, three different situations of dependency is considered by choosing $\rho = 0$, 0.15 and 0.25.

After generating latent profiles and item parameters, we simulate the observed responses for 500 independent replications. Even though the data are generated under the DINA and LCDM, the true models are assumed to be *unknown* during the estimation.

5.1 Exploratory estimation of the whole Q-matrix

In this study we estimate the Q-matrix completely from the data. In the case of $K = 3$, the following crossover design is applied for the considered two models, three sample sizes, and three attribute dependent levels: $\{\text{DINA}, \text{LCDM}\} \otimes \{N = 500, 1000, 2000\} \otimes \{\rho = 0, 0.15, 0.25\}$.

Table 3 presents the simulation results. The column "Matrix" shows matrix-level estimation results and gives the proportion of the entire Q-matrix correctly recovered by the estimation method among 500 replications. The column "Item" is the item-level estimation results and it shows the averaged proportion of the item Q-vectors being correctly estimated. For the entry-level results, the column "TPR" is the proportion of true connections between attribute and item being correctly detected, i.e., the 1's in the true Q-matrix correctly estimated; and "FPR" is the proportion of irrelevant item-attribute pairs specified as relevant, i.e., the 0's in the true Q-matrix estimated as 1's. For comparison, we have also performed the estimation method using the Lasso penalty. Multiple starting values are used. Table 3 shows that the proposed truncated L_1 method outperforms the L_1 regularized estimation in most cases. Both methods perform better when sample size increases and attributes are less correlated. The correct recovery rate of the Q-matrix is higher for the DINA model than that for the LCDM. This is because in the DINA model each item has only one non-zero and non-intercept coefficient, which is relatively large and easier to detect.

			DINA				LCDM			
ρ	N		Matrix	Item	TPR	FPR	Matrix	Item	TPR	FPR
0	500	TLP	0.958	0.998	1.000	0.002	0.566	0.972	0.987	0.005
		L1	0.948	0.997	1.000	0.002	0.552	0.972	0.988	0.007
	1000	TLP	0.980	0.999	1.000	0.001	0.938	0.996	1.000	0.002
		L1	0.980	0.999	1.000	0.001	0.926	0.996	0.999	0.003
	2000	TLP	0.990	0.999	1.000	0.000	0.980	0.999	1.000	0.001
		L1	0.990	0.999	1.000	0.000	0.978	0.999	1.000	0.001
0.15	500	TLP	0.920	0.992	0.999	0.005	0.562	0.970	0.990	0.010
		L1	0.920	0.995	1.000	0.004	0.532	0.966	0.988	0.011
	1000	TLP	0.958	0.996	1.000	0.003	0.900	0.995	0.999	0.003
		L1	0.958	0.996	1.000	0.003	0.900	0.995	1.000	0.003
	2000	TLP	0.972	0.997	1.000	0.002	0.974	0.999	1.000	0.001
		L1	0.970	0.997	1.000	0.003	0.972	0.998	1.000	0.001
0.25	500	TLP	0.910	0.990	0.998	0.006	0.516	0.966	0.988	0.011
		L1	0.886	0.992	1.000	0.006	0.432	0.959	0.986	0.013
	1000	TLP	0.958	0.998	1.000	0.002	0.866	0.993	0.999	0.005
		L1	0.930	0.996	1.000	0.003	0.826	0.990	0.999	0.006
	2000	TLP	0.964	0.996	0.999	0.003	0.980	0.999	1.000	0.001
		L1	0.958	0.995	1.000	0.004	0.974	0.999	1.000	0.001

Table 1: Exploratory estimation results for $K = 3$. The column "Matrix" is the proportion of the entire Q-matrix correctly recovered. "Item" is the proportion of the item vectors correctly estimated. TPR is the proportion of the 1's in the true Q-matrix correctly detected. FPR is the proportion of the 0's in the true Q-matrix falsely estimated as 1's.

