Journal of Machine Learning Research 20 (2019) 1-58 Submitted 3/19; Revised 6/19; Published 6/19

Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models

Yuqi Gu YUQIGU@UMICH.EDU
Gongjun Xu GONGJUN@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Animashree Anandkumar

Abstract

Structured latent attribute models (SLAMs) are a special family of discrete latent variable models widely used in social and biological sciences. This paper considers the problem of learning significant attribute patterns from a SLAM with potentially high-dimensional configurations of the latent attributes. We address the theoretical identifiability issue, propose a penalized likelihood method for the selection of the attribute patterns, and further establish the selection consistency in such an overfitted SLAM with a diverging number of latent patterns. The good performance of the proposed methodology is illustrated by simulation studies and two real datasets in educational assessments.

1. Introduction

Structured Latent Attribute Models (SLAMs) are widely used statistical and machine learning tools in modern social and biological sciences. SLAMs offer a framework to achieve fine-grained inference on individuals' latent attributes based on their observed multivariate responses, and also to obtain the latent subgroups of a population based on the inferred attribute patterns. In practice, each latent attribute is often assumed to be discrete and has particular scientific interpretation, such as mastery or deficiency of some targeted skill in educational assessments (Junker and Sijtsma, 2001; de la Torre, 2011), presence or absence of some underlying mental disorder in psychiatric diagnosis (Templin and Henson, 2006; de la Torre et al., 2018), and the existence or nonexistence of some disease pathogen in subjects' biological samples (Wu et al., 2017). In these scenarios, the framework of SLAMs enables one to simultaneously achieve the machine learning task of clustering, and the scientific purpose of diagnostic inference.

Different from the exploratory nature of traditional latent variable models, SLAMs often have some additional scientific information for model fitting. In particular, the observed variables are assumed to have certain structured dependence on the unobserved latent attributes, where the dependence is introduced through a binary design matrix to respect the scientific context. The rich structure and nice interpretability of SLAMs make them popular in many scientific disciplines, such as cognitive diagnosis in educational assessment (Junker and Sijtsma, 2001; von Davier, 2008; Henson et al., 2009; Rupp et al., 2010; de la Torre, 2011), psychological and psychiatric measurement for diagnosis of mental disorders

(Templin and Henson, 2006; de la Torre et al., 2018), and epidemiological and medical studies for scientifically constrained clustering (Wu et al., 2017, 2018).

One challenge in modern applications of SLAMs is that the number of discrete latent attributes could be large, leading to a high-dimensional space for all the possible configurations of the attributes, i.e., a high-dimensional space for latent attribute patterns. In many applications, the number of potential patterns is much larger than the sample size. For scientific interpretability and practical use, it is often assumed that not all the possible attribute patterns exist in the population. Examples with a large number of potential latent patterns and a moderate sample size can be found in educational assessments (Lee et al., 2011; Choi et al., 2015; Yamaguchi and Okada, 2018) and the epidemiological diagnosis of disease etiology (Wu et al., 2017). For instance, Example 1 in Section 2 presents a dataset from Trends in International Mathematics and Science Study (TIMSS), which has 13 binary latent attributes (i.e., $2^{13} = 8192$ possible latent attribute patterns) while only 757 students' responses are observed. In cognitive diagnosis, it is of interest to select the significant attribute patterns among these $2^{13} = 8192$ ones. In such high-dimensional scenarios, existing estimation methods often tend to over select the number of latent patterns, and may not scale to datasets with a huge number of patterns. Moreover, theoretical questions remain open on whether and when the "sparse" latent attribute patterns are identifiable and can be consistently learned from data.

Identifiability of SLAMs has long been an issue in the literature (e.g., von Davier, 2008; DeCarlo, 2011; Maris and Bechger, 2009; von Davier, 2014; Xu and Zhang, 2016). SLAMs can be viewed as a special family of restricted latent class models and their identifiability has a close connection with the study of tensor decompositions, by noting that the probability distribution of a SLAM can be viewed as a mixture of specially structured tensor products. In the literature, it is known that unrestricted latent class models are not identifiable (Gyllenberg et al., 1994). Nonetheless, Carreira-Perpinán and Renals (2000) showed through extensive simulations that they are almost always identifiable, which the authors termed as practical identifiability. Allman et al. (2009) further established *generic* identifiability of various latent variable models, including latent class models. Generic identifiability is weaker than strict identifiability, and it implies that the model parameters are almost surely identifiable with respect to the Lebesgue measure of the parameter space. The study of Allman et al. (2009) is based on an identifiability result of the three-way tensor decomposition in Kruskal (1977). Other analysis of tensor decompositions has also been conducted to study the identifiability of various latent variable models (e.g., Drton et al., 2007; Hsu and Kakade, 2013; Anandkumar et al., 2014; Bhaskara et al., 2014; Anandkumar et al., 2015; Jaffe et al., 2018). However, the structural constraints imposed by the design matrix make these results not directly applicable to SLAMs.

With the aid of the structural constraints, strict identifiability of SLAMs has been obtained under certain conditions on the design matrix (Xu, 2017; Xu and Shang, 2018; Gu and Xu, 2019a,b). However, these works either make the strong assumption that all the possible configurations of the attributes exist in the population with positive probabilities (Xu, 2017; Xu and Shang, 2018), or assume these significant attribute patterns are known *a priori* (Gu and Xu, 2019b). These assumptions are difficult to meet in practice for SLAMs with high-dimensional attributes patterns, and the fundamental learnability issue of the sparse patterns in SLAMs remains unaddressed.

In terms of estimation, learning sparse attribute patterns from a high-dimensional space is related to learning the significant mixture components in a highly overfitted mixture model. Researchers have shown that the estimation of the mixing distributions in overfitted mixture models is technically challenging and it usually leads to nonstandard convergence rate (e.g., Chen, 1995; Ho and Nguyen, 2016; Heinrich and Kahn, 2018). Estimating the number of components in the mixture model goes beyond only estimating the parameters of a mixture, by learning at least the order of the mixing distribution (Heinrich and Kahn, 2018). This problem was also studied in Rousseau and Mengersen (2011) from a Bayesian perspective; however, the Bayesian estimator in Rousseau and Mengersen (2011) may not guarantee the frequentist selection consistency, as to be shown in Section 3. In the setting of SLAMs with the structural constraints and a large number (larger than sample size) of potential latent attribute patterns, it is not clear how to consistently select the significant patterns.

Our contributions in this paper contain the following aspects. First, we characterize the identifiability requirement needed for a SLAM with an arbitrary subset of attribute patterns to be learnable, and establish mild identifiability conditions. Our new identifiability conditions significantly extends the results of previous works (Xu, 2017; Xu and Shang, 2018) to more general and practical settings. Second, we propose a statistically consistent method to perform attribute pattern selection. In particular, we establish theoretical guarantee for selection consistency in the setting of high dimensional latent patterns, where both the sample size and the number of latent patterns can go to infinity. Our analysis also shows that imposing the popular Dirichlet prior on the population proportions would fail to select the true model consistently, when the convergence rate of the SLAM is slower than the usual root-N rate. As for computation, we develop two approximation algorithms to maximize the penalized likelihood for pattern selection. In addition, we propose a fast screening strategy for SLAMs as a preprocessing step that can scale to a huge number of potential patterns, and establish its sure screening property.

The rest of the paper is organized as follows. Section 2 introduces the general setup of structured latent attribute models and motivates our study. Section 3 investigates the learnability requirement and proposes mild sufficient conditions for learnability. Section 4 proposes the estimation methodology and establishes theoretical guarantee for the proposed methods. Section 5 and Section 6 include simulations and real data analysis, respectively. The proofs of all the theoretical results and additional experimental results are included in the Appendix.

2. Model Setup and Motivation

In this section, we first describe the model setup of SLAMs and present several examples. Then we describe the motivation for our study and introduce the problem of interest.

2.1. Structured Latent Attribute Models and Examples

We first introduce the general setup of SLAMs. Consider a SLAM with J designed items which depend on the K latent attributes of interest. There are two types of subject-specific variables in the model, the observed responses to items $\mathbf{R} = (R_1, \ldots, R_J)$ and the latent attribute pattern $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, both assumed to be binary vectors in this work. The J-

dimensional vector $\mathbf{R} \in \{0,1\}^J$ denotes the observed binary responses to the set of J items. The K-dimensional vector $\boldsymbol{\alpha} \in \{0,1\}^K$ denotes a profile of existence or non-existence of the K attributes.

A key structure that specifies how the observed responses depend on the latent attributes is called the Q-matrix, which is a $J \times K$ matrix with binary entries. We denote $Q = (q_{j,k})$ and $q_{j,k} \in \{1,0\}$ reflects whether or not the response to item j has statistical dependence on attribute k. In the context of an educational assessment, $q_{j,k} = 1$ implies the jth test item requires the mastery of the kth skill attribute to answer correctly. We denote the jth row vector of Q by $\boldsymbol{q}_j$, then the K-dimensional binary vector $\boldsymbol{q}_j$ reflects the full attribute requirements of item j. For an attribute pattern $\boldsymbol{\alpha}$, we say $\boldsymbol{\alpha}$ possesses all the required attributes of item j, if $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$, where $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$ denotes $\alpha_k \geq q_{j,k}$ for all $k = 1, \ldots, K$. Example 1 below gives an example of the Q-matrix.

Example 1 *Trends in International Mathematics and Science Study (TIMSS) is a large scale cross-country educational assessment. TIMSS evaluates the mathematics and science abilities of fourth and eighth graders every four years since 1995. Researchers have used SLAMs to analyze the TIMSS data (e.g., Lee et al., 2011; Choi et al., 2015; Yamaguchi and Okada, 2018). For example, a* 23×13 *Q-matrix constructed by mathematics educators was specified for the TIMSS 2003 eighth grade mathematics assessment (Choi et al., 2015). Thirteen attributes (*$K = 13$*) are identified, which fall in five big categories of skill domains measured by the exam, Number, Algebra, Geometry, Measurement, and Data. Table 1 shows the first and last three rows of the Q-matrix, i.e.,* $\{\boldsymbol{q}_j : j = 1, 2, 3, 21, 22, 23\}$.

Item	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}
1	1	0	0	0	0	0	0	0	0	0	1	0	1
2	0	0	0	0	0	1	0	0	0	0	0	0	0
3	0	1	0	0	0	0	1	0	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	0	0	0	0	1	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	1	0	0	0	0	1	0	0	0	0

Table 1: Q-matrix, TIMSS 2003 8th Grade Data

The Q-matrix constrains the model parameters in a certain way to reflect the scientific assumptions. We next introduce the model parameters and how the Q-matrix impose constraints on them in general. Conditional on a subject's latent attribute pattern $\boldsymbol{\alpha} \in \{0,1\}^K$, his/her responses to the J items are assumed to be independent Bernoulli random variables with parameters $\theta_{1,\boldsymbol{\alpha}}, \ldots, \theta_{J,\boldsymbol{\alpha}}$. Specifically, $\theta_{j,\boldsymbol{\alpha}} = \mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha})$ denotes the positive response probability, and is also called an item parameter of item j. We collect all the item parameters in the matrix $\boldsymbol{\Theta} = (\theta_{j,\boldsymbol{\alpha}})$, which has size $J \times 2^K$ with rows indexed by the J items and columns by the 2^K attribute patterns. For pattern $\boldsymbol{\alpha} \in \{0,1\}^K$, we denote its corresponding column vector in $\boldsymbol{\Theta}$ by $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}}$.

One key assumption in SLAMs is that for a latent attribute pattern $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and item j, the parameter $\theta_{j,\boldsymbol{\alpha}}$ is only determined by whether $\boldsymbol{\alpha}$ possesses the attributes

in the set $\mathcal{K}_j = \{k \in \{1, \ldots, K\} : q_{j,k} = 1\}$; that is, those attributes related to item j as specified in the Q-matrix. We will sometimes call the attributes in $\mathcal{K}_j$ the *required attributes of item* j. Under this assumption, all latent attribute patterns in the set

$$\mathcal{C}_j = \{\boldsymbol{\alpha} \in \{0,1\}^K : \boldsymbol{\alpha} \succeq \boldsymbol{q}_j\} \tag{1}$$

share the same value of $\theta_{j,\boldsymbol{\alpha}}$; namely,

$$\max_{\boldsymbol{\alpha} \in \mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha} \in \mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}} \text{ for any } j \in \{1, \ldots, J\}. \tag{2}$$

We will call the set $\mathcal{C}_j$ a *constraint set*. Thus, the Q-matrix puts constraints on $\boldsymbol{\Theta}$ by forcing certain entries of it to be the same. Different SLAMs model the dependence of $\theta_{j,\boldsymbol{\alpha}}$ on the required attributes in $\mathcal{K}_j$ differently to encode different scientific assumptions; please see Examples 2 and 3.

In addition to (2), another key assumption in SLAMs is the monotonicity assumption that

$$\theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'} \text{ for any } \boldsymbol{\alpha} \in \mathcal{C}_j,\ \boldsymbol{\alpha}' \notin \mathcal{C}_j. \tag{3}$$

Constraint (3) is commonly used in our motivating applications of cognitive diagnosis in educational assessments, where (3) indicates subjects mastering all required attributes of an item are more "capable" of giving a positive response to it (i.e., with a larger Bernoulli parameter $\theta_{j,\boldsymbol{\alpha}}$), than those who lack some required attributes. Nonetheless, our theoretical results of model learnability in Section 3 also applies if (3) is relaxed to

$$\theta_{j,\boldsymbol{\alpha}} \neq \theta_{j,\boldsymbol{\alpha}'} \text{ for any } \boldsymbol{\alpha} \in \mathcal{C}_j,\ \boldsymbol{\alpha}' \notin \mathcal{C}_j.$$

This allows more flexibility in the model assumptions of SLAMs used in other applications.

Next we introduce some popular SLAMs in educational and psychological applications. These models are also called Cognitive Diagnosis Models in the psychometrics literature. The first type of SLAMs have exactly two item parameters associated with each item.

Example 2 (two-parameter SLAM) *The two-parameter SLAM specifies exactly two item parameters for each item* j*, which we denote by* θ_j^+ *and* θ_j^-*, with* $\theta_j^+ > \theta_j^-$*. The popular Deterministic Input Noisy output "And" gate (DINA) model introduced in Junker and Sijtsma (2001) is a two-parameter SLAM. It specifies the general form of* $\theta_{j,\boldsymbol{\alpha}}$ *can be rewritten as*

$$\theta_{j,\boldsymbol{\alpha}}^{\textit{two-param.}} = \begin{cases} \theta_j^+, & \textit{if } \boldsymbol{\alpha} \in \mathcal{C}_j, \\ \theta_j^-, & \textit{if } \boldsymbol{\alpha} \notin \mathcal{C}_j. \end{cases}$$

In the application of the two-parameter SLAM in educational assessments, the item parameters θ_j^+ *and* θ_j^- *have the following interpretations. The* $1 - \theta_j^+$ *is called the slipping parameter, denoting the probability of a "capable" subject slips the correct answer, despite mastering all the required attributes of the test item* j*; and* θ_j^- *is called the guessing parameter, denoting the probability of a "non-capable" subject coincidentally giving the correct answer by guessing, despite lacking some required attributes of item* j*. In this case, the unique item parameters in matrix* $\boldsymbol{\Theta}$ *reduce to* $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$*, where* $\boldsymbol{\theta}^+ = (\theta_1^+, \ldots, \theta_J^+)^\top$ *and* $\boldsymbol{\theta}^- = (\theta_1^-, \ldots, \theta_J^-)^\top$*. Under the two-parameter SLAM, the constraint set of each item* j *takes the form of* (1) *and satisfies* (2) *and* (3)*.*

Another family of SLAMs are the multi-parameter models, which allow each item to have multiple levels of item parameters.

Example 3 (multi-parameter SLAMs) *Multi-parameter SLAMs can be categorized into two general types, the main-effect models and the all-effect models. The main-effect models assume the main effects of the required attributes in $\mathcal{K}_j$ play a role in distinguishing the item parameters, which can be written as*

$$\theta_{j,\boldsymbol{\alpha}}^{\textit{main-eff}} = f\Big(\beta_{j,0} + \sum\nolimits_{k\in\mathcal{K}_j}\beta_{j,k}\alpha_k\Big), \tag{4}$$

where $f(\cdot)$ is a link function. Different link functions $f(\cdot)$ lead to different models, including the popular reduced Reparameterized Unified Model (reduced-RUM; DiBello et al., 1995) with $f(\cdot)$ being the exponential function, the Linear Logistic Model (LLM; Maris, 1999) with $f(\cdot)$ being the sigmoid function, and the Additive Cognitive Diagnosis Model (ACDM; de la Torre, 2011) with $f(\cdot)$ the identity function.