We also consider the cases with the number of latent attributes $K = 4$ and $K = 5$. We use the non-correlated attributes and two sample size $N = 1000$ and 2000. Table 2 shows the simulation results for 500 replications. Due to the fact that the size of parameters in the saturated model increases with K exponentially, the estimation becomes more difficult, particularly for the LCDM. However, the item-level ("Item") and entry-level ("TPR" and "FPR") estimation results are quite accurate with more than 98% of the item Q-vectors and almost all entries correctly estimated when $N = 2000$. Overall, the TLP outperforms the Lasso method. As in the case of $K = 3$, the DINA model has better estimation results than the LCDM due to the sparser and stronger signals. We also note that under the DINA model, the TLP estimation results for $K = 5$ are slightly better than $K = 4$; this might be due to the Monte Carlo error and the selection of tuning parameters during the estimation.

Estimation results of the parameters $\Theta = (\theta_{j,\boldsymbol{\alpha}})_{J\times 2^K}$ are presented in Table 3 with $K \in \{3, 4, 5\}$ and $\rho = 0$. The correlated cases with $\rho \neq 0$ are similar and therefore not reported here. Recall that $\theta_{j,\boldsymbol{\alpha}}$ denotes the correct response probability to the jth item for latent class $\boldsymbol{\alpha}$. Two methods are compared. For the proposed method, the $\hat{\theta}$'s are calculated from the refitted $\hat{\beta}$ values under the estimated model structure (column "TLP"). For the true model, the θ's are estimated under the true Q-matrix and the true diagnostic model assumption (column "True"). We report the averaged absolute values of the estimated biases of $\hat{\theta}$'s (column "aBias") and the averaged squared-root mean squared error (column "RMSE"). Table 3 shows that the proposed method gives similar estimation results to those under the true model, which is consistent with the theoretical results in Propositions 1 and 2. Please also refer to Figure A.2 in the Appendix for the box plots of the MSEs.

			DINA				LCDM			
K	N		Matrix	Item	TPR	FPR	Matrix	Item	TPR	FPR
$K = 4$	1000	TLP	0.956	0.998	1.000	0.001	0.600	0.973	0.997	0.010
		L1	0.950	0.997	1.000	0.001	0.566	0.969	0.998	0.012
	2000	TLP	0.960	0.998	1.000	0.001	0.890	0.994	1.000	0.003
		L1	0.956	0.998	1.000	0.001	0.884	0.994	1.000	0.003
$K = 5$	1000	TLP	0.970	0.998	1.000	0.001	0.342	0.944	0.991	0.013
		L1	0.916	0.995	1.000	0.002	0.136	0.905	0.998	0.029
	2000	TLP	0.974	0.998	1.000	0.001	0.712	0.982	1.000	0.005
		L1	0.954	0.997	1.000	0.001	0.662	0.980	1.000	0.006

Table 2: Exploratory estimation results for $K = 4$ and 5.

	N	DINA				LCDM			
		TLP		True		TLP		True	
		aBias	RMSE	aBias	RMSE	aBias	RMSE	aBias	RMSE
$K = 3$	500	0.004	0.029	0.004	0.028	0.007	0.044	0.007	0.045
	1000	0.003	0.020	0.003	0.020	0.005	0.029	0.005	0.032
	2000	0.002	0.014	0.002	0.014	0.004	0.019	0.004	0.022
$K = 4$	1000	0.004	0.023	0.004	0.022	0.008	0.039	0.007	0.039
	2000	0.003	0.016	0.003	0.016	0.006	0.024	0.005	0.027
$K = 5$	1000	0.004	0.026	0.004	0.026	0.010	0.048	0.008	0.047
	2000	0.003	0.018	0.003	0.017	0.006	0.027	0.005	0.032

Table 3: Estimation results for Θ. "aBias" is averaged absolute values of estimated biases of θ's and "RMSE" is averaged squared-root mean squared error. "TLP" is the re-fitted estimate under the estimated model structure; "True" is the estimate under the true model structure.

5.2 Stepwise estimation with a provisional Q-matrix

In this simulation study, we aim to estimate the Q-matrix when a provisional Q-matrix is available. The provisional Q-matrix is designed to be misspecified at two levels: 10% and 20% on the item level. The misspecified J_0 items are selected randomly from the $J = 20$ items and the Q-vector of a misspecified item is uniformly selected from the 2^K possible vectors expect the true one and the zero vector.

We first consider $K = 3$ and use a crossover design of two models, two degrees of misspecification levels, three sample sizes, and three attribute dependent levels: $\{\text{DINA}, \text{LCDM}\} \otimes \{\text{Misspecification } 10\%, 20\%\} \otimes \{N = 500, 1000, 2000\} \otimes \{\rho = 0, 0.15, 0.25\}$. For each case, we compare the performance of proposed method with the GMDI method (de la Torre and Chiu, 2016). The simulation results are summarized in Table 4 for low (10%) and Table A.1 (presented in the Appendix) for high (20%) misspecification levels. The column "Total" shows the proportion of correctly estimated item vectors for each method; note that by the design of our simulation, the baseline "Total" value of the initial $Q^{(0)}$-matrix is 0.9 for low-misspecified case and 0.8 for high-misspecified case. The column "TPR" (true positive rate) shows the proportion of those misspecified entries/vectors in the initial $Q^{(0)}$ that are correctly detected, and "FPR" (false positive rate) is the proportion of those correctly-specified entries/vectors in the provisional $Q^{(0)}$ that are being falsely detected. The results show that the proposed method outperforms the GMDI method for all simulation conditions. The TPR of proposed method tends to 1 as sample size increases while such trend is not significant for the GMDI. Moreover, the performance of our method declines only slightly as the misspecification level increases from 10% to 20%, while the GMDI approach is affected more significantly. With the same sample size, the stepwise procedure works better for the simpler DINA model while the performance is similar for different dependence levels.

We use the proposed stopping rule based on the BIC. We also report the results using a fixed step number J_0, which is the number of misspecified items in the initial $Q^{(0)}$, in the brackets. It shows that the proposed sequential method performs similarly to that with fixed J_0 steps, which indicates that our method detects most of the misspecified items within the first J_0 steps. We also consider the cases with $K = 4$ and $K = 5$, and the results are presented in the Appendix.

6 Data analysis

We consider a data set that has been used for educational assessment. This dataset contains responses from 536 middle school students to a set of fraction subtraction items. Various subsets of the items with different numbers of attributes have been analyzed in the literature, such as Tatsuoka (2002), de la Torre and Douglas (2004), de la Torre (2011), Chen et al. (2015), and de la Torre and Chiu (2016). We follow the setting in Chen et al. (2015) and study 17 items. The item contents and the Q-matrix with 8 attributes given by de la Torre and Douglas (2004) are presented in the left of Table 5. The attributes are defined as follows: (A1) Convert a whole number to a fraction; (A2) Separate a whole number from a fraction; (A3) Simplify before subtracting; (A4) Find a common denominator; (A5) Borrow from whole number part; (A6) Column borrow to subtract the 2nd numerator from the 1st; (A7) Subtract numerators; and (A8) Reduce answer to simplest form.