Another type of multi-parameter SLAMs are the all-effect models. The item parameter of an all-effect model can be written as

$$\theta_{j,\boldsymbol{\alpha}}^{\textit{all-eff}} = f\Big(\sum\nolimits_{S\subseteq\mathcal{K}_j}\beta_{j,S}\prod\nolimits_{k\in S}\alpha_k\Big). \tag{5}$$

When $f(\cdot)$ is the identity function, (5) *gives the Generalized DINA (GDINA) model proposed by de la Torre (2011); and when $f(\cdot)$ is the sigmoid function,* (5) *gives the Log-linear Cognitive Diagnosis Models (LCDMs) proposed by Henson et al. (2009); see also the General Diagnostic Models (GDMs) proposed in von Davier (2008).*

Under the multi-parameter SLAMs, the constraint set of each item j also takes the form of (1). Those attribute patterns in $\mathcal{C}_j$ still share the same value of item parameters by the definition; and what is different from the two-parameter counterpart is that those $\boldsymbol{\alpha}$ not in $\mathcal{C}_j$ can have different levels of item parameters. We next give another example of multi-parameter SLAMs.

Example 4 (Deep Boltzmann Machines) *The Restricted Boltzmann Machine (RBM) (Smolensky, 1986; Goodfellow et al., 2016) is a popular neural network model. RBM is an undirected probabilistic graphical model, with one layer of latent (hidden) binary variables, one layer of observed (visible) binary variables, and a bipartite graph structure between the two layers. We denote variables in the observed layer by $\mathbf{R}$ and variables in the latent layer by $\boldsymbol{\alpha}$, with lengths J and K, respectively. Under an RBM, the probability mass function of $\mathbf{R}$ and $\boldsymbol{\alpha}$ is $\mathbb{P}(\mathbf{R},\boldsymbol{\alpha}) \propto \exp(-\mathbf{R}^\top \boldsymbol{W}^Q\boldsymbol{\alpha} - \boldsymbol{f}^\top\mathbf{R} - \boldsymbol{b}^\top\boldsymbol{\alpha})$, where $\boldsymbol{f}$, $\boldsymbol{b}$, and $\boldsymbol{W}^Q = (w_{j,k})$ are the parameters. The binary Q-matrix then specifies the sparsity structure in $\boldsymbol{W}^Q$, by constraining $w_{j,k} \neq 0$ only if $q_{j,k} \neq 0$. The Deep Boltzmann Machine (DBM) is a generalization of RBM by allowing multiple latent layers. Consider a DBM with two latent layers $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ of length K_1 and K_2, respectively. The probability mass function of $(\mathbf{R},\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)})$ in this DBM can be written as*

$$\mathbb{P}(\mathbf{R},\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)}) \propto \exp\Big(-\mathbf{R}^\top\boldsymbol{W}^Q\boldsymbol{\alpha}^{(1)} - (\boldsymbol{\alpha}^{(1)})^\top\boldsymbol{U}\boldsymbol{\alpha}^{(2)} - \boldsymbol{f}^\top\mathbf{R} - \boldsymbol{b}_1^\top\boldsymbol{\alpha}^{(1)} - \boldsymbol{b}_2^\top\boldsymbol{\alpha}^{(2)}\Big), \tag{6}$$

where $\boldsymbol{f} \in \mathbb{R}^J$, $\boldsymbol{b}_i \in \mathbb{R}^{K_i}$ *for* $i = 1, 2$, *and* $\boldsymbol{W}^Q = (w_{j,k}) \in \mathbb{R}^{J \times K_1}$, $\boldsymbol{U} \in \mathbb{R}^{K_1 \times K_2}$ *are model parameters; Figure 1 gives an example of a DBM with a* 5×4 *Q-matrix. For* $\boldsymbol{f} = (f_1, \ldots, f_J)^\top$ *and* $\boldsymbol{\alpha}^{(1)} = (\alpha_1^{(1)}, \ldots, \alpha_{K_1}^{(1)})$, *the conditional distribution of an observed variable* R_j *given the latent variables is*

$$\mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \cdots) = \mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha}^{(1)}) = \frac{\exp\left(\sum_{k=1}^{K_1} w_{j,k}\alpha_k^{(1)} + f_j\right)}{1 + \exp\left(\sum_{k=1}^{K_1} w_{j,k}\alpha_k^{(1)} + f_j\right)}, \tag{7}$$

where "$\cdots$*" represents deeper latent layers that potentially exist in a DBM. Moreover, from* (6) *we have* $\mathbb{P}(\mathbf{R} \mid \boldsymbol{\alpha}^{(1)}) = \prod_{j=1}^{J} \mathbb{P}(R_j \mid \boldsymbol{\alpha}^{(1)})$, *so a DBM satisfies the local independence assumption that the* R_j*'s are conditionally independent given the* $\boldsymbol{\alpha}^{(1)}$. *Therefore, a DBM can be viewed as a multi-parameter main-effect SLAM in* (4) *with a sigmoid link function. Viewing a DBM in this way,* (7) *gives the item parameter* $\theta_{j,\boldsymbol{\alpha}^{(1)}}$, *and the constraint set of each item* j *also takes the form* $\mathcal{C}_j = \{\boldsymbol{\alpha}^{(1)} \in \{0,1\}^{K_1} : \boldsymbol{\alpha}^{(1)} \succeq \boldsymbol{q}_j\}$.

$$Q = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix};$$

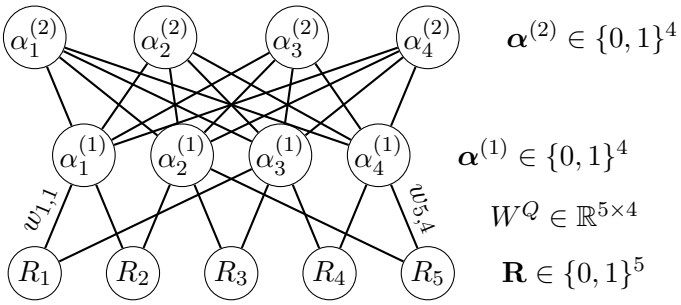


Figure 1: Deep Boltzmann Machine

2.2. Motivation and Problem

One challenge in modern applications of SLAMs is that the number of potential latent attribute patterns 2^K increases exponentially with K and could be much larger than the sample size N. It is often assumed that a relatively small portion of attribute patterns exist in the population. For instance, Example 1 has $2^K = 2^{13} = 8192$ different configurations of attribute patterns; for the limited sample size 757 there, it is desirable to learn the potentially small set of significant attribute patterns from data.

Another motivation for assuming a small number of attribute patterns exist in the population results from the possible hierarchical structure among the targeted attributes. For instance, in an educational assessment of a set of underlying latent skill attributes, some attributes often serve as prerequisites for some others (Leighton et al., 2004; Templin and Bradshaw, 2014). Specifically, the prerequisite relationship depicts the different level of difficulty of the skill attributes, and also reveals the order in which these skills are learned in the population of students. For instance, if attribute α_1 is a prerequisite for attribute α_2, then the attribute pattern $(\alpha_1 = 0, \alpha_2 = 1)$ does not exist in the population, naturally resulting in a sparsity structure of the existence of attribute patterns. When the number of attributes is large and the underlying hierarchy structure is complex and unknown, it is

desirable to learn the hierarchy of attributes directly from data. In such cases with attribute hierarchy, the number of patterns respecting the hierarchy could be far fewer than 2^K.

The problem of interest is that, given a moderate sample size, how to consistently estimate the small set of latent attribute patterns among all the possible 2^K ones. As discussed in the introduction, in the high-dimensional case when the total number of attribute patterns is large or even larger than the sample size, the questions of when the true model with the significant latent patterns are learnable from data, and how to perform consistent pattern selection, remain open in the literature.

This problem is equivalent to selecting the nonzero elements of the population proportion parameters $\boldsymbol{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)$, where $p_{\boldsymbol{\alpha}}$ denotes the proportion of the subjects with latent pattern $\boldsymbol{\alpha}$ in the population. The $\boldsymbol{p}$ satisfies $p_{\boldsymbol{\alpha}} \in [0,1]$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$ and $\sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} = 1$. In this work, we will treat the latent attribute patterns $\boldsymbol{\alpha}$ as random variables (random effects). For any subject, his/her attribute pattern is a random vector $\mathbf{A} \in \{0,1\}^K$ that (marginally) follows a categorical distribution with population proportion parameters $\boldsymbol{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)$. One main reason for this random effect assumption is that, when the number of observed variables per subject (i.e., J) does not increase with the sample size N asymptotically, the counterpart fixed effect model can not consistently estimate the model parameters. As a consequence, the fixed effect approach can not give consistent selection of significant attribute patterns. This scenario with relatively small J but larger N and 2^K is commonly seen in the motivating applications in educational and psychological assessments.

We would like to point out that we give the joint distribution of the attributes full flexibility by modeling it as a categorical distribution with 2^K-1 free proportion parameters $p_{\boldsymbol{\alpha}}$'s. Modeling in this way allows those "sparse" significant attribute patterns to have arbitrary structures among the 2^K possibilities. On the contrary, any simpler parametric model of the distribution of $\boldsymbol{\alpha}$ with fewer parameters would fail to capture all the possibilities of the attributes' dependency.

Under the introduced notations, the probability mass function of a subject's response vector $\mathbf{R} = (R_1, \ldots, R_J)^\top$ can be written as $\mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^J \theta_{j,\boldsymbol{\alpha}}^{r_j}(1-\theta_{j,\boldsymbol{\alpha}})^{1-r_j}$, for $\boldsymbol{r} \in \{0,1\}^J$. Alternatively, the responses can be viewed as a J-th order tensor and the probability mass function of $\mathbf{R}$ can be written as a probability tensor

$$\mathbb{P}(\mathbf{R} \mid \boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{l=1}^{2^K} p_{\boldsymbol{\alpha}_l} \begin{pmatrix} \theta_{1,\boldsymbol{\alpha}_l} \\ 1-\theta_{1,\boldsymbol{\alpha}_l} \end{pmatrix} \circ \begin{pmatrix} \theta_{2,\boldsymbol{\alpha}_l} \\ 1-\theta_{2,\boldsymbol{\alpha}_l} \end{pmatrix} \circ \cdots \circ \begin{pmatrix} \theta_{J,\boldsymbol{\alpha}_l} \\ 1-\theta_{J,\boldsymbol{\alpha}_l} \end{pmatrix}, \tag{8}$$

where "$\circ$" denotes the tensor outer product and $\theta_{j,\boldsymbol{\alpha}}$'s are constrained by (2) and (3).

In the following sections, we first investigate the learnability requirement of learning a SLAM with an arbitrary set of true latent patterns, and provide identifiability conditions in Section 3. Then in Section 4, we propose a penalized likelihood method to select the latent attribute patterns, and establish theoretical guarantee for the proposed method.

3. Learnability Requirement and Conditions

To facilitate the discussion on identifiability of SLAMs, we need to introduce a new notation, the Γ-matrix. We first introduce the $J \times 2^K$ constraint matrix Γ^{all} that is entirely determined

by the Q-matrix. The rows of Γ^{all} are indexed by the J items, and columns by the 2^K latent attribute patterns in $\{0,1\}^K$. The $(j,\boldsymbol{\alpha})$th entry of $\Gamma^{\text{all}}_{j,\boldsymbol{\alpha}}$ is defined as

$$\Gamma^{\text{all}}_{j,\boldsymbol{\alpha}} = I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j) = I(\boldsymbol{\alpha} \in \mathcal{C}_j), \quad j \in \{1,\ldots,J\},\ \boldsymbol{\alpha} \in \{0,1\}^K, \tag{9}$$

which is a binary indicator of whether attribute pattern $\boldsymbol{\alpha}$ possess all the required attributes of item j. We will also call Γ^{all} the *constraint matrix*, since its entries indicate what latent patterns are constrained to have the highest level of Bernoulli parameters for each item. For example, consider the 2×2 Q-matrix in the following (10). Then its corresponding Γ-matrix Γ^{all} with a saturated set of attribute patterns takes the following form.

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \implies \Gamma^{\text{all}} = \begin{array}{c} \begin{array}{cccc} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3 & \boldsymbol{\alpha}_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{c} \\ \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \end{array}. \tag{10}$$

More generally, we generalize the definition of the constraint matrix Γ^{all} in (9) to an arbitrary subset of latent patterns $\mathcal{A} \subseteq \{0,1\}^K$, and an arbitrary set of items $S \subseteq [J]$. For $S \subseteq [J]$ and $\mathcal{A} \subseteq \{0,1\}^K$, we simply denote by $\Gamma^{(S,\mathcal{A})}$ the $|S| \times |\mathcal{A}|$ submatrix of Γ^{all} with row indices from S and column indices from $\mathcal{A}$. When $S = \{1,\ldots,J\}$, we will sometimes just denote $\Gamma^{(S,\mathcal{A})}$ by $\Gamma^{\mathcal{A}}$ for simplicity. Then $\Gamma^{\mathcal{A}}$ itself can be viewed as the constraint matrix for a SLAM with attribute pattern space $\mathcal{A}$, and $\Gamma^{\mathcal{A}}$ directly characterizes how the items constrain the positive response probabilities of latent attribute patterns in $\mathcal{A}$.

Given the Q-matrix, we denote by $\mathcal{A}_0 \subseteq \{0,1\}^K$ the set of true attribute patterns existing in the population, i.e., $\mathcal{A}_0 = \{\boldsymbol{\alpha} \in \{0,1\}^K : p_{\boldsymbol{\alpha}} > 0\}$. In knowledge space theory (Düntsch and Gediga, 1995), the set $\mathcal{A}_0$ of patterns corresponds to the *knowledge structure* of the population. We further denote by $\boldsymbol{\Theta}^{\mathcal{A}_0}$ the item parameter matrix respecting the constraints imposed by $\Gamma^{\mathcal{A}_0}$; specifically, $\boldsymbol{\Theta}^{\mathcal{A}_0} = (\theta_{j,\boldsymbol{\alpha}})$ has the same size as $\Gamma^{\mathcal{A}_0}$, with rows and columns indexed by the J items and the attribute patterns in $\mathcal{A}_0$, respectively. For any positive integer $k \leq 2^K$, we let $\mathcal{T}^{k-1}$ be the k-dimensional simplex, i.e., $\mathcal{T}^{k-1} = \{(x_1, x_2, \ldots, x_k) : x_i \geq 0,\ \sum_{i=1}^k x_k = 1\}$. We denote the true proportion parameters by $\boldsymbol{p}^{\mathcal{A}_0} = (p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}_0) \in \mathcal{T}^{|\mathcal{A}_0|-1}$, then $\boldsymbol{p}^{\mathcal{A}_0} \succ \mathbf{0}$ by the definition of $\mathcal{A}_0$.

The following toy example illustrates why we need to establish identifiability guarantee for pattern selection.

Example 5 *Consider the 2×2 Q-matrix together with its corresponding 2×4 Γ-matrix in Equation (10). Consider two attribute pattern sets, the true set $\mathcal{A}_0 = \{\boldsymbol{\alpha}_1 = (0,0), \boldsymbol{\alpha}_2 = (0,1)\}$ and an alternative set $\mathcal{A}_1 = \{\boldsymbol{\alpha}_2 = (0,1), \boldsymbol{\alpha}_3 = (1,0)\}$. Under the two-parameter SLAM, for any valid item parameters $\boldsymbol{\Theta}$ restricted by Γ and any proportion parameters $\boldsymbol{p} = (p_{\boldsymbol{\alpha}_1}, p_{\boldsymbol{\alpha}_2}, p_{\boldsymbol{\alpha}_3}, p_{\boldsymbol{\alpha}_4})$ such that $p_{\boldsymbol{\alpha}_1} = p_{\boldsymbol{\alpha}_3}$, we have $\mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}^{\mathcal{A}_0}, (p_{\boldsymbol{\alpha}_1}, p_{\boldsymbol{\alpha}_2})) = \mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}^{\mathcal{A}_1}, (p_{\boldsymbol{\alpha}_3}, p_{\boldsymbol{\alpha}_2}))$. This is because $\Gamma^{\mathcal{A}_0} = \Gamma^{\mathcal{A}_1}$ from (10) and hence $\boldsymbol{\Theta}^{\mathcal{A}_0} = \boldsymbol{\Theta}^{\mathcal{A}_1}$; and also $(p_{\boldsymbol{\alpha}_1}, p_{\boldsymbol{\alpha}_2}) = (p_{\boldsymbol{\alpha}_3}, p_{\boldsymbol{\alpha}_2})$ by our construction that $p_{\boldsymbol{\alpha}_1} = p_{\boldsymbol{\alpha}_3}$. This implies even if one knows exactly there are two latent attribute patterns in the population, one can never tell which two patterns those are based on the likelihood function. In this sense, $\mathcal{A}_0$ is not identifiable, due to the fact that $\Gamma^{\mathcal{A}_0}$ and $\Gamma^{\mathcal{A}_1}$ do not lead to distinguishable distributions of responses under the two-parameter SLAM.*

From the above example, to make sure the set of true attribute patterns $\mathcal{A}_0$ is learnable from the observed multivariate responses, we need the $\Gamma^{\mathcal{A}_0}$-matrix to have certain structures. We state the formal definition of (strict) learnability of $\mathcal{A}_0$.

Definition 1 (strict learnability of $\mathcal{A}_0$) *Given Q, the set $\mathcal{A}_0$ is said to be (strictly) learnable, if for any constraint matrix $\Gamma^{\mathcal{A}}$ of size $J \times |\mathcal{A}|$ with $|\mathcal{A}| \leq |\mathcal{A}_0|$, any valid item parameters $\boldsymbol{\Theta}^{\mathcal{A}}$ respecting constraints given by $\Gamma^{\mathcal{A}}$, and any proportion parameters $\boldsymbol{p}^{\mathcal{A}} \in \mathcal{T}^{|\mathcal{A}|-1}$, $\boldsymbol{p}^{\mathcal{A}} \succ \mathbf{0}$, the following equality*

$$\mathbb{P}(\mathbf{R} \mid \boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0}) = \mathbb{P}(\mathbf{R} \mid \boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) \tag{11}$$

implies $\mathcal{A} = \mathcal{A}_0$. Moreover, if (11) implies $(\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) = (\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$, then we say the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ are (strictly) identifiable.

Next we further introduce some notations and definitions about the constraint matrix Γ and then present the needed identifiability result. Consider an arbitrary subset of items $S \subseteq \{1, \ldots, J\}$. For $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \mathcal{A}$, we denote $\boldsymbol{\alpha} \succeq_S \boldsymbol{\alpha}'$ under $\Gamma^{\mathcal{A}}$, if for each $j \in S$ there is $\Gamma^{\mathcal{A}}_{j,\boldsymbol{\alpha}} \geq \Gamma^{\mathcal{A}}_{j,\boldsymbol{\alpha}'}$. If viewing $\Gamma_{j,\boldsymbol{\alpha}} = 1$ as $\boldsymbol{\alpha}$ being "capable" of item j, then $\boldsymbol{\alpha} \succeq_S \boldsymbol{\alpha}'$ would mean $\boldsymbol{\alpha}$ is at least as capable as $\boldsymbol{\alpha}'$ of items in set S. Then under Γ, any subset of items S defines a partial order "$\succeq_S$" on the set of latent attribute patterns $\mathcal{A}$. For two item sets S_1 and S_2, we say " $\succeq_{S_1}$ " = " $\succeq_{S_2}$ " under $\Gamma^{\mathcal{A}}$, if for any $\boldsymbol{\alpha}', \boldsymbol{\alpha} \in \mathcal{A}$, there is $\boldsymbol{\alpha} \succeq_{S_1} \boldsymbol{\alpha}'$ under $\Gamma^{\mathcal{A}}$ if and only if $\boldsymbol{\alpha} \succeq_{S_2} \boldsymbol{\alpha}'$ under $\Gamma^{\mathcal{A}}$. The next theorem gives conditions that ensure the constraint matrix Γ as well as the Γ-constrained model parameters are jointly identifiable.

Theorem 2 (conditions for strict learnability) *Consider a SLAM with an arbitrary set of true attribute patterns $\mathcal{A}_0 \subseteq \{0,1\}^K$, and a corresponding constraint matrix $\Gamma^{\mathcal{A}_0}$. If this true $\Gamma^{\mathcal{A}_0}$ satisfies the following conditions, then $\mathcal{A}_0$ is identifiable.*

- *A. There exist two disjoint item sets S_1 and S_2, such that $\Gamma^{(S_i,\mathcal{A}_0)}$ has distinct column vectors for $i = 1, 2$ and "$\succeq_{S_1}$=$\succeq_{S_2}$" under $\Gamma^{\mathcal{A}_0}$.*
- *B. For any $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}' \in \mathcal{A}_0$ where $\boldsymbol{\alpha}' \succeq_{S_i} \boldsymbol{\alpha}$ under $\Gamma^{\mathcal{A}_0}$ for $i = 1$ or 2, there exists some $j \in (S_1 \cup S_2)^c$ such that $\Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}} \neq \Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}'}$.*
- *C. Any column vector of $\Gamma^{\mathcal{A}_0}$ is different from any column vector of $\Gamma^{\mathcal{A}_0^c}$, where $\mathcal{A}_0^c = \{0,1\}^K \setminus \mathcal{A}_0$.*

Recall that each column in the Γ-matrix corresponds to a latent attribute pattern, then Conditions A and B help ensure the Γ-matrix of the true patterns $\Gamma^{\mathcal{A}_0}$ contains enough information to distinguish between these true patterns. Specifically, Condition A requires $\Gamma^{\mathcal{A}_0}$ to contain two vertically stacked submatrices corresponding to item sets S_1 and S_2, each having distinct columns, i.e., each being able to distinguish between the true patterns; and Condition B requires the remaining submatrix of $\Gamma^{\mathcal{A}_0}$ to distinguish those pairs of true patterns that have some order ($\boldsymbol{\alpha}' \succeq_{S_i} \boldsymbol{\alpha}$) based on the first two item sets S_1 or S_2. Condition C is necessary for identifiability of $\mathcal{A}_0$ by ensuring that any true pattern would have a different column vector in Γ^{all} from that of any false pattern. Condition C is satisfied for any $\mathcal{A}_0 \subseteq \{0,1\}^K$ if the Q-matrix contains an identity submatrix I_K, because such a Q-matrix will give a Γ^{all} that has all the 2^K columns distinct.

We would like to point out that our identifiability conditions in Theorem 2 do not depend on the unknown parameters (e.g., $\boldsymbol{\Theta}$ and $\boldsymbol{p}$), but only rely on the structure of the constraint matrix Γ. The Γ-matrix with respect to the true set of patterns $\mathcal{A}_0$ is the key quantity that defines the latent structure of a SLAM. Generally, it is hard to establish identifiability conditions that only depend on the cardinality of $\mathcal{A}_0$ but not on $\Gamma^{\mathcal{A}_0}$. For instance, in Example 5, the two sets $\mathcal{A}_0$ and $\mathcal{A}_1$ have the same cardinality but can not be distinguished under the conditions there; indeed further conditions on Q (and the resulting Γ) are needed to guarantee identifiability.

The developed identifiability conditions generally apply to any SLAM satisfying the constraints (2) and (3) introduced in Section 2.1. If one makes further assumptions on $\boldsymbol{\Theta}$, such as assuming each item $j \in [J]$ has exactly two item parameters to make it a two-parameter model, then the conditions in Theorem 2 may be further relaxed. For example, in the saturated case with $\mathcal{A}_0 = \{0,1\}^K$, the sufficient identifiability conditions developed in Xu (2017) for a general SLAM require Q to contain two copies of I_K as submatrices, while the necessary and sufficient conditions established in Gu and Xu (2019a) for the two-parameter SLAM require Q to have just one submatrix I_K. We expect that in the current case with an arbitrary $\mathcal{A}_0 \subseteq \{0,1\}^K$, the conditions in Theorem 2 can also be relaxed under the two-parameter model in a technically nontrivial way. For the reason of generality, we focus on SLAMs under the general constraints (2) and (3) in this work.

When the conditions in Theorem 2 are satisfied, $\mathcal{A}_0$ is identifiable; and from Theorem 4.1 in Gu and Xu (2019b), the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ associated with $\mathcal{A}_0$ are also identifiable.

Corollary 3 *Under the conditions in Theorem 2, the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ associated with $\mathcal{A}_0$ are identifiable.*

Note that the result of Theorem 2 differs from the existing works Xu (2017), Xu and Shang (2018) and Gu and Xu (2019b) in that those works assume $\mathcal{A}_0$ is known *a priori* and study the identifiability of $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$, while in the current work $\mathcal{A}_0$ is unknown and we focus on the identifiability of $\mathcal{A}_0$ itself. This is crucially needed in order to guarantee that we can learn the set of true attribute patterns.

Remark 4 *The identifiability results in Theorem 2 and Corollary 3 are related to the uniqueness of tensor decomposition. As shown in* (8)*, the probability mass function of the multivariate responses of each subject can be viewed as a higher order tensor with constraints on entries of the tensor, and unique decomposition of the tensor correspond to identification of the constraint matrix as well as the model parameters. The identifiability conditions in Theorem 2 are weaker than the general conditions for uniqueness of three-way tensor decomposition in Kruskal (1977), which is a celebrated result in the literature. Kruskal's conditions require the tensor can be decomposed as a Khatri-Rao product of three matrices, two having full-rank and the other having Kruskal rank at least two (Kruskal rank of a matrix is the largest number T such that every set of T columns of it are linearly independent). Consider an example with $J = 5$, $K = 2$, $\mathcal{A}_0 = \{\boldsymbol{\alpha}_2 = (0,1),\ \boldsymbol{\alpha}_3 = (1,0)\}$, and the corresponding $\Gamma^{\mathcal{A}_0}$ in the form of* (12)*. Then we can set $S_1 = \{1,2\}$, $S_2 = \{3,4\}$ and Condition A in Theorem 2 is satisfied. Further, Condition B is also satisfied since $\boldsymbol{\alpha}_2 \not\succeq_{S_i} \boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_3 \not\succeq_{S_i} \boldsymbol{\alpha}_2$ under $\Gamma^{\mathcal{A}_0}$. Therefore, Theorem 1 guarantees the set $\mathcal{A}_0$ is identifiable,*

and further guarantees the parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ *are identifiable. On the contrary, results based on Kruskal's conditions for unique three-way tensor decomposition can not guarantee identifiability, because other than two full rank structures given by the items in* S_1 *and* S_2*, the remaining item 5 in* $(S_1 \cup S_2)^c$ *corresponds to a structure with Kruskal rank only one.*

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \hline 1 & 0 \\ 0 & 1 \\ \hline 1 & 1 \end{pmatrix} \quad \Longrightarrow \quad \Gamma^{\mathcal{A}_0} = \begin{array}{c} \begin{matrix} \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3 \\ (0,1) & (1,0) \end{matrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \hline 1 & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{pmatrix} \end{array}. \tag{12}$$

We next discuss two extensions of the developed identifiability theory. First, Theorem 2 guarantees the strict learnability of $\mathcal{A}_0$. Under a multi-parameter SLAM, these conditions can be relaxed if the aim is to obtain the so-called generic joint identifiability of $\mathcal{A}_0$, which means that $\mathcal{A}_0$ is learnable with the true model parameters ranging almost everywhere in the constrained parameter space except a set with Lebesgue measure zero. Specifically, we have the following definition.

Definition 5 (generic learnability of the true model) *Denote the parameter space of* $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ *constrained by* $\Gamma^{\mathcal{A}_0}$ *by* Ω*. We say* $\mathcal{A}_0$ *is generically identifiable, if there exists a subset* $\mathcal{V}$ *of* Ω *that has Lebesgue measure zero, such that for any* $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0}) \in \Omega \setminus \mathcal{V}$*, Equation* (11) *implies* $\mathcal{A} = \mathcal{A}_0$*. Moreover, if for any* $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0}) \in \Omega \setminus \mathcal{V}$*, Equation* (11) *implies* $(\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) = (\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$*, we say the model parameters* $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ *are generically identifiable.*

The generic learnability result is presented in the next theorem.

Theorem 6 (conditions for generic learnability) *Consider a multi-parameter SLAM with the set of true attribute patterns* $\mathcal{A}_0$ *and the* $J \times |\mathcal{A}_0|$ *constraint matrix* $\Gamma^{\mathcal{A}_0}$*. If* $\Gamma^{\mathcal{A}_0}$ *satisfies Condition C and also the following conditions, then* $\mathcal{A}_0$ *is generically identifiable.*

- $A^\star$. *There exist two disjoint item sets* S_1 *and* S_2*, such that altering some entries from 0 to 1 in* $\Gamma^{(S_1 \cup S_2, \mathcal{A}_0)}$ *can yield a* $\widetilde{\Gamma}^{(S_1 \cup S_2, \mathcal{A}_0)}$ *satisfying Condition A. That is,* $\widetilde{\Gamma}^{(S_i, \mathcal{A}_0)}$ *has distinct columns for* $i = 1, 2$ *and* " $\succeq_{S_1}$ " = " $\succeq_{S_2}$ " *under* $\widetilde{\Gamma}^{(S_1 \cup S_2, \mathcal{A}_0)}$*.*
- $B^\star$. *For any* $\boldsymbol{\alpha}$*,* $\boldsymbol{\alpha}' \in \mathcal{A}_0$ *where* $\boldsymbol{\alpha}' \succeq_{S_i} \boldsymbol{\alpha}$ *under* $\widetilde{\Gamma}^{(S_1 \cup S_2, \mathcal{A}_0)}$ *for* $i = 1$ *or* 2*, there exists some* $j \in (S_1 \cup S_2)^c$ *such that* $\Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}} \neq \Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}'}$*.*

We also have the following corollary, where the identifiability requirements are directly characterized by the structure of the Q-matrix, instead of Γ.

Corollary 7 *If the* Q*-matrix satisfies the following conditions, then for any true set of attribute patterns* $\mathcal{A}_0 \subseteq \{0,1\}^K$ *such that* $\Gamma^{\mathcal{A}_0}$ *satisfies Condition C, the set* $\mathcal{A}_0$ *is generically identifiable.*

(A**) The Q contains two $K \times K$ sub-matrices Q_1, Q_2, such that for $i = 1, 2$,

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q' \end{pmatrix}_{J \times K} ; \quad Q_i = \begin{pmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & 1 \end{pmatrix}_{K \times K}, \quad i = 1, 2, \tag{13}$$

where each '$*$' can be either zero or one.

(B**) With Q in the form of (13), there is $\sum_{j=2K+1}^{J} q_{j,k} \geq 1$ for each $k \in \{1, \dots, K\}$.

Remark 8 *When the conditions in Theorem 7 are satisfied, $\mathcal{A}_0$ is generically identifiable and from Theorem 4.3 in Gu and Xu (2019b), the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ are also generically identifiable. Corollary 7 differs from Theorem 4.3 in Gu and Xu (2019b) in that, here we allow the true set of attribute patterns $\mathcal{A}_0$ to be unknown and arbitrary, and study its identifiability, while Gu and Xu (2019b) assumes $\mathcal{A}_0$ is pre-specified and studies the identifiability of the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$.*

Remark 9 *Under the conditions for generic identifiability in Theorem 6 or Corollary 7, we can obtain the explicit forms of the measure zero set $\mathcal{V}$ ($\mathcal{V} \subseteq \Omega$) where the non-identifiability may occur. Under either Theorem 6 or Corollary 7, the set $\mathcal{V}$ is characterized by the zero set of certain polynomials about the parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ (see the proofs for details). The zero set of these polynomials indeed defines a lower-dimensional manifold in the parameter space. Therefore, Theorem 6 and Corollary 7 supplement Theorem 2 by relaxing the original conditions and establishing identifiability when $(\boldsymbol{\Theta}, \boldsymbol{p})$ satisfy certain shape constraints, i.e., $(\boldsymbol{\Theta}, \boldsymbol{p})$ do not fall on that manifold $\mathcal{V}$ in the parameter space.*

The above generic identifiability results of $\mathcal{A}_0$ ensure that nonidentifiability happens only in a measure zero set in the parameter space. Next, we develop a second extension of Theorem 2 for scenarios where nonidentifiability cases occupy a positive measure set in the parameter space. This situation happens when certain latent attribute patterns always have the same item parameters across all the items, i.e., $\boldsymbol{\Theta}_{\cdot,\boldsymbol{\alpha}} = \boldsymbol{\Theta}_{\cdot,\boldsymbol{\alpha}'}$ for some $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$. We define $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ to be in the same equivalence class if $\boldsymbol{\Theta}_{\cdot,\boldsymbol{\alpha}} = \boldsymbol{\Theta}_{\cdot,\boldsymbol{\alpha}'}$. For instance, still consider the following 2×2 Q-matrix under the two-parameter SLAM introduced in Example 2,

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \tag{14}$$

then attribute patterns $\boldsymbol{\alpha}_1 = (0, 0)$ and $\boldsymbol{\alpha}_3 = (1, 0)$ are equivalent under the two-parameter SLAM, as can be seen from the Γ^{all} in (10). Therefore the two latent patterns $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$ are not identifiable, no matter which values the true model parameters take.