			DINA					LCDM				
				Entry		Vector			Entry		Vector	
ρ	N		Total	TPR	FPR	TPR	FPR	Total	TPR	FPR	TPR	FPR
0	500	Proposed	0.998	0.986	0.000	0.980	0.001	0.963	0.860	0.009	0.850	0.024
			(0.996)	(0.981)	(0.001)	(0.975)	(0.002)	(0.973)	(0.884)	(0.005)	(0.850)	(0.013)
		GMDI	0.923	0.252	0.000	0.230	0.000	0.923	0.252	0.000	0.235	0.000
	1000	Proposed	0.999	0.996	0.000	0.985	0.000	0.993	0.950	0.000	0.935	0.001
			(0.999)	(0.996)	(0.000)	(0.985)	(0.000)	(0.994)	(0.968)	(0.001)	(0.950)	(0.002)
		GMDI	0.932	0.345	0.000	0.325	0.000	0.932	0.340	0.000	0.320	0.000
	2000	Proposed	0.999	1.000	0.000	0.995	0.001	0.998	0.993	0.000	0.985	0.001
			(1.000)	(1.000)	(0.000)	(0.995)	(0.000)	(0.997)	(0.986)	(0.000)	(0.970)	(0.001)
		GMDI	0.924	0.290	0.000	0.245	0.000	0.924	0.281	0.000	0.240	0.000
0.15	500	Proposed	0.998	0.991	0.000	0.985	0.001	0.961	0.877	0.009	0.840	0.026
			(0.998)	(0.991)	(0.000)	(0.985)	(0.001)	(0.972)	(0.874)	(0.005)	(0.835)	(0.013)
		GMDI	0.924	0.247	0.000	0.240	0.000	0.923	0.248	0.000	0.240	0.001
	1000	Proposed	0.999	0.991	0.000	0.985	0.000	0.994	0.962	0.001	0.955	0.002
			(0.998)	(0.991)	(0.000)	(0.985)	(0.001)	(0.995)	(0.973)	(0.001)	(0.965)	(0.002)
		GMDI	0.932	0.345	0.000	0.325	0.000	0.931	0.326	0.000	0.315	0.000
	2000	Proposed	1.000	1.000	0.000	1.000	0.000	0.999	0.993	0.000	0.990	0.000
			(1.000)	(1.000)	(0.000)	(0.995)	(0.000)	(0.999)	(0.991)	(0.000)	(0.985)	(0.000)
		GMDI	0.924	0.284	0.000	0.245	0.000	0.924	0.279	0.000	0.240	0.000
0.25	500	Proposed	0.998	0.991	0.000	0.985	0.001	0.967	0.902	0.008	0.875	0.023
			(0.997)	(0.986)	(0.001)	(0.980)	(0.002)	(0.975)	(0.888)	(0.004)	(0.855)	(0.012)
		GMDI	0.922	0.234	0.000	0.225	0.000	0.920	0.235	0.001	0.225	0.003
	1000	Proposed	0.999	0.995	0.000	0.990	0.000	0.994	0.963	0.001	0.955	0.002
			(0.999)	(0.995)	(0.000)	(0.990)	(0.000)	(0.994)	(0.971)	(0.001)	(0.960)	(0.003)
		GMDI	0.932	0.333	0.000	0.325	0.000	0.931	0.326	0.000	0.315	0.000
	2000	Proposed	1.000	1.000	0.000	1.000	0.000	1.000	0.995	0.000	0.995	0.000
			(1.000)	(1.000)	(0.000)	(0.995)	(0.000)	(0.999)	(0.991)	(0.000)	(0.985)	(0.000)
		GMDI	0.924	0.272	0.000	0.245	0.000	0.924	0.275	0.000	0.245	0.000

Table 4: Low misspecification with $K = 3$. "Total" is the proportion of correctly estimated items with the initial baseline 0.9. "TPR" is true positive rate and "FPR" is the false positive rate. For the proposed method, results after the first 2 steps are presented in brackets.

We first apply the proposed stepwise estimation method in Section 4.2. Note that attribute A7 is required by all the items and for the reason of identifiability, we focus on the other 7 attributes. The estimation result suggests to update the highlighted entries in Table 5. In particular, it suggests that the attributes A2 and A3 should be required by item 10 and A4 required by item 11 while A2 not needed for item 11. Such changes appear difficult to interpret under the definitions of the latent attributes. This may be due to the false discoveries of the sequential estimation method with only 536 observations for 8 attributes and the nonidentifiablity issue with the Q-matrix. To better control the false detection,