In this case where both strict and generic identifiability do not hold, we study the $\boldsymbol{p}$-partial identifiability, a concept introduced in Gu and Xu (2019b). Specifically, when some attribute patterns have the same item parameters across all items, we define the set of these attribute patterns as an equivalence class, and aim to identify the proportion of this equivalence class, instead of the separate proportions of these equivalent patterns, in the population. For instance, in the above example in (14), because $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$ are

equivalent, there are three equivalence classes: $\{\boldsymbol{\alpha}_1 = (0,0), \boldsymbol{\alpha}_3 = (1,0)\}$, $\{\boldsymbol{\alpha}_2 = (0,1)\}$, and $\{\boldsymbol{\alpha}_4 = (1,1)\}$. We denote these three equivalence classes by $[\boldsymbol{\alpha}_1]$ (or $[\boldsymbol{\alpha}_3]$, since $[\boldsymbol{\alpha}_1] = [\boldsymbol{\alpha}_3]$), $[\boldsymbol{\alpha}_2]$ and $[\boldsymbol{\alpha}_4]$, since $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_4$ form a complete set of representatives of the equivalence classes. For any Q, we denote the induced set of equivalence classes by $\mathcal{A}^{\text{equiv}} = \{[\boldsymbol{\alpha}_1], \ldots, [\boldsymbol{\alpha}_C]\}$, where $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C$ form a complete set of representatives of the equivalence classes. In this case, the pattern selection problem of interest is to learn which equivalence classes in $\mathcal{A}^{\text{equiv}}$ are significant.

For the two-parameter SLAM introduced in Example 2, two attribute patterns $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ are in the same equivalence class if and only if $\Gamma^{\mathcal{A}}_{\cdot,\boldsymbol{\alpha}_1} = \Gamma^{\mathcal{A}}_{\cdot,\boldsymbol{\alpha}_2}$. This is because under the two-parameter SLAM, the Γ-matrix determined by the Q-matrix with $\Gamma_{j,\boldsymbol{\alpha}} = I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j)$ fully captures the model structure in the sense that $\theta_{j,\boldsymbol{\alpha}} = \theta_j^+ \Gamma_{j,\boldsymbol{\alpha}} + \theta_j^- (1 - \Gamma_{j,\boldsymbol{\alpha}})$. Therefore under a two-parameter SLAM, we can obtain a complete set of representatives of the equivalence classes directly from the $\boldsymbol{q}$-vectors, which are

$$\mathcal{A}_Q = \{\vee_{j \in S} \boldsymbol{q}_j : S \subseteq \{1, \ldots, J\}\}, \tag{15}$$

where $\vee_{j \in S} \boldsymbol{q}_j = (\max_{j \in S} q_{j,1}, \ldots, \max_{j \in S} q_{j,K})$. For $S = \varnothing$, we define the vector $\vee_{j \in S} \boldsymbol{q}_j$ to be $\mathbf{0}_K$, the all-zero attribute pattern. The reasons for $\mathcal{A}_Q$ being a complete set of representatives are that, first, $\Gamma^{\mathcal{A}_Q}$ has distinct columns and contains all the unique column vectors in Γ^{all}; and second, for any other pattern not in $\mathcal{A}_Q$, there is some pattern in $\mathcal{A}_Q$ such that the two patterns have identical column vectors in Γ^{all}. It is not hard to see that $\mathcal{A}_Q = \{0,1\}^K$ if and only if the Q-matrix contains a submatrix I_K.

For multi-parameter SLAMs introduced in Example 3, two attribute patterns $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ are in the same equivalence class if $\Gamma_{\cdot,\boldsymbol{\alpha}_1} = \Gamma_{\cdot,\boldsymbol{\alpha}_2} = \mathbf{1}$. This can be seen by considering $\Gamma_{\cdot,\boldsymbol{\alpha}_1} = \Gamma_{\cdot,\boldsymbol{\alpha}_1} \neq \mathbf{1}$, i.e., $\Gamma_{j,\boldsymbol{\alpha}_1} = \Gamma_{j,\boldsymbol{\alpha}_2} = 0$ for some item j. Then different from the two-parameter SLAMs, for such item j, the $\theta_{j,\boldsymbol{\alpha}_1}$ and $\theta_{j,\boldsymbol{\alpha}_2}$ are not always the same by the modeling assumptions of multi-parameter SLAMs. Indeed, under a multi-parameter SLAM, for item j, patterns in the set $\mathcal{A}_0 \setminus \mathcal{C}_j$ can have multiple levels of item parameters.

We have the following corollary of Theorem 2 on identifiability, when certain attribute patterns are not distinguishable. Denote the set of significant equivalence classes by $\mathcal{A}_0^{\text{equiv}} = \{[\boldsymbol{\alpha}_{\ell_1}], \ldots, [\boldsymbol{\alpha}_{\ell_m}]\}$, which is a subset of the saturated set $\mathcal{A}^{\text{equiv}} = \{[\boldsymbol{\alpha}_1], \ldots, [\boldsymbol{\alpha}_C]\}$. Denote the set of representative patterns of the significant equivalence classes by $\{\boldsymbol{\alpha}_{\ell_1}, \ldots, \boldsymbol{\alpha}_{\ell_m}\} = \mathcal{A}^{\text{rep}}$.

Corollary 10 *If the matrix $\Gamma^{\mathcal{A}^{\text{rep}}}$ satisfies Conditions A, B and C, $\mathcal{A}_0^{\text{equiv}}$ is identifiable.*

Remark 11 *Under the two-parameter SLAM with $\mathcal{A}^{equiv} = \{[\boldsymbol{\alpha}_1], \ldots, [\boldsymbol{\alpha}_C]\}$, the Γ-matrix $\Gamma^{\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C\}}$ by definition would have distinct column vectors. Therefore any column vector of $\Gamma^{\mathcal{A}^{\text{rep}}}$ in Corollary 10 must be different form any column vector of $\Gamma^{\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C\} \setminus \mathcal{A}^{\text{rep}}}$. In this case, Condition C is automatically satisfied. And in order to identify $\mathcal{A}_0^{\text{equiv}}$, one only needs to check if $\Gamma^{\mathcal{A}^{\text{rep}}}$ satisfies Conditions A and B.*

4. Penalized Likelihood approach to pattern selection

In this section, we first present the method of shrinkage estimation, and then describe a screening approach as a preprocessing step.

4.1. Shrinkage Estimation

The developed identifiability conditions guarantee that the true set of patterns can be distinguished from any alternative set that has not more than $|\mathcal{A}_0|$ patterns, since they would lead to different probability mass functions of the responses. As $\mathcal{A}_0 = \{\boldsymbol{\alpha} \in \{0,1\}^K : p_{\boldsymbol{\alpha}} > 0\}$, we know that learning the significant attribute patterns is equivalent to selecting the nonzero elements of the population proportion vector $\boldsymbol{p}$. In practice, if we directly overfit the data with all the 2^K possible attribute patterns, the corresponding maximum likelihood estimator (MLE) can not correctly recover the sparsity structure of the vector $\boldsymbol{p}$. In this case, we propose to impose some regularization on the proportion parameters $\boldsymbol{p}$, and perform pattern selection through maximizing a penalized likelihood function.

In general, we denote by $\mathcal{A}_{\text{input}}$ the set of candidate attribute patterns given to the shrinkage estimation method as input. If the saturated space of all the possible attribute patterns are considered, then $\mathcal{A}_{\text{input}} = \{0,1\}^K$ and it contains all the 2^K possible configurations of attributes. When $2^K \gg N$, we propose to use a preprocessing step that returns a proper subset $\mathcal{A}_{\text{input}}$ of the saturated set $\{0,1\}^K$ as candidate attribute patterns, and then perform the shrinkage estimation (please see Section 4.2 for the preprocessing procedure).

We first introduce the general data likelihood of a structured latent attribute model. Given a sample of size N, we denote the ith subject's response by $\mathbf{R}_i = (R_{i,1}, \ldots, R_{i,J})^\top$, $i = 1, \ldots, N$. We further use $\mathcal{R}$ to denote the $N \times J$ data matrix $(\mathbf{R}_1^\top, \ldots, \mathbf{R}_N^\top)^\top$. The marginal likelihood can be written as

$$L(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R}) = \prod_{i=1}^{N} \Big[\sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{R_{i,j}} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big], \tag{16}$$

where the constraints on $\boldsymbol{\Theta}$ imposed by Q are made implicit. We denote the corresponding log likelihood by $\ell(\boldsymbol{\Theta}, \boldsymbol{p}) = \log L(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R})$.

As the proportion parameters $\boldsymbol{p}$ belongs to a simplex, in order to encourage sparsity of $\boldsymbol{p}$, we propose to use a log-type penalty with a tuning parameter $\lambda < 0$. Specifically, we use the following penalized likelihood as the objective function,

$$\ell^\lambda(\boldsymbol{\Theta}, \boldsymbol{p}) \ = \ell(\boldsymbol{\Theta}, \boldsymbol{p}) + \lambda \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} \log_{\rho_N}(p_{\boldsymbol{\alpha}}), \quad \lambda \in (-\infty, 0), \tag{17}$$

where $\log_{\rho_N}(p_{\boldsymbol{\alpha}}) = \log(p_{\boldsymbol{\alpha}}) \cdot I(p_{\boldsymbol{\alpha}} > \rho_N) + \log(\rho_N) \cdot I(p_{\boldsymbol{\alpha}} \leq \rho_N)$ and ρ_N is a small threshold parameter that is introduced to circumvent the singularity issue of the log function at zero. Specifically, we take

$$\rho_N \asymp N^{-d} \tag{18}$$

for some constant $d \geq 1$, where for two sequences $\{a_N\}$ and $\{b_N\}$, we denote $a_N \lesssim b_N$ if $a_N = O(b_N)$ and $a_N \asymp b_N$ if $a_N \lesssim b_N$ and $b_N \lesssim a_N$. Any attribute pattern $\boldsymbol{\alpha}$ whose estimated $p_{\boldsymbol{\alpha}} < \rho_N$ will be considered as 0, and hence not selected. The tuning parameter $\lambda \in (-\infty, 0)$ controls the sparsity level of the estimated proportion vector $\boldsymbol{p}$, and a smaller λ leads to a sparser solution (with more estimated proportion $p_{\boldsymbol{\alpha}}$ falling below ρ_N). Given a $\lambda \in (-\infty, 0)$, we denote the estimated set of patterns by $\widehat{\mathcal{A}}^\lambda = \{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}} : \widehat{p}_{\boldsymbol{\alpha}} > \rho_N, (\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}}) = \arg\max_{\boldsymbol{\Theta}, \boldsymbol{p}} \ell^\lambda(\boldsymbol{\Theta}, \boldsymbol{p})\}$.

Remark 12 *In the literature, Chen et al. (2001) and Chen et al. (2004) used a similar form of penalty as the summation term in our* (17)*, but instead imposed* $\lambda > 0$ *to avoid sparse solutions of the proportion parameters. These works used that penalty in order to avoid singularity when performing restricted likelihood ratio test. While our goal here is to encourage sparsity of* $\boldsymbol{p}$ *so that significant attribute patterns can be selected.*

The formulation of (17) *can also be interpreted in a Bayesian way, where the penalty term regarding the proportions* $\boldsymbol{p}$ *is the logarithm of the Dirichlet prior density with hyperparameter* $\beta = \lambda + 1$ *over the proportions. But note that when* $\beta < 0$*, the penalty term is not a proper prior density. Our later Proposition 15 reveals that, under nonstandard convergence rate of the mixture model, the traditional Bayesian way of imposing a proper Dirichlet prior over proportions is not sufficient for selecting significant attribute patterns consistently. Instead, this classical procedure will yield too many false patterns being selected. Therefore, our novelty of allowing* λ *in* (17) *to be negative with arbitrarily large magnitude is crucial to selection consistency.*

Other than the nice connection to the Dirichlet prior density in the Bayesian literature, the log-type penalty in (17) *also facilitates the computation based on modified EM and variational EM algorithms, as shown in our Algorithms 1 and 2. For such reasons, this work uses the log-type penalty. There are also alternative ways of imposing penalty on the proportion parameters* $\boldsymbol{p}$ *that would lead to selection consistency, such as the truncated* L_1 *penalty used in Shen et al. (2012) for high-dimensional feature selection.*

We denote the MLE obtained from directly maximizing $L(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R})$ in (16) by $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{p}}$, and denote the "oracle" MLE of the parameters obtained by maximizing the likelihood constrained to the true set of attribute patterns by $(\widehat{\boldsymbol{\Theta}}^{\mathcal{A}_0}, \widehat{\boldsymbol{p}}^{\mathcal{A}_0})$. We denote the rate of convergence of $\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$ to $\ell(\widehat{\boldsymbol{\Theta}}^{\mathcal{A}_0}, \widehat{\boldsymbol{p}}^{\mathcal{A}_0})$ by $\delta \in (0, 1]$, that is,

$$\big[\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}}) - \ell(\widehat{\boldsymbol{\Theta}}^{\mathcal{A}_0}, \widehat{\boldsymbol{p}}^{\mathcal{A}_0})\big]/N = O_P(N^{-\delta}). \tag{19}$$

When $\delta = 1$, (19) implies $\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$ converges with the usual root-N rate, and $\delta < 1$ would imply a slower convergence rate. In the literature, Ho and Nguyen (2016) and Heinrich and Kahn (2018) have studied the technically involved problem of convergence rate of the mixing distribution of certain mixture models, and showed these models may not have the standard root-N rate. As implied by these works, for complicated models like SLAMs, the convergence rate of the mixing distribution is likely to be slower than root-N, so as the convergence rate of $\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$.

For a set $\mathcal{A}$, denote its cardinality by $|\mathcal{A}|$. We have the following theorem.

Theorem 13 (selection consistency) *Suppose the true constraint matrix* $\Gamma^{\mathcal{A}_0}$ *associated with* $\mathcal{A}_0$ *satisfies conditions A, B and C in Theorem 2. The true parameters satisfy*

$$\min_{\boldsymbol{\alpha} \in \mathcal{A}_0} p_{\boldsymbol{\alpha}} > c_0; \quad \theta_{j,\boldsymbol{\alpha}^\star} - \max_{\boldsymbol{\alpha}: \Gamma_{j,\boldsymbol{\alpha}} = 0} \theta_{j,\boldsymbol{\alpha}} \geq c_1, \ \forall\ j = 1, \ldots, J \text{ and } \boldsymbol{\alpha}^\star \in \mathcal{C}_j, \tag{20}$$

where $c_0, c_1 > 0$ *are some constants. Assume* $\log |\mathcal{A}_{\text{input}}| = o(N)$ *and* $|\mathcal{A}_{\text{input}}| \cdot \rho_N = O(N^{-\delta})$*. Then there exist a sequence of tuning parameters* $\{\lambda_N\}$ *satisfying* $N^{1-\delta}/|\log \rho_N| \lesssim -\lambda_N \lesssim N/|\log \rho_N|$ *such that* $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} = \mathcal{A}_0) \to 1$ *as* $N \to \infty$*.*

Remark 14 *Together with our identifiability result in Theorem 2, the assumption* (20) *helps distinguish the true patterns from any alternative set of patterns with no larger cardinality, and further helps establish selection consistency. It is possible to further extend the current result and relax the constant lower bound assumption, though identifiability conditions would need to be adapted carefully to the case with a growing number of significant patterns and a shrinking magnitude of the proportions; we leave this for future work.*

The proof of Theorem 13 also reveals that if the convergence rate of U_N are slower than $\sqrt{N}$ with $\delta < 1$ in (19), then the tuning parameter λ in (17) has to satisfy $\lambda < -1$ in order to have pattern selection consistency; otherwise the issue of over selecting exists. Under the Bayesian interpretation as discussed in Remark 12, this result implies that imposing the popular Dirichlet prior with a proper hyperparameter $\beta = \lambda + 1 \in (0, 1)$ is not sufficient for consistent selection of the significant mixture components (i.e., latent attribute patterns). Therefore, the approach proposed by Rousseau and Mengersen (2011) would not yield frequentist selection consistency in this considered scenario. We state this in the following proposition.

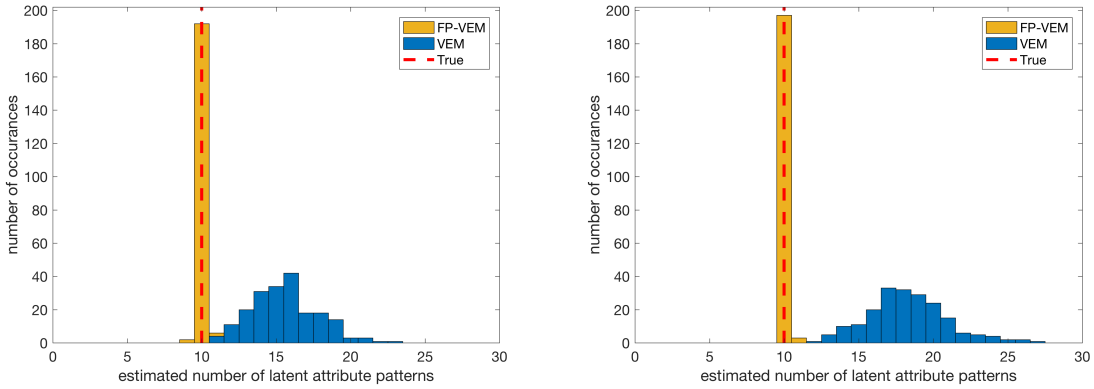
Proposition 15 (selection inconsistency of Dirichlet prior) *Suppose $\delta < 1$ in* (19)*, i.e., the rate of convergence of $\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$ is slower than the usual $\sqrt{N}$-rate. Then there does not exist a sequence of $\{\lambda_N,\ N = 1, 2, ...\} \subseteq [-1, 0)$ such that $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} = \mathcal{A}_0) \to 1$ as $N \to \infty$.*

Example 6 *To visualize how the numbers of selected patterns differ for our proposed method based on maximizing* (17) *with $\beta = \lambda + 1 \in (-\infty, 1)$, and the variational EM algorithm resulting from imposing a proper Dirichlet prior over the proportions, we conduct a simulation study. In a simulation setting of $K = 10$ and $J = 30$, for each sample size $N = 500$ and* 1000*, we carry out* 200 *independent runs and in each run record the number of selected attribute patterns given by the proposed method, and that by the variational EM algorithm. We plot the histogram corresponding to the proposed method (FP-VEM, see Section 4 for details), together with that corresponding to Variational EM (VEM) with a small Dirichlet parameter $\beta = 0.01$. For both algorithms, we use the same threshold $\rho_N = 1/(2N)$ for selecting attribute patterns in the end of the algorithm, by only keeping patterns whose posterior means exceeds ρ_N. Here we did not plot the results corresponding to VEM with β smaller than* 0.01*, because we found the VEM algorithm with smaller β values can have convergence issues and in many cases it fails to converge but just jumps between several solutions. One can see from Figure 2 that the proposed method selects 10 patterns for most of datasets, which are indeed the 10 true patterns; while VEM over selects the patterns.*

We next propose two algorithms to perform pattern selection, one being a modification of an EM algorithm, and the other being a variational EM algorithm resulting from an alternative formulation of the problem.

4.1.1. Modified EM algorithm.

We first consider using an EM algorithm with a slight modification in the E step to maximize (17). For each subject $i = 1, \ldots, N$, denote his/her latent attribute pattern by $\mathbf{A}_i =$



(a) $K = 10,\ N = 500$ (b) $K = 10,\ N = 1000$

Figure 2: Histograms of estimated number of latent attribute patterns. VEM represents Variational EM with $\beta = \lambda + 1 = 0.01$, and FP-VEM represents the proposed Algorithm 2 in Section 4. The true number of latent attribute patterns is $|\mathcal{A}_0| = 10$.

$(A_{i,1}, \ldots, A_{i,K})$, then $\mathbf{A}_i \in \{0,1\}^K$. The complete log likelihood corresponding to (17) is

$$\ell^{\lambda}_{\text{comp}}(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R}, \mathbf{A}) = \sum_{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}}} \Big(\sum_i I(\mathbf{A}_i = \boldsymbol{\alpha}_l) + \lambda \Big) \log_{\rho_N}(p_{\boldsymbol{\alpha}_l}) \tag{21}$$
$$+ \sum_{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}}} \sum_i I(\mathbf{A}_i = \boldsymbol{\alpha}_l) \sum_j \Big[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}_l}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}_l}) \Big],$$

where $I(\cdot)$ denotes the binary indicator function. Following the standard formulation of the EM algorithm (Dempster et al., 1977), in the E step of the $(t+1)$-th iteration, conditional expectations of $\ell^{\lambda}_{\text{comp}}(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R}, \mathbf{A})$ is evaluated with respect to the posterior distribution of latent variables $\mathbf{A}_i$'s given the current iterates of parameters $\boldsymbol{\Theta}^{(t)}$ and $\boldsymbol{p}^{(t)}$. Specifically, in the E step we replace the indicator $I(\mathbf{A}_i = \boldsymbol{\alpha}_l)$ in (21) by the probability $\varphi_{i,l} = \mathbb{P}(\mathbf{A}_i = \boldsymbol{\alpha}_l \mid \boldsymbol{\Theta}^{(t)}, \boldsymbol{p}^{(t)})$; and this is equivalent to updating

$$Q(\boldsymbol{\Theta}, \boldsymbol{p} \mid \boldsymbol{\Theta}^{(t)}, \boldsymbol{p}^{(t)}) := \mathbb{E}\Big[\ell^{\lambda}_{\text{comp}}(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R}, \mathbf{A}) \ \Big|\ \boldsymbol{\Theta}^{(t)}, \boldsymbol{p}^{(t)} \Big].$$

In the M step, we update $(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{p}^{(t+1)}) = \arg\max Q(\boldsymbol{\Theta}, \boldsymbol{p} \mid \boldsymbol{\Theta}^{(t)}, \boldsymbol{p}^{(t)})$. Note that directly using a negative λ in the EM algorithm may yield an invalid E step, due to potentially negative updates for some proportion parameters (e.g., $p_{\boldsymbol{\alpha}}$'s). When this happens, we do a thresholding in the E step as an approximation by replacing the probably negative class potential (Δ_l in Algorithm 1) with a pre-specified small constant $c > 0$. In practice, Algorithm 1's performance appears not sensitive to small values of c, and we take $c = 0.01$ in our numerical experiments; see Appendix B for a sensitivity study of the parameter c.

Remark 16 *Under the two-parameter SLAM, the DINA model in Example 2, or the identity-link multi-parameter all-effect SLAM, the GDINA model in Example 3, the M-step of updating the item parameters $\{\theta_{j,\boldsymbol{\alpha}}\}$'s in Algorithm 1 has closed forms. Specifically, under*

Algorithm 1: PEM: Penalized EM for log-penalty with $\lambda \in (-\infty, 0)$

Data: Q, responses $\mathcal{R}$, and candidate attribute patterns $\mathcal{A}_{\text{input}}$.

Initialize $\boldsymbol{\Delta} = (\Delta_1^{(0)}, \ldots, \Delta_{|\mathcal{A}_{\text{input}}|}^{(0)})$.

while *not converged* **do**

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

for $(i, l) \in [N] \times [|\mathcal{A}_{\text{input}}|]$ **do**

$$\varphi_{i,\boldsymbol{\alpha}_l}^{(t+1)} = \frac{\Delta_l^{(t)} \cdot \exp\Big\{\sum_j \Big[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}_l}^{(t)}) + (1-R_{i,j})\log(1-\theta_{j,\boldsymbol{\alpha}_l}^{(t)})\Big]\Big\}}{\sum_m \Delta_m^{(t)} \cdot \exp\Big\{\sum_j \Big[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}_m}^{(t)}) + (1-R_{i,j})\log(1-\theta_{j,\boldsymbol{\alpha}_m}^{(t)})\Big]\Big\}};$$

for $l \in [|\mathcal{A}_{\text{input}}|]$ **do**

$\Delta_l^{(t+1)} = \max\{c,\ \lambda + \sum_{i=1}^N \varphi_{i,\boldsymbol{\alpha}_l}^{(t+1)}\}$; $(c > 0$ is pre-specified$)$;

$\boldsymbol{p}^{(t+1)} \leftarrow \boldsymbol{\Delta}^{(t+1)}/(\sum_l \Delta_l^{(t+1)})$;

for $j \in [J]$ **do**

$$\boldsymbol{\Theta}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} \Big\{\sum_{\boldsymbol{\alpha}_l}\sum_i \varphi_{i,\boldsymbol{\alpha}_l}^{(t+1)} \sum_j \Big[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}_l}) + (1-R_{i,j})\log(1-\theta_{j,\boldsymbol{\alpha}_l})\Big]\Big\};$$

After the total T iterations,

Output: $\{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}} :\ p_{\boldsymbol{\alpha}_l}^{(T)} > \rho_N\}$.

DINA, for any item j the update for the unique parameters (θ_j^+, θ_j^-) takes the form

$$(\theta_j^+)^{(t+1)} = \frac{\sum_i \sum_{\boldsymbol{\alpha}} R_{i,j}\Gamma_{j,\boldsymbol{\alpha}}\varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}{\sum_i \sum_{\boldsymbol{\alpha}} \Gamma_{j,\boldsymbol{\alpha}}\varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}, \quad (\theta_j^-)^{(t+1)} = \frac{\sum_i \sum_{\boldsymbol{\alpha}} R_{i,j}(1-\Gamma_{j,\boldsymbol{\alpha}})\varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}{\sum_i \sum_{\boldsymbol{\alpha}} (1-\Gamma_{j,\boldsymbol{\alpha}})\varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}.$$

Under GDINA, for item j, the update for the unique parameters $\theta_{j,\{k_1,\ldots,k_l\}}$ with $\{k_1, \ldots, k_l\} \subseteq \mathcal{K}_j$ takes the following form,

$$\theta_{j,\{k_1,\ldots,k_l\}}^{(t+1)} = \frac{\sum_i \sum_{\boldsymbol{\alpha}} I(\{k \in \mathcal{K}_j :\ \alpha_k = 1\} = \{k_1, \ldots, k_l\}) R_{i,j}\varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}{\sum_i \sum_{\boldsymbol{\alpha}} I(\{k \in \mathcal{K}_j :\ \alpha_k = 1\} = \{k_1, \ldots, k_l\})\varphi_{i,\boldsymbol{\alpha}}^{(t+1)}}.$$

In addition, when certain latent patterns are not distinguishable as discussed earlier in Corollary 10, we can easily modify Algorithm 1 from selecting attribute patterns to selecting equivalence classes of attribute patterns. For instance, under a two-parameter SLAM, given the row vectors $\{\boldsymbol{q}_j, j \in [J]\}$ of Q, we first obtain the representatives of the Q-induced equivalence classes: $\mathcal{A}_Q = \{\vee_{j\in S}\, \boldsymbol{q}_j :\ S \subseteq \{1, \ldots, J\}\}$, then get the ideal response matrix of $\mathcal{A}_Q$, namely $\Gamma(\cdot, \mathcal{A}_Q) = (\gamma_{j,l})_{J\times|\mathcal{A}_Q|}$ where $\gamma_{j,l} = I(\boldsymbol{\alpha}_l \succeq \boldsymbol{q}_j)$ for $\boldsymbol{\alpha}_l \in \mathcal{A}_Q$ and $j \in [J]$. After initializing $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_{|\mathcal{A}_Q|})$, we just follow the same iterative procedure as that of Algorithm 1 for the two-parameter SLAM. In the end of the algorithm, after calculating $\nu_{[\boldsymbol{\alpha}_l]} = \Delta_l/(\sum_m \Delta_m)$, we select those $[\boldsymbol{\alpha}_l]$ with proportion $\nu_{[\boldsymbol{\alpha}_l]}$ above a pre-specified threshold. From the selected equivalence classes of attribute profiles, we can go back to obtain their representatives which are combinations of the $\boldsymbol{q}$-vectors from $\mathcal{A}_Q$ defined in Equation (15).

In practice when applying the PEM algorithm, we recommend using a sequential procedure with a range of λ values $\lambda_1 > \lambda_2 > \cdots > \lambda_B$, where $\lambda_1 > -1$ is close to 0 and λ_B should be less than -1. Specifically, we start with the relatively large λ_1 and use the estimated parameters from PEM with λ_1 as initial values for the next round of PEM with λ_2. We do this sequentially with estimates from PEM with λ_b serving as initializations for PEM with λ_{b+1}. When this sequential procedure ends, we choose the final model from the total number of B estimated ones using certain information criterion.

Given the large model space, we propose to use the Extended Bayesian Information Criterion (EBIC) introduced in Chen and Chen (2008) to select the tuning parameter. Recall that we denote by $\mathcal{A}^\lambda$ the selected set of attribute patterns obtained by maximizing the penalized likelihood function (17) with the specific tuning parameter λ. And we denote the item parameters and proportion parameters defined on this $\mathcal{A}^\lambda$ by $\boldsymbol{\Theta}^{\mathcal{A}^\lambda}$ and $\boldsymbol{p}^{\mathcal{A}^\lambda}$, respectively. The EBIC family have the following information criterion

$$\mathrm{BIC}_\gamma(\mathcal{A}^\lambda) = -2\ell(\boldsymbol{\Theta}^{\mathcal{A}^\lambda}, \boldsymbol{p}^{\mathcal{A}^\lambda}) + |\mathcal{A}^\lambda| \log N + 2\gamma \log \binom{|\mathcal{A}_{\text{input}}|}{|\mathcal{A}^\lambda|},$$

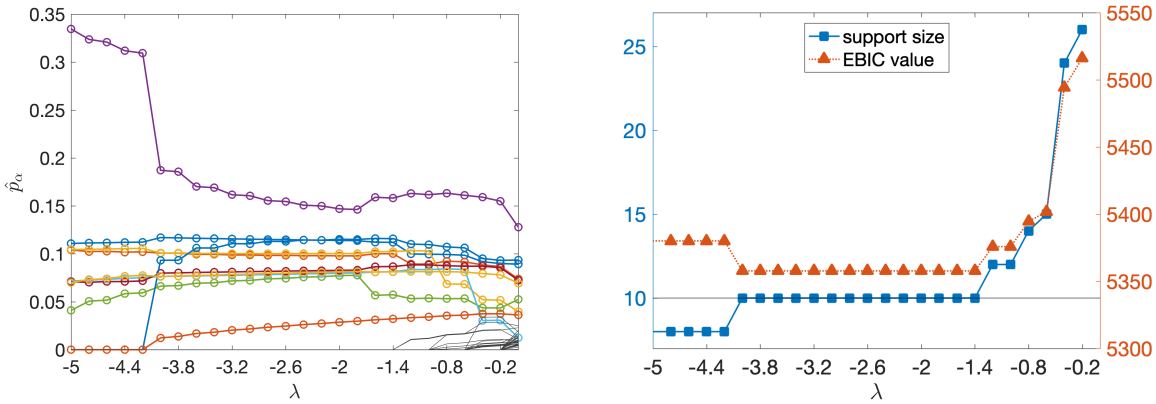
with the EBIC parameter $\gamma \in [0, 1]$. A smaller EBIC value implies a more favorable model. Selection consistency of the EBIC for high-dimensional model is established in Theorem 1 of Chen and Chen (2008) for γ greater than a certain threshold. When $\gamma = 0$, EBIC becomes the the classical BIC. Generally, larger γ yields a more parsimonious model. Here we choose $\gamma = 1$, for which the condition in Theorem 1 for selection consistency in Chen and Chen (2008) is satisfied.