		Pre-specified								Bootstrap Aggregation						
Item	Content	A1	A2	A3	A4	A5	A6	A7	A8	A1	A2	A3	A4	A5	A6	A8
1	$\frac{5}{3}-\frac{3}{4}$	0	0	0	1	0	1	1	0	0.00	0.00	0.00	1.00	0.00	1.00	0.00
2	$\frac{3}{4}-\frac{3}{8}$	0	0	0	1	0	0	1	0	0.00	0.03	0.00	1.00	0.00	0.00	0.00
3	$\frac{5}{6}-\frac{1}{9}$	0	0	0	1	0	0	1	0	0.01	0.00	0.00	1.00	0.00	0.00	0.00
4	$3\frac{1}{2}-2\frac{2}{3}$	0	1	1	0	1	0	1	0	0.00	0.97	1.00	0.01	0.98	0.00	0.01
5	$1\frac{1}{8}-\frac{1}{8}$	0	0	0	0	0	0	1	1	0.00	0.03	0.00	0.00	0.02	0.00	1.00
6	$3\frac{4}{5}-3\frac{2}{5}$	0	1	0	0	0	0	1	0	0.00	1.00	0.00	0.00	0.00	0.00	0.00
7	$4\frac{5}{7}-1\frac{4}{7}$	0	1	0	0	0	0	1	0	0.00	1.00	0.00	0.00	0.01	0.01	0.00
8	$4\frac{3}{5}-3\frac{4}{10}$	0	1	0	1	0	0	1	1	0.02	0.99	0.00	1.00	0.00	0.01	0.98
9	$3-2\frac{1}{5}$	1	1	0	0	0	0	1	0	1.00	0.95	0.06	0.00	0.06	0.00	0.00
10	$2-\frac{1}{3}$	1	**0**	**0**	0	0	0	1	0	1.00	**0.03**	**0.03**	0.00	0.00	0.00	0.00
11	$4\frac{4}{12}-2\frac{7}{12}$	0	**1**	0	**0**	1	0	1	1	0.00	**0.92**	0.00	**0.13**	1.00	0.00	0.94
12	$4\frac{1}{3}-2\frac{4}{3}$	0	1	0	0	1	0	1	0	0.00	0.99	0.01	0.00	1.00	0.00	0.00
13	$7\frac{3}{5}-\frac{4}{5}$	0	1	0	0	1	0	1	0	0.01	1.00	0.00	0.00	1.00	0.00	0.00
14	$4\frac{1}{10}-2\frac{8}{10}$	0	1	0	0	1	1	1	0	0.00	1.00	0.00	0.00	1.00	1.00	0.00
15	$4-1\frac{4}{3}$	1	1	1	0	1	0	1	0	0.98	1.00	0.89	0.00	0.91	0.00	0.02
16	$4\frac{1}{3}-1\frac{5}{3}$	0	1	1	0	1	0	1	0	0.00	1.00	1.00	0.00	1.00	0.00	0.00
17	$3\frac{3}{8}-2\frac{5}{6}$	0	1	0	1	1	0	1	0	0.00	1.00	0.00	1.00	1.00	0.00	0.00

Table 5: The left is the Q-matrix in de la Torre and Douglas (2004). The highlighted entries are detected from the stepwise method. The right is the bootstrap aggregating result.

we conduct the proposed bootstrap bagging method. The aggregated estimation, shown in Table 5, suggests none of the four detected entries should be changed. The result confirms the validity of the original Q-matrix in de la Torre and Douglas (2004).

We further consider a simpler Q-matrix proposed in Chen et al. (2015) with $K = 3$. The Q-matrix is demonstrated on the left of Table 6 with the three attributes interpreted as: Attribute 1 finding common denominator; Attribute 2 writing integer as fraction; and Attribute 3 subtraction of fraction numbers with integers involved. We perform the proposed sequential approach and it suggests to update item 9 and 10. Both corrections are further confirmed by the bootstrap bagging results in the middle of Table 6. If we check the item content, solving $3-2\frac{1}{5}$ and $2-\frac{1}{3}$ does involve the process of writing integers as fractions, hence should require Attribute 2. Therefore, our results are more consistent with the definition of the attributes and the two detected entries are recommended to be updated.

Exploratory estimation of the latent structure is also conducted using the proposed ap-

		Pre-specified			Bootstrap Aggregation			Exploratory Estimate		
Item	Content	Attr1	Attr2	Attr3	Attr1	Attr2	Attr3	Attr1	Attr2	Attr3
1	$\frac{5}{3}-\frac{3}{4}$	1	0	0	1.00	0.00	0.00	1	0	0
2	$\frac{3}{4}-\frac{3}{8}$	1	0	0	1.00	0.03	0.08	1	0	0
3	$\frac{5}{6}-\frac{1}{9}$	1	0	0	1.00	0.00	0.05	1	0	0
4	$3\frac{1}{2}-2\frac{2}{3}$	0	1	0	0.03	1.00	0.00	0	1	0
5	$1\frac{1}{8}-\frac{1}{8}$	0	0	1	0.04	0.05	1.00	0	0	1
6	$3\frac{4}{5}-3\frac{2}{5}$	0	0	1	0.06	0.05	1.00	0	0	1
7	$4\frac{5}{7}-1\frac{4}{7}$	0	0	1	0.10	0.07	1.00	0	0	1
8	$4\frac{3}{5}-3\frac{4}{10}$	1	0	1	1.00	0.05	0.99	1	0	1
9	$3-2\frac{1}{5}$	1	**0***	1	0.99	**0.51***	0.95	1	1	0
10	$2-\frac{1}{3}$	1	**0***	1	1.00	**0.55***	1.00	1	1	1
11	$4\frac{4}{12}-2\frac{7}{12}$	0	1	1	0.18	1.00	0.97	1	1	0
12	$4\frac{1}{3}-2\frac{4}{3}$	0	1	1	0.00	1.00	0.99	0	1	0
13	$7\frac{3}{5}-\frac{4}{5}$	0	1	1	0.00	1.00	1.00	0	1	0
14	$4\frac{1}{10}-2\frac{8}{10}$	0	1	1	0.03	1.00	1.00	0	1	1
15	$4-1\frac{4}{3}$	0	1	1	0.36	1.00	1.00	1	1	1
16	$4\frac{1}{3}-1\frac{5}{3}$	0	1	1	0.00	1.00	0.99	0	1	0
17	$3\frac{3}{8}-2\frac{5}{6}$	1	1	1	1.00	1.00	1.00	1	1	1

Table 6: Fraction subtraction data with $K = 3$. The left is the Q-matrix from Chen et al. (2015). Entries in blue are detections from the stepwise estimation. The middle is the bagging result. Entries with bootstrap significance are labeled with "$*$". The right is the exploratory estimation result.

proach for $K = 3$. The result is shown on the right of Table 6 and it agrees with Chen et al. (2015) on the first 8 items. Note that Chen et al. (2015) assumes the specific DINA model when conduct the estimation but our approach does not make such model assumption. Consequentially the interpretation of the three attributes should be different from theirs. To further compare the three Q-matrices in Table 6, we calculate their BIC values: the BIC of the Q-matrix in Chen et al. (2015) is 7846, the BIC of the sequential updated one is 7837, and the exploratory one is 7793, which shows that both proposed methods give better goodness of fit than the initial Q-matrix. We also perform the exploratory analysis with other K values. In particular, the estimated Q-matrix with $K = 5$ is given in Table 7. Compared with the Q-matrix in Table 5, the estimated first attribute can be interpreted as "(A1) Convert a whole number to a fraction", the second as "(A4) Find common denominator", the third one as "(A5) Borrow from whole number part", while the last two attributes shall be interpreted differently from theirs due to the dimension reduction. The BIC of the estimated model is 7485, which gives a better fit than the Q-matrices in Tables 5 and 6. Nevertheless, it shall be noted that the estimation only serves as a data-driven guide of constructing the finial Q-matrix, and researchers need to further validate the estimates based on their understanding of the diagnostic items.

Item	Content	Attr1	Attr2	Attr3	Attt4	Attr5
1	$\frac{5}{3} - \frac{3}{4}$	0	1	0	0	0
2	$\frac{3}{4} - \frac{3}{8}$	0	1	0	0	0
3	$\frac{5}{6} - \frac{1}{9}$	0	1	0	0	0
4	$3\frac{1}{2} - 2\frac{2}{3}$	0	0	1	0	0
5	$1\frac{1}{8} - \frac{1}{8}$	0	0	0	1	1
6	$3\frac{4}{5} - 3\frac{2}{5}$	0	0	0	1	1
7	$4\frac{5}{7} - 1\frac{4}{7}$	0	0	0	1	1
8	$4\frac{3}{5} - 3\frac{4}{10}$	0	1	0	1	1
9	$3 - 2\frac{1}{5}$	1	0	1	0	0
10	$2 - \frac{1}{3}$	1	0	1	0	1
11	$4\frac{4}{12} - 2\frac{7}{12}$	0	0	1	1	1
12	$4\frac{1}{3} - 2\frac{4}{3}$	0	0	1	0	1
13	$7\frac{3}{5} - \frac{4}{5}$	0	0	1	0	1
14	$4\frac{1}{10} - 2\frac{8}{10}$	0	0	1	0	1
15	$4 - 1\frac{4}{3}$	1	0	1	0	1
16	$4\frac{1}{3} - 1\frac{5}{3}$	0	0	1	0	1
17	$3\frac{3}{8} - 2\frac{5}{6}$	0	1	1	0	1

Table 7: Exploratory analysis results of Fraction Subtraction Data with $K = 5$.