Example 7 *Figure 3 presents an illustration of the solution paths of the estimated proportions versus λ based on a simulated dataset with $N = 150$, $K = 10$, and $J = 30$. The Q-matrix $Q = (Q_1^\top, Q_2^\top, Q_3^\top)^\top$ with Q_i in the following form,*

$$Q_1 = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 1 \end{pmatrix}, \quad Q_3 = \begin{pmatrix} 1 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 1 \end{pmatrix}. \tag{22}$$

When generating the data, 10 attribute patterns are randomly selected from the $2^{10} = 1024$ possible ones as true patterns, and the proportion of each of them is set to be 0.1. The item parameters are set to $1 - \theta_j^+ = \theta_j^- = 0.2$ for each j under a two-parameter SLAM. In the current setting with $K = 10$, we take the set of patterns as input to the PEM algorithm to be $\mathcal{A}_{\text{input}} = \{0, 1\}^K$. Figure 3(a) plots the solution paths of the estimated proportions of all the $2^{10} = 1024$ attribute patterns as λ varies in $\{-0.2, -0.4, \cdots, -4.8, -5.0\}$. The 10 true attribute patterns are plotted with colored lines with circles while the remaining $2^{10} - 10$ attribute patterns are plotted with black solid lines. Figure 3(b) plots the estimated support size of $\boldsymbol{p}$ versus λ, and the EBIC value versus λ. We observe that when $\lambda \in [-4.4, -1.4]$, Algorithm 1 selects the correct model with 10 true attribute patterns. This interval of λ corresponds to a "stable window" of the estimation algorithm that gives the correct selection and also has the smallest EBIC value. For this specific dataset, the proposed method along with EBIC succeeds in selecting the true model. Please see Section 5 for more

simulation results which show that the proposed methods combined with EBIC indeed have good performance in general.



(a) solution paths versus λ (b) EBIC values and support sizes versus λ

Figure 3: PEM solution paths and EBIC values in one trial, $N = 150$.

4.1.2. Variational EM algorithm from an alternative formulation.

In the following, we discuss an alternative formulation of the objective function (17) and propose a variational EM algorithm for estimation, by treating the proportion parameters $\boldsymbol{p}$ as latent random variables. As discussed in Remark 12, for the objective function (17) with $\lambda \in (-\infty, -1]$, the penalty term $\prod_{l=1}^{2^K} p_{\boldsymbol{\alpha}_l}^{\lambda}$ does not correspond to a proper Dirichlet distribution density. However, for any arbitrarily small λ value, the objective function (17) can be replaced by the following alternative formulation:

$$\ell_{\text{pseudo}}^{\lambda,\Upsilon}(\boldsymbol{\Theta}, \boldsymbol{p}) = \Upsilon \cdot \ell(\boldsymbol{\Theta}, \boldsymbol{p}) + (\beta - 1) \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} \log_{\rho_N}(p_{\boldsymbol{\alpha}}) \quad \text{for } \beta \in (0,1),\ \Upsilon \in (0,1]. \tag{23}$$

where we introduce a new parameter $\Upsilon \in (0, 1]$ and replace λ with $\beta - 1$ to respect the conventional notation of a Dirichlet distribution with hyperparameter $\beta \in (0, 1)$ to encourage sparsity. With $\beta \in (0, 1)$ and $\Upsilon \in (0, 1]$, the ratio $(1 - \beta)/\Upsilon$ can be arbitrarily large when Υ is arbitrarily close to zero, therefore making (23) equivalent to (17).

In the new objective function (23), the penalty term $\prod_{l=1}^{2^K} p_{\boldsymbol{\alpha}_l}^{\beta-1}$, $\beta \in (0, 1)$, can be viewed as a well-defined Dirichlet density function for the latent variables $\boldsymbol{p}$. In (23), the first term is the logarithm of the likelihood function raised to a fractional power $\Upsilon \in (0, 1]$. One intuition behind (23) is that given a moderate sample size and a large number of potential latent patterns, one needs to downweight the influence of the data likelihood and magnify the prior information encoded by the Dirichlet prior, in order to have the sufficient extent of shrinkage. The fractional-powered likelihood multiplied by the Dirichlet density can then be treated as a loss function to minimize. The idea of assigning a fractional power to the likelihood was also used in the Bayesian literature, such as Bissiri et al.

(2016) and Holmes and Walker (2017) for Bayesian learning under model misspecification, and Yang et al. (2019) and Chérief-Abdellatif and Alquier (2018) for variational Bayesian inference. Different from these works, here we use the alternative formulation (23) of the original objective function (17) in order to consistently select the significant latent attribute patterns.

The formulation (23) allows for a variational EM algorithm for obtaining the item parameters $\boldsymbol{\Theta}$ and the posterior means of the latent variables $\boldsymbol{p}$. Here we treat $\boldsymbol{\Theta}$ still as model parameters, then we follow the general derivation of variational algorithms in Blei et al. (2017) to derive Algorithm 2. We denote the digamma function by $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$ for $x \in (0, \infty)$. In particular, the complete log likelihood is

$$\ell_{\text{comp}}^{\lambda,\Upsilon}(\boldsymbol{\Theta} \mid \mathcal{R}, \mathbf{A}, \boldsymbol{p}) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}}} \Big\{\Upsilon \cdot \Big[\sum_i I(\mathbf{A}_i = \boldsymbol{\alpha})\Big] + \beta - 1\Big\} \log_{\rho_N}(p_{\boldsymbol{\alpha}}) \tag{24}$$
$$+ \Upsilon \cdot \Big\{ \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}}} \sum_i I(\mathbf{A}_i = \boldsymbol{\alpha}) \sum_j \Big[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}}) + (1 - R_{i,j})\log(1 - \theta_{j,\boldsymbol{\alpha}})\Big]\Big\}.$$

In the variational E step, we first obtain the conditional probability of $I(\mathbf{A}_i = \boldsymbol{\alpha}_l)$ for each individual i and each input attribute pattern $\boldsymbol{\alpha}_l$, which we denote by $\varphi_{i,\boldsymbol{\alpha}_l}$. In updating this $\varphi_{i,\boldsymbol{\alpha}_l}$, the variational posterior distribution of the $p_{\boldsymbol{\alpha}}$'s are used, which is still a Dirichlet distribution with mean parameters $(\Delta_1, \ldots, \Delta_{|\mathcal{A}_{\text{input}}|})$ updated in the previous E step (or from initializations if in the first iteration). Then we update the mean parameters for the variational posterior distribution of $p_{\boldsymbol{\alpha}_l}$'s based on the obtained $\varphi_{i,\boldsymbol{\alpha}_l}$, following the conventional derivation in variational inference. After finishing this E step, in the M step we maximize the complete likelihood with respect to $\boldsymbol{\Theta}$, by substituting the $I(\mathbf{A}_i = \boldsymbol{\alpha}_l)$'s with $\varphi_{i,\boldsymbol{\alpha}_l}$'s. Note that taking the derivatives of (24) with respect to $\theta_{j,\boldsymbol{\alpha}_l}$'s does not involve either terms of $p_{\boldsymbol{\alpha}_l}$ or terms of Υ and β, so only $\varphi_{i,\boldsymbol{\alpha}_l}$ are used in the M step for updating $\boldsymbol{\Theta}$. Indeed, the M step of updating $\boldsymbol{\Theta}$ in the current Algorithm 2 takes the same form as that of Algorithm 1.

Similar to Algorithm 1, in the practical use of Algorithm 2 for pattern selection, we recommend using a sequential fitting procedure. For a small fixed $\beta > 0$, we choose a sequence of Υ values $1 > \Upsilon_1 > \Upsilon_2 > \cdots > \Upsilon_B > 0$ where Υ_1 should be close to 1 and Υ_B should be relatively small. In our simulation studies, we found a $\Upsilon_B = 0.3$ is sufficient in most of cases. Then we sequentially run Algorithm 2 for B times with fractional powers $\Upsilon_1, \ldots, \Upsilon_B$ respectively and use estimated parameters from FP-VEM with Υ_b as initial values for FP-VEM with Υ_{b+1}. In the end, we also use EBIC to select the best Υ. Since β and Υ can be viewed as acting together through the term $(1 - \beta)/\Upsilon$, in terms of practical parameter tuning, we recommend fixing β to a relatively small value, say $\beta = 0.01$, and let the fractional power $\Upsilon \in (0, 1]$ vary to control the sparsity level of the proportion parameters.

4.2. Screening as a preprocessing step when $2^K \gg N$

In many applications of SLAMs, the number of attribute patterns 2^K could be much larger than N. This is especially the case in the application of SLAMs in epidemiological and medical diagnosis (Wu et al., 2017, 2018). In such scenarios, given a sample with size of several thousands or hundreds, it is desirable to develop an efficient screening procedure to

Algorithm 2: FP-VEM: Fractional Power Variational EM for $\Upsilon \in (0, 1]$

Data: Q, $\mathcal{R}$, and candidate attribute patterns $\mathcal{A}_{\text{input}}$.

Initialize $\boldsymbol{\Delta} = (\Delta_1^{(0)}, \ldots, \Delta_{|\mathcal{A}_{\text{input}}|}^{(0)}) = (\beta, \ldots, \beta)$.

while *not converged* **do**

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

for $(i, l) \in [N] \times [|\mathcal{A}_{\text{input}}|]$ **do**

$$\varphi_{i,\boldsymbol{\alpha}_l}^{(t+1)} = \frac{\exp\left\{\Psi(\Delta_l^{(t)}) + \Upsilon \cdot \sum_j \left[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}_l}^{(t)}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}_l}^{(t)})\right]\right\}}{\sum_m \exp\left\{\Psi(\Delta_m^{(t)}) + \Upsilon \cdot \sum_j \left[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}_m}^{(t)}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}_m}^{(t)})\right]\right\}};$$

for $l \in [|\mathcal{A}_{\text{input}}|]$ **do**

$\Delta_l^{(t+1)} \leftarrow \beta + \Upsilon \times \sum_{i=1}^N \varphi_{i,l}^{(t+1)}$;

for $j \in [J]$ **do**

$$\boldsymbol{\Theta}^{(t+1)} = \arg\max_{\boldsymbol{\Theta}} \left\{ \sum_{\boldsymbol{\alpha}_l} \sum_i \varphi_{i,\boldsymbol{\alpha}_l}^{(t+1)} \sum_j \left[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}_l}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}_l})\right]\right\}$$

After the total T iterations,

for $\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}}$ **do**

$p_{\boldsymbol{\alpha}_l} \leftarrow \Delta_l^{(T)} / (\sum_m \Delta_m^{(T)})$.

output: $\{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}} : \ p_{\boldsymbol{\alpha}_l} > \rho_N\}$.

bring down the number of candidate attribute patterns, and then perform the shrinkage estimation.

We next describe our screening approach. Recall that for each subject $i = 1, \ldots, N$, we denote his/her latent attribute pattern by $\mathbf{A}_i = (A_{i,1}, \ldots, A_{i,K}) \in \{0, 1\}^K$. In the screening stage we jointly estimate the item parameters $\boldsymbol{\Theta}$ and the $\{\mathbf{A}_i, i \in [N]\}$ to get a rough estimation of each subject i's attribute pattern, and gather all the N estimated attribute profiles as candidate patterns. The estimation of $\boldsymbol{p}$ is postponed to the estimation stage. Under the basic two-parameter SLAM, the complete log likelihood involving the latent variables $\{\mathbf{A}_i, i \in [N]\}$ takes the form

$$\begin{aligned}\ell_{\text{complete}}(\boldsymbol{\Theta}, \mathbf{A}) = \sum_{i=1}^N \sum_{j=1}^J \Big[& R_{i,j} \Big(\prod_k A_{i,k}^{q_{j,k}} \log \theta_j^+ + (1 - \prod_k A_{i,k}^{q_{j,k}}) \log \theta_j^- \Big) \Big] \\ & + (1 - R_{i,j}) \Big(\prod_k A_{i,k}^{q_{j,k}} \log(1 - \theta_j^+) + (1 - \prod_k A_{i,k}^{q_{j,k}}) \log(1 - \theta_j^-) \Big) \Big].\end{aligned}$$

We next derive an algorithm with a stochastic EM flavor to estimate the posterior mean of each latent variable $A_{i,k}$, denoted by a matrix $(\widehat{a}_{i,k})$ of size $N \times K$, where $\widehat{a}_{i,k} = \mathbb{E}[A_{i,k} \mid \cdot]$. In the end of the algorithm, we obtain the binary matrix W containing the candidate attribute patterns by defining $W = (w_{i,k})_{N \times K}$ with $w_{i,k} = I(\widehat{a}_{i,k} > 1/2)$. In such a screening procedure, we first use the dependency among the K attributes in iterative updates, then partly ignore the dependency in the last step through applying Bayes' rule to each subject

i's each single attribute k. This results in fast and valid screening of attribute patterns. Viewing the ith row vector of W as the estimated attribute pattern of subject i, the unique row vectors in W are the roughly selected attribute patterns output by the screening stage. We denote this set of candidate patterns by $\widehat{\mathcal{A}}_{\text{screen}}$. As long as the screening has the nice property of "no false exclusion", meaning the rows in W contain all the true attribute patterns, then the screening stage is considered successful. The selected candidate patterns are passed along to the shrinkage estimation stage as input patterns.

We say the screening procedure has the *sure screening property* if as N goes to infinity, the probability of all the true attribute patterns included in $\widehat{\mathcal{A}}_{\text{screen}}$ goes to one. The next theorem establishes the sure screening property of the proposed screening procedure.

Theorem 17 (sure screening property) *Suppose the identifiability conditions in Theorem 2 and the constraints* (20) *are satisfied. The screening procedure applied to a SLAM that covers the two-parameter SLAM as a submodel has the sure screening property. Specifically, there exists a constant* $\beta_{\min} > 0$ *such that* $\mathbb{P}(\widehat{\mathcal{A}}_{\text{screen}} \supseteq \mathcal{A}_0) \geq 1 - |\mathcal{A}_0| \exp(-N\beta_{\min}) \to 1$ *as* $N \to \infty$.

Theorem 17 shows that the probability of the screening procedure failing to include all true patterns has an exponential decay with the sample size N. We point out that despite having the nice property of sure screening, the screening procedure does not guarantee consistency in selecting exactly the set $\mathcal{A}_0$ of true patterns, if the number of observed variables per subject J is not large enough. Generally speaking, as N goes large but J does not, the set $\widehat{\mathcal{A}}_{\text{screen}}$ will include many false attribute patterns, although it will contain the true set $\mathcal{A}_0$ with probability tending to one. Therefore the shrinkage estimation approach in Section 4.1 is still essential to performing pattern selection.

In Algorithm 3, we present the proposed screening algorithm with stochastic approximations based on a number of M_{eff} Gibbs samples of $\mathbf{A}$ in the E step. Alternatively, we can also use an even faster screening procedure by just updating the conditional probability of each subject possessing each attribute (i.e., the conditional posterior mean of each $A_{i,k}$) in each E step, conditioning on everything else; we term this alternative procedure the variational screening procedure. As stated before, the screening algorithm is derived based on the log-likelihood of the two-parameter SLAM, but can be applied to a multi-parameter SLAM that covers the two-parameter SLAM as a submodel. After the screening stage, the set of attribute patterns as input to the shrinkage Algorithms 1 or 2 is taken as $\mathcal{A}_{\text{input}} = \widehat{\mathcal{A}}_{\text{screen}}$. Screening drastically lowers down the computational cost of the subsequent shrinkage estimation, and the number of candidate patterns fed to the shrinkage stage is kept at the order of N, even if the original number of possible configurations $2^K \gg N$.

Remark 18 *The screening algorithm can be modified to be more conservative in order to reduce the risk of excluding true patterns. In particular, after each stochastic E step in the screening algorithm, based on the current iterate of* $\mathbf{A}^{\text{ave}}$ *we can obtain a* $N \times K$ *binary matrix with the* (i,k)*th entry being* $I(A^{\text{ave}}_{i,k}) > 1/2$. *The unique row vectors of this intermidiate binary matrix can be viewed as the current candidate latent patterns. To make the screening procedure more conservative, we recommend saving this set of candidate patterns after every* M *stochastic EM iterations (*M *is a positive integer), and take the union of these saved sets in the end of the algorithm to form* $\widehat{\mathcal{A}}_{screen}$ *as the output. We call this*

Algorithm 3: Stochastic Approximation Gibbs Screening

Data: Q, $\mathbf{R}$

Result: Candidate attribute patterns $\widehat{\mathcal{A}}_{\text{screen}}$.

Initialize latent attribute patterns $\mathbf{A} = (A_{i,k})_{N\times K} \in \{0,1\}^{N\times K}$, and $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$.

Set $t = 1$, $\mathbf{A}^{\text{ave}} = \mathbf{0}$, $\mathbf{I}^{\text{ave}} = \mathbf{0}$.

while *not converged* **do**

$\mathbf{A}^{\text{s}} \leftarrow \mathbf{0}$, $\mathbf{I}^{\text{s}} \leftarrow \mathbf{0}$, $M_{\text{eff}} \leftarrow 0$.

for $r \in [M_{\max}]$ **do**

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

Draw $A_{i,k} \sim$ Bernoulli$\Big(\text{logit}^{-1}\Big(\sum_j q_{j,k} \prod_{m\neq k} A_{i,m}^{q_{j,m}} \Big[R_{i,j} \log \frac{\theta_j^+}{\theta_j^-} + (1-R_{i,j}) \log \frac{1-\theta_j^+}{1-\theta_j^-}\Big]\Big)\Big)$.

if $r \geq M_{\max} - M_{\text{eff}}$ **then**

$\mathbf{A}^{\text{s}} \leftarrow \mathbf{A}^{\text{s}} + \mathbf{A}$, $\mathbf{I}^{\text{s}} \leftarrow \mathbf{I}^{\text{s}} + \Big(\prod_k A_{i,k}^{q_{j,k}}\Big)_{N\times J}$.

$\mathbf{A}^{\text{ave}} \leftarrow \frac{1}{t}\mathbf{A}^{\text{s}}/M_{\text{eff}} + \Big(1-\frac{1}{t}\Big)\mathbf{A}^{\text{ave}}$, $\mathbf{I}^{\text{ave}} \leftarrow \frac{1}{t}\mathbf{I}^{\text{s}}/M_{\text{eff}} + \Big(1-\frac{1}{t}\Big)\mathbf{I}^{\text{ave}}$, $t = t+1$.

for $j \in [J]$ **do**

$\theta_j^+ \leftarrow (\sum_i R_{i,j} I_{i,j}^{\text{ave}})/(\sum_i I_{i,j}^{\text{ave}})$, $\theta_j^- \leftarrow (\sum_i R_{i,j}(1-I_{i,j}^{\text{ave}}))(\sum_i (1-I_{i,j}^{\text{ave}}))$.

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

$w_{i,k} \leftarrow I(A_{i,k}^{\text{ave}} > \frac{1}{2})$.

Output: include all the unique row vectors of W in the set $\widehat{\mathcal{A}}_{\text{screen}}$.

strategy "screening enhanced by Gibbs exploration", since it takes advantage of the latent patterns that the Gibbs sampling explores along the stochastic EM iterations.

5. Simulation Studies

We next present simulation results with the two-parameter SLAM and the multi-parameter all-effect SLAM, respectively.

Two-parameter SLAM. Consider the two-parameter SLAM with a $3K \times K$ Q-matrix $Q = (Q_1^\top, Q_2^\top, Q_3^\top)^\top$, where the three submatrices Q_1, Q_2 and Q_3 are specified in (22). We consider three dimensions of possible attribute patterns with $2^K = 2^{10}, 2^{15}$, and 2^{20}, three sample sizes with $N = 150, 500$ and 1000, and two different signal levels with true item parameters: $\{\theta_j^+ = 0.8, \theta_j^- = 0.2;\ j \in [J]\}$, the relatively weak signals; and $\{\theta_j^+ = 0.9, \theta_j^- = 0.1;\ j \in [J]\}$, the relatively strong signals. We randomly generate the set of true attribute patterns $\mathcal{A}_0 \subseteq \{0,1\}^K$ with cardinality $|\mathcal{A}_0| = 10$ and set $p_{\boldsymbol{\alpha}} = 0.1$ for all $\boldsymbol{\alpha} \in \mathcal{A}_0$. In the simulations, for $K = 10$ the $\mathcal{A}_{\text{input}}$ is taken to be $\{0,1\}^K$; while for $K = 15$ and 20, the $\mathcal{A}_{\text{input}}$ is taken to be $\widehat{\mathcal{A}}_{\text{screen}}$, i.e., the set of candidate patterns output by the screening method.

In each scenario we perform 200 independent replications. For shrinkage estimation, we apply the proposed Algorithm 1 "Penalized EM (PEM)" and Algorithm 2 "Fractional Power

Variational EM (FP-VEM)", and also apply the plain EM algorithm with thresholding for comparison. When running PEM we compute a solution path by varying λ in the range of $\lambda \in \{-0.2, -0.4, \ldots, -3.8, -4.0\}$, and select the λ that gives the smallest EBIC. When running FP-VEM we fix $\beta = \lambda + 1 = 0.01$ and compute a solution path by varying Υ in $\{1.0, 0.9, \ldots, 0.4, 0.3\}$ and also select Υ using EBIC. We use the threshold value $\rho_N = 1/(2N)$ for the estimated proportions in the last step for all three shrinkage algorithms to select patterns (other smaller ρ_N values give similar results).

signal strength	2^K	N	1−FDR			TPR		
			EM	Algo. 1	Algo. 2	EM	Algo. 1	Algo. 2
$\theta_j^+ = 0.8$, $\theta_j^- = 0.2$.	2^{10}	150	0.139	0.883	0.896	0.930	0.885	0.895
		500	0.115	0.995	0.992	1.000	1.000	0.999
		1000	0.100	1.000	0.996	1.000	1.000	1.000
	2^{15}	150	0.049	0.523	0.544	0.539	0.530	0.543
		500	0.089	0.924	0.928	0.934	0.930	0.932
		1000	0.078	0.984	0.988	0.991	0.991	0.991
	2^{20}	150	0.019	0.213	0.264	0.270	0.255	0.271
		500	0.019	0.609	0.633	0.636	0.641	0.642
		1000	0.038	0.816	0.848	0.864	0.864	0.863
$\theta_j^+ = 0.9$, $\theta_j^- = 0.1$.	2^{10}	150	0.323	0.909	1.000	1.000	1.000	1.000
		500	0.208	1.000	1.000	1.000	1.000	1.000
		1000	0.167	1.000	1.000	1.000	1.000	1.000
	2^{15}	150	0.317	0.989	0.974	0.993	0.991	0.992
		500	0.220	1.000	0.995	1.000	1.000	1.000
		1000	0.205	1.000	0.994	1.000	1.000	1.000
	2^{20}	150	0.232	0.968	0.941	0.972	0.971	0.970
		500	0.159	1.000	0.999	1.000	1.000	1.000
		1000	0.146	1.000	0.997	1.000	1.000	1.000

Table 2: Pattern selection accuracies for two-parameter SLAM. Tuning parameter $\lambda \in \{-0.2, -0.4, \ldots, -3.8, -4.0\}$ in PEM (Algorithm 1) and $\Upsilon \in \{1.0, 0.9, \ldots, 0.4, 0.3\}$ in FP-VEM (Algorithm 2) are selected based on EBIC.

The simulation results on selection accuracies are presented in Table 2. The "TPR" stands for True Positive Rate, which denotes the proportion of true patterns that are selected. The "1-FDR" stands for "1−False Discovery Rate (FDR)", which denotes the proportion of selected patterns that are true patterns. Table 2 shows the proposed PEM and FP-VEM yield good selection results in various scenarios, while the EM algorithm with direct thresholding at ρ_N suffers from high FDR, i.e., selecting too many non-existing attribute patterns. We would like to point out that the plain VEM as presented in Example 6 is a special case of the proposed FP-VEM, by just taking the fractional power Υ to be $\Upsilon = 1$. So in each simulation run, the result given by VEM is included in the solution path given

by FP-VEM with $\Upsilon \in \{1.0, 0.9, \ldots, 0.4, 0.3\}$, and in the final step EBIC selects the best Υ from the entire solution path. Indeed, in all our simulations about FP-VEM, the result given by $\Upsilon = 1$ is never selected by EBIC, which means the selection result given by plain VEM is never favored over the proposed FP-VEM. We also remark here that the proposed methods are computationally efficient. All the algorithms are implemented in Matlab. In particular, in the case of relatively strong signal with $1 - \theta_j^+ = \theta_j^- = 0.10$, screening and computing an entire solution path for $(2^K, N) = (2^{20}, 1000)$ takes < 2 minutes on average on a laptop with a 2.8 GHz processor, and yields almost perfect pattern selection results, as shown in the last row of Table 2.

We give some discussions on the comparison of the PEM and the FP-VEM algorithms. The estimation accuracies presented in Table 2 generally show the two algorithms have comparable performance on pattern selection. In terms of selecting the tuning parameter, the FP-VEM can be easier to tune because the fractional power Υ is always between 0 and 1, while the PEM algorithm has a negative tuning parameter $\lambda \in (-\infty, 0)$ that can have an arbitrarily large magnitude. Specifically, the scenario of an increasing sparsity corresponds to $\Upsilon \to 0$ and $\lambda \to -\infty$, and when extremal sparsity exists, the FP-VEM needs to choose Υ close to zero with a small magnitude and the PEM needs to choose λ with a large magnitude. Therefore, in such cases the tuning of PEM may take more time, since $\lambda < 0$ needs to be searched over a relatively large interval; an exponential grid search might be of help in this case, while further investigation into how to best specify the grid for searching tuning parameters would be needed. Meanwhile, we find in simulation studies that choosing a small Υ in FP-VEM too close to zero may result in the algorithm to be less stable in some cases. In practice, if the computation time is not a primary concern, we recommend first considering the PEM algorithm for the better stability.

We further conduct a simulation study to investigate how the threshold value ρ_N for the estimated proportions impact the pattern selection results of different methods. In the setting with $1 - \theta_j^+ = \theta_j^- = 0.2$ and $N = 150$ (the same setting as the first line in Table 2), we simulate 200 independent datasets, and apply the proposed PEM (Algorithm 1), FP-VEM (Algorithm 2) and the usual EM algorithm with various thresholds $\rho_N \in \{1/(50N)\} \cup \{i/(2N),\ i = 1, 3, 5, \ldots, 15\}$. Figure 4 plots the average "TPR" and average "1−FDR" versus the threshold values. It can be seen that directly thresholding the MLE of the proportions (corresponding to the thresholding after EM) does not yield good selection results. For a small threshold $\rho_N = 1/(2N)$, the FDR of thresholded EM is quite high. When further decreasing the threshold ρ_N from $1/(2N)$ to $1/(50N)$, the FDR of thresholded EM becomes worse while the proposed methods have stable performance. On the other hand, as the threshold ρ_N increases from $1/(2N)$ to larger values, the TPR of EM quickly decreases. In contrast, the proposed methods PEM and FP-VEM give reasonably good selection results across all the threshold values, and have slightly better performance for smaller thresholds. Even the best selection result given by thresholding EM corresponding to the threshold $\rho_N = 7/(2N)$ is not comparable to those given by the proposed methods.

We next evaluate the performance of the screening procedure. We find that the screening procedure drastically reduces the computational cost in the subsequent shrinkage estimation stage. For instance, in the setting $(N, K) = (150, 15)$ when noise rate is $1 - \theta_j^+ = \theta_j^- = 20\%$, based on 200 runs, the variational screening procedure takes 1.55 seconds on average, and

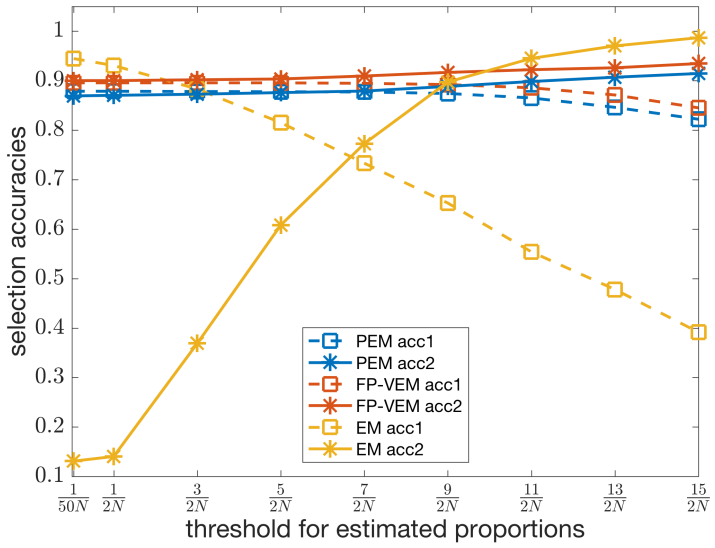
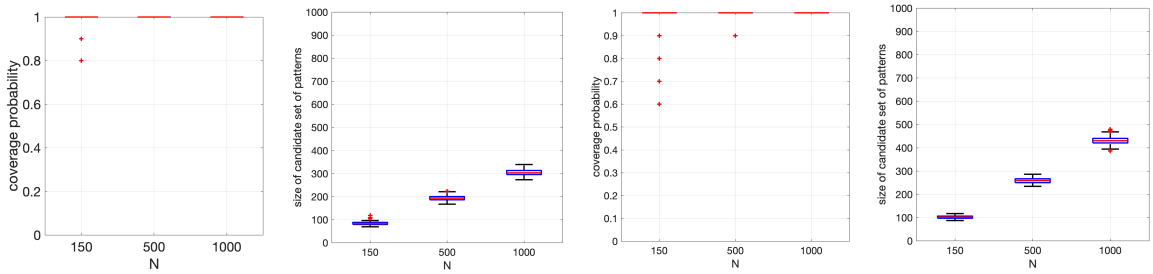


Figure 4: Selection accuracies versus thresholds for the two-parameter SLAM with $1-\theta_j^+ = \theta_j^- = 0.2$ and $N = 150$. For each method, "acc1" denotes the True Positive Rate (TPR), the proportion of true patterns that are selected; and "acc2" denotes "1−False Discovery Rate (FDR)", the proportion of selected patterns that are true.

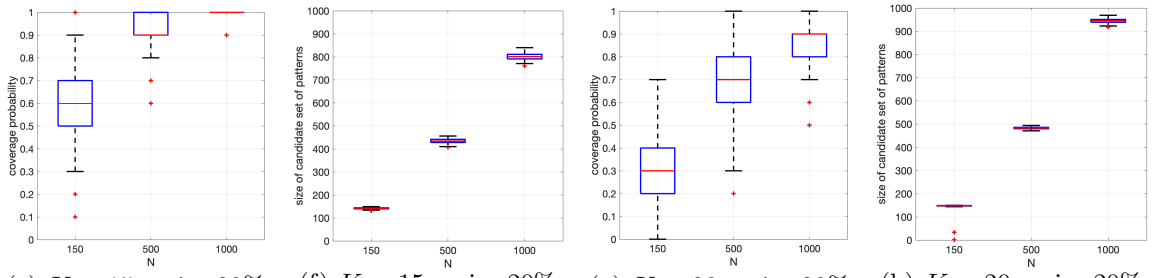
the subsequent PEM algorithm takes 6.42 seconds on average; while if no screening is performed, the PEM algorithm takes 7.96×10^3 seconds on average.

As described earlier, the screening is considered successful if all true patterns are included in the candidate set $\widehat{\mathcal{A}}_{\text{screen}}$. Under each simulation scenario in Table 2 corresponding to $K = 15$ or $K = 20$, we record the coverage probabilities of the true patterns for each of 200 runs, where in each run $\sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} I(\boldsymbol{\alpha} \in \widehat{\mathcal{A}}_{\text{screen}})/|\mathcal{A}_0|$ is recorded as the coverage probability. The boxplots of coverage probabilities under these scenarios are presented in Figure 5(a), (c), (e) and (g). We also record the size of $\widehat{\mathcal{A}}_{\text{screen}}$, i.e., the number of candidate patterns given by the screening procedure in each run, and present their boxplots in Figure 5(b), (d), (f) and (h). The screening procedure generally has good performance. On the other hand, Figure 5(e) and (g) show that for the relatively large noise rate and small sample size, the screening accuracy is not very high.