7 Discussion

This paper aims to identify and estimate the Q-matrix in a family of restricted latent class models. Based on the identifiability results, we develop a likelihood-based estimation method that can be applied to two different cases in practice: estimation of the whole Q-matrix in exploratory analysis and misspecification detection for a provisional Q-matrix. The simulation studies show that our method is able to recover the true latent structure with high accuracy. The real data study demonstrates that our method can construct interpretable latent structure and provide reasonable updates to the existing Q-matrix.

The capability of adapting with or without the prior information of the Q-matrix is one advantage of the proposed approach. The exploratory analysis in Section 4.1 provides researcher with useful information on the tests and helps them to explore the features of new items. The stepwise estimation method in Section 4.2 would serve as a reliable tool to detect the possible misspecifications of a provisional Q-matrix. It should be noted that the Q-matrix that statistically fit the data best may not agree with the one having best practical interpretation. It is always recommended that researchers and test designers further validate the estimation results.

One future research direction is to estimate the number of latent attributes in exploratory analysis. In this study, the latent dimension K is assumed to be known. It is of interest to select the latent dimension according to the model fit and model complexity. Another future work is to establish the partial identification of the Q-matrix when the identifiability conditions are not satisfied. This would be of practical importance, especially when single attribute items are difficult to design and therefore the completeness condition may not be satisfied. Moreover, we assume binary responses in this study while in practice there could be various types of responses data; for instance, identifiablity of multinomial response was recently studied in Fang et al. (2017) using the result in Kruskal (1977). Lastly, the Q-matrix based cognitive diagnosis models provide the basis for cognitive diagnosis computerized adaptive testing (CD-CAT); the proposed method can be extended to the CD-CAT setting to calibrate new designed items and estimate their Q-vectors and item parameters.

Acknowledgement

The authors thank the editor, the associate editor, and three reviewers for many helpful and constructive comments. This work is partially supported by NSF grant SES1659328 and IES grant R305D160010.

References

Agresti, A. (2013), *Categorical data analysis*, Wiley, 3rd ed.

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

— (2011), "Parameter identifiability in a class of random graph mixture models," *Journal of Statistical Planning and Inference*, 141, 1719–1736.

Barnes, T. (2010), "Novel derivation and application of skill matrices: The Q-matrix method," in *Handbook on Educational Data Mining*, Boca Raton: CRC Press, pp. 159–172.

Chen, J. and Chen, Z. (2008), "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, 95, 759–771.

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

Chiu, C.-Y. (2013), "Statistical refinement of the Q-matrix in cognitive diagnosis," *Applied Psychological Measurement*, 37, 598–618.

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

de la Torre, J. (2008), "An empirically-based method of Q-matrix validation for the DINA model: Development and applications," *Journal of Educational Measurement*, 45, 343–362.

— (2011), "The generalized DINA model framework," *Psychometrika*, 76, 179–199.

de la Torre, J. and Chiu, C.-Y. (2016), "A General Method of Empirical Q-matrix Validation," *Psychometrika*, 81, 253–273.

de la Torre, J. and Douglas, J. A. (2004), "Higher order latent trait models for cognitive diagnosis," *Psychometrika*, 69, 333–353.

DeCarlo, L. T. (2011), "On the analysis of fraction subtraction data: the DINA model, classification, class sizes, and the Q-matrix," *Applied Psychological Measurement*, 35, 8–26.

— (2012), "Recognizing Uncertainty in the Q-Matrix via a Bayesian Extension of the DINA Model," *Applied Psychological Measurement*, 36, 447–468.

DiBello, L. V., Stout, W. F., and Roussos, L. A. (1995), "Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques," in *Cognitively diagnostic assessment*, eds. Nichols, P. D., Chipman, S. F., and Brennan, R. L., Hillsdale, NJ: Erlbaum Associates, pp. 361–390.

Elmore, R., Hall, P., and Neeman, A. (2005), "An application of classical invariant theory to identifiability in nonparametric mixtures," in *Annales de l'institut Fourier*, vol. 55, pp. 1–28.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007), "Describing disability through individual-level mixture models for multivariate binary data," *The Annals of Applied Statistics*, 1, 502–537.

Fan, Y. and Tang, C. Y. (2013), "Tuning parameter selection in high dimensional penalized likelihood," *Journal of the Royal Statistical Society: Series B*, 75, 531–552.

Fang, G., Liu, J., and Ying, Z. (2017), "On the Identifiability Diagnostic Classification Models," arXiv Preprint.

Goodman, L. A. (1974), "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, 61, 215–231.

Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994), "Non-uniqueness in probabilistic numerical identification of bacteria," *Journal of Applied Probability*, 542–548.

Hagenaars, J. A. (1993), *Loglinear Models with Latent Variables*, vol. 94, Sage.

Hall, P., Neeman, A., Pakyari, R., and Elmore, R. (2005), "Nonparametric inference in multivariate mixtures," *Biometrika*, 92, 667–678.

Hall, P. and Zhou, X.-H. (2003), "Nonparametric estimation of component distributions in a multivariate mixture," *Annals of Statistics*, 31, 201–224.

Hartz, S. and Roussos, L. (2008), "The fusion model for skills diagnosis: Blending theory with practicality," in *Educational Testing Service, Research Report, RR-08-71*, Princeton, NJ: Educational Testing Service.

Henry, M., Kitamura, Y., and Salanié, B. (2014), "Partial identification of finite mixtures in econometric models," *Quantitative Economics*, 5, 123–144.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009), "Defining a family of cognitive diagnosis models using log-linear models with latent variables," *Psychometrika*, 74, 191–210.

Junker, B. W. and Sijtsma, K. (2001), "Cognitive assessment models with few assumptions, and connections with nonparametric item response theory," *Applied Psychological Measurement*, 25, 258–272.

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, 95–138.

Liu, J., Xu, G., and Ying, Z. (2012), "Data-Driven Learning of Q-Matrix," *Applied Psychological Measurement*, 36, 548–564.

— (2013), "Theory of Self-Learning Q-Matrix," *Bernoulli*, 19(5A), 1790–1817.

Maris, E. (1999), "Estimating multiple classification latent class models," *Psychometrika*, 64, 187–212.

Maris, G. and Bechger, T. M. (2009), "Equivalent diagnostic classification models," *Measurement*, 7, 41–46.

Reckase, M. (2009), *Multidimensional item response theory*, vol. 150, Springer.

Rupp, A. A., Templin, J. L., and Henson, R. A. (2010), *Diagnostic Measurement: Theory, Methods, and Applications*, New York: Guilford Press.

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

Tatsuoka, C. (2002), "Data-analytic methods for latent partially ordered classification models," *Applied Statistics (JRSS-C)*, 51, 337–350.

— (2009), "Diagnostic models as partially ordered sets," *Measurement*, 7, 49–53.

Tatsuoka, K. K. (2009), *Cognitive Assessment: An Introduction to the Rule Space Method*, New York: Routledge.

Templin, J. L. and Henson, R. A. (2006), "Measurement of psychological disorders using cognitive diagnosis models," *Psychological Methods*, 11, 287–305.

von Davier, M. (2008), "A general diagnostic model applied to language testing data," *British Journal of Mathematical and Statistical Psychology*, 61, 287–307.

— (2014), "The DINA model as a constrained general diagnostic model: Two variants of a model equivalency," *British Journal of Mathematical and Statistical Psychology*, 67, 49–71.

Wang, S., Lin, H., Chang, H.-H., and Douglas, J. (2016), "Hybrid Computerized Adaptive Testing: From Group Sequential Design to Fully Sequential Design," *Journal of Educational Measurement*, 53, 45–62.

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

Xu, G., Wang, C., and Shang, Z. (2016), "On initial item selection in cognitive diagnostic computerized adaptive testing," *British Journal of Mathematical and Statistical Psychology*, 69, 291–315.

Xu, G. and Zhang, S. (2016), "Identifiability of diagnostic classification models," *Psychometrika*, 81, 625–649.

Zhang, S. and Chang, H.-H. (2016), "From smart testing to smart learning: how testing technology can assist the new generation of education," *International Journal of Smart Technology and Learning*, 1, 67–92.

Zhang, Y. and Shen, X. (2010), "Model selection procedure for high-dimensional data," *Statistical analysis and data mining*, 3, 350–358.