To improve the performance of screening, we apply the strategy of *screening enhanced by Gibbs exploration* described in Remark 18 and take $M = 3$. That is, along the stochastic EM iterations of the screening algorithm, after every three iterations we add the current set of latent patterns to the candidate set $\widehat{\mathcal{A}}_{\text{screen}}$. The resulting screening accuracies and sizes of $\widehat{\mathcal{A}}_{\text{screen}}$ are presented in Figure 6. Compared to the second row of plots in Figure 5, one can clearly see that the enhancing procedure improves the screening accuracy significantly, while the size of $\mathcal{A}_{\text{screen}}$ also increases but still remains quite manageable. Under the noise rate $1-\theta_j^+ = \theta_j^- = 20\%$, the size of $\mathcal{A}_{\text{screen}}$ is always below N for screening without enhancing, while for screening with enhancing, the size of $\mathcal{A}_{\text{screen}}$ is around $2N$ for $K = 15$ and around $3N$ for $K = 20$. The enhancing by Gibbs exploration would not sacrifice the efficiency of

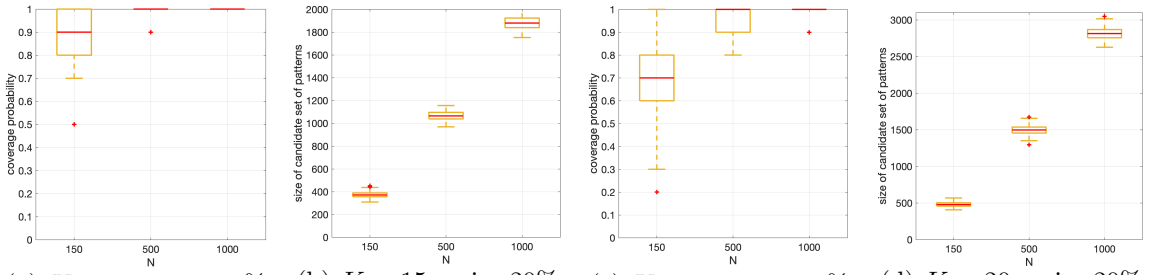


(a) $K = 15$, noise 10%; screening accuracy (b) $K = 15$, noise 10%; size of $\widehat{\mathcal{A}}_{\text{screen}}$ (c) $K = 20$, noise 10%; screening accuracy (d) $K = 20$, noise 10%; size of $\widehat{\mathcal{A}}_{\text{screen}}$



(e) $K = 15$, noise 20%; screening accuracy (f) $K = 15$, noise 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$ (g) $K = 20$, noise 20%; screening accuracy (h) $K = 20$, noise 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$

Figure 5: Screening: plots (a), (c), (e) and (g) are coverage probabilities of the true patterns, from the screening procedure under the two-parameter SLAM; plots (b), (d), (f) and (h) are sizes of $\widehat{\mathcal{A}}_{\text{screen}}$. The "noise" refers to the value of $1 - \theta_j^+ = \theta_j^-$.



(a) $K = 15$, noise 20%; screening accuracy (b) $K = 15$, noise 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$ (c) $K = 20$, noise 20%; screening accuracy (d) $K = 20$, noise 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$

Figure 6: Screening enhanced by Gibbs exploration: screening accuracy and size of $\widehat{\mathcal{A}}_{\text{screen}}$. Noise rate is $1 - \theta_j^+ = \theta_j^- = 20\%$.

the screening procedure itself, though it results in a larger set of $\widehat{\mathcal{A}}_{\text{screen}}$ which incurs higher computational cost in the shrinkage stage. In practice, one should leverage this tradeoff according to the sample size. Specifically, when sample size N is small, choosing a more conservative screening procedure (with a smaller integer M) is recommended, because this would increase the screening accuracy without causing much computational burden for the shrinkage algorithm. With the enhanced screening procedure, in the relatively weak signal case $1 - \theta_j^+ = \theta_j^- = 0.2$ and under $(K, N) = (15, 150)$, the two accuracy measures 1−FDR and TPR for the PEM algorithm, become $(0.850, 0.860)$ (previously it was $(0.523, 0.530)$

in Table 2), and those under the FP-VEM algorithm become $(0.839, 0.853)$ (previously $(0.544, 0.543)$ in Table 2). Under $(K, N) = (20, 150)$, the two accuracy measures for the PEM become $(0.608, 0.648)$ (previously $(0.213, 0.255)$ in Table 2) and those for the FP-VEM become $(0.620, 0.634)$ (previously $(0.264, 0.271)$ in Table 2).

Multi-parameter all-effect SLAM. We next consider the multi-parameter all-effect SLAM introduced in (5) in Example 3 with an identity link function $f(\cdot)$. Let the Q-matrix be in the form $Q = (Q_1^\top, Q_2^\top, Q_2^\top)$ with Q_1 and Q_2 specified in (22). Similar to the two-parameter simulation study, we consider three dimensions of possible attribute patterns with $2^K = 2^{10}, 2^{15}$, and 2^{20}, and three sample sizes with $N = 150, 500$ and 1000. For each item, we set the baseline probability, the positive response probability of the all-zero attribute pattern $\boldsymbol{\alpha} = \mathbf{0}_K$, to 0.2 (i.e., $\theta_{j,\mathbf{0}_K} = 0.2$), and the positive response probability of $\boldsymbol{\alpha} = \mathbf{1}_K$ to 0.8 (i.e., $\theta_{j,\mathbf{1}_K} = 0.8$). And we set all the main effects and interaction effects parameters of the item to be equal (i.e., $\beta_{j,S_1} = \beta_{j,S_2}$ for any $\varnothing \neq S_1, S_2 \subset \mathcal{K}_j$ for the β-coefficients in (5)). We randomly generate the set of true attribute patterns, $\mathcal{A}_0 \subseteq \{0,1\}^K$ with cardinality $|\mathcal{A}_0| = 10$ and set $p_{\boldsymbol{\alpha}} = 0.1$ for all $\boldsymbol{\alpha} \in \mathcal{A}_0$.

2^K	N	1−FDR			TPR		
		EM	Algo. 1	Algo. 2	EM	Algo. 1	Algo. 2
2^{10}	150	0.277	0.983	0.953	0.996	0.980	0.974
	500	0.214	0.988	0.976	1.000	1.000	1.000
	1000	0.193	0.992	0.986	1.000	1.000	1.000
2^{15}	150	0.198	0.900	0.893	0.904	0.902	0.902
	500	0.166	0.999	0.997	1.000	1.000	1.000
	1000	0.134	1.000	0.996	1.000	1.000	1.000
2^{20}	150	0.109	0.723	0.741	0.739	0.734	0.743
	500	0.129	0.980	0.981	0.980	0.982	0.983
	1000	0.104	1.000	0.998	1.000	1.000	1.000

Table 3: Pattern selection accuracies for multi-parameter all-effect SLAM. Tuning parameter $\lambda \in \{-0.2, -0.4, \ldots, -4.0\}$ in PEM (Algorithm 1) and $\Upsilon \in \{1.0, 0.9, \ldots, 0.3\}$ in FP-VEM (Algorithm 2) are selected using EBIC. Signal strengths are $\theta_{j,\mathbf{0}_K} = 0.1,\ \theta_{j,\mathbf{1}_K} = 0.9$.

Similar to the observations in Table 2, Table 3 shows that the proposed methods also have good pattern selection performance for the more complicated multi-parameter all-effect model. The approximate screening algorithm based on the likelihood of the two-parameter submodel is quite effective here for obtaining candidate patterns under the multi-parameter model. And similarly to the two-parameter case, the EM algorithm tends to severely overselects the attribute patterns. Please see Appendix B for additional results on the performance of the screening procedure.

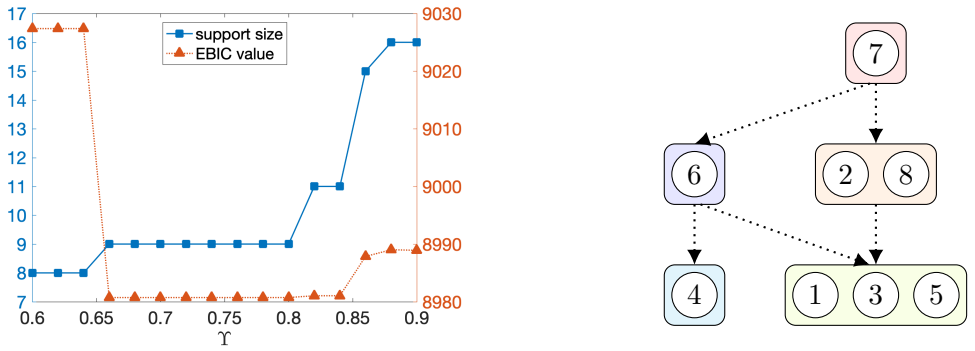
6. Data Analysis

In this section, we apply the proposed methodology to two real world datasets in educational assessments to uncover the knowledge structure of the student population.

Analysis of Fraction Subtraction Data. The fraction subtraction dataset is widely analyzed in the psychometrics literature (de la Torre and Douglas, 2004; DeCarlo, 2011; Henson et al., 2009; de la Torre, 2011). The dataset contains $N = 536$ middle school students' binary (correct or wrong) responses to 20 questions that were designed for the diagnostic assessment of 8 skill attributes related to fraction and subtraction. Table 4 presents the Q-matrix specified in de la Torre and Douglas (2004). The eight attributes are (α_1) Convert a whole number to a fraction; (α_2) Separate a whole number from a fraction; (α_3) Simplify before subtracting; (α_4) Find a common denominator; (α_5) Borrow from whole number part; (α_6) Column borrow to subtract the second numerator from the first; (α_7) Subtract numerators; (α_8) Reduce answers to simplest form.

Item ID	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
1	0	0	0	1	0	1	1	0
2	0	0	0	1	0	0	1	0
3	0	0	0	1	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	0	1	0	0	1	1	1	0
19	1	1	1	0	1	0	1	0
20	0	1	1	0	1	0	1	0

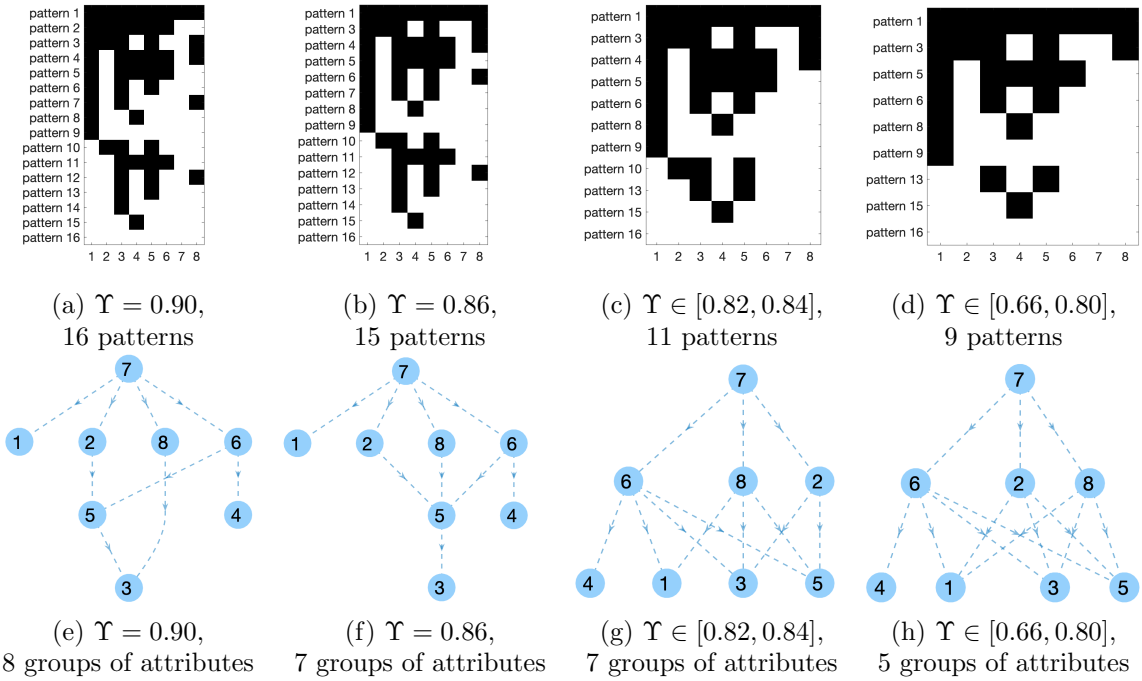
Table 4: Q-matrix, Fraction Subtraction Data



(a) EBIC values and support sizes versus Υ (b) attribute structure selected by EBIC

Figure 7: Results of Fraction Subtraction Data analyzed using two-parameter SLAM.

Many studies in the literature use the two-parameter SLAM to fit the dataset, mostly due to that it is reasonable to assume the required attributes of each item act together to form a "capable" knowledge state and an "incapable" knowledge state. This results in two levels of item parameters for each item. We first use the two-parameter model to



(a) $\Upsilon = 0.90$, 16 patterns (b) $\Upsilon = 0.86$, 15 patterns (c) $\Upsilon \in [0.82, 0.84]$, 11 patterns (d) $\Upsilon \in [0.66, 0.80]$, 9 patterns

(e) $\Upsilon = 0.90$, 8 groups of attributes (f) $\Upsilon = 0.86$, 7 groups of attributes (g) $\Upsilon \in [0.82, 0.84]$, 7 groups of attributes (h) $\Upsilon \in [0.66, 0.80]$, 5 groups of attributes

Figure 8: Fraction Subtraction Data: different sets of estimated patterns (a)–(d) (black for "0" and white for "1") and the corresponding attribute structures (e)–(h) under various Υ's in Algorithm 2. Plot (h) here is equivalent to Figure 7(b).

analyze the data. Given this 20×8 Q-matrix, the number of equivalence classes induced by the Q-matrix $Q_{20\times 8}$ under the two-parameter model is $|\{\vee_{j\in S}\boldsymbol{q}_j : S \subseteq \{1, \ldots, J\}\}| = 58$. We apply Algorithm 2, the FP-VEM algorithm with a sequence of fractional power values $\Upsilon \in \{0.90, 0.89, \cdots, 0.60\}$ and use EBIC to select the tuning parameter Υ while keeping the Dirichlet hyperparameter $\beta = 0.01$. Figure 7(a) plots the EBIC values and the support sizes of $\boldsymbol{p}$, both against the Υ values. It can be seen that $\Upsilon = 0.8$ yields the smallest EBIC value 8.98×10^3, and it is the largest Υ value in the flat window of $[0.66, 0.8]$ that gives 9 equivalence classes of attribute patterns. We also use the multi-parameter all-effect model introduced in Example 3 to fit the dataset. For a range of values of the tuning parameters Υ, the smallest EBIC value is above 1.02×10^4, which is much higher than the smallest EBIC 8.98×10^3 given by the two-parameter model. This also aligns with the results in the literature that the two-parameter model fits the fraction subtraction dataset better than other models (DeCarlo, 2011; de la Torre and Douglas, 2004). Therefore next we only present and discuss the results given by the two-parameter model.

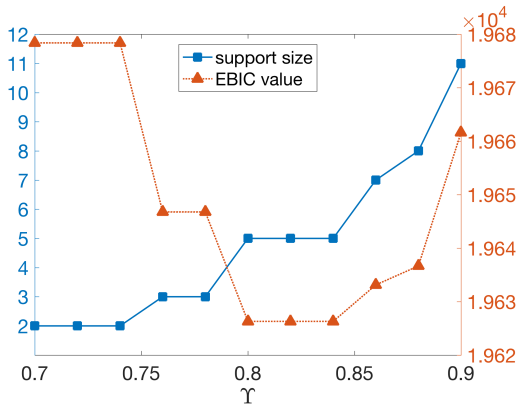
Figure 7(b) plots the attribute structure corresponding to the 9 equivalence classes of attribute patterns selected by EBIC. We obtain this attribute structure using the following procedure. First, we obtain the representatives of these 9 equivalence classes and construct a 9×8 matrix of selected attribute patterns. We denote this 9×8 matrix by $\widehat{\mathbf{A}}$, with each row of $\widehat{\mathbf{A}}$ a 8-dimensional binary vector denoting one selected knowledge state (i.e., attribute pattern). We next examine the partial orders among the columns of this matrix to determine

the relationships among attributes. In particular, if $\widehat{\mathbf{A}}(\cdot, k_1) \succeq \widehat{\mathbf{A}}(\cdot, k_2)$, then attribute k_1 is considered as a prerequisite for attribute k_2. Examining these 9 selected knowledge states, we find that the total number of 8 attributes are separated into 5 groups $G_1 = \{7\}$, $G_2 = \{2, 8\}$, $G_3 = \{6\}$ and $G_4 = \{4\}$ and $G_5 = \{1, 3, 5\}$, such that the attributes in the same group play the same role in clustering the students population into the 9 knowledge states. In particular, based on the observed data, attributes 2 and 8 are equivalent in distinguishing the students population's knowledge states; and so are attributes 1, 3, 5. The estimated prerequisite relationship among these 5 groups is depicted in Figure 7(b). Figure 7(b) implies that attribute (α_7) *Subtract numerators*, is a quite basic skill attribute and serves as prerequisite for all the remaining attributes. This suits the common sense that in the problems about fraction and subtraction, the ability of subtracting integers should be the most basic. Figure 7 also shows that attributes (α_2), (α_6), (α_8) are middle level skills that only has one prerequisite attribute (α_7), and serve as prerequisites for multiple other skills. Finally, the remaining attributes (α_4), (α_1), (α_3) and (α_5) are high level skills in the hierarchical structure. We would like to point out that the directed edges in the attribute hierarchy in Figure 7(b) (and also in the later Figure 9 for the TIMSS dataset) do not necessarily correspond to causal relations between the skill attributes. Instead, the attribute hierarchy results from the learned subset of attribute patterns, and it just reflects the estimated cognitive structure of the students being measured.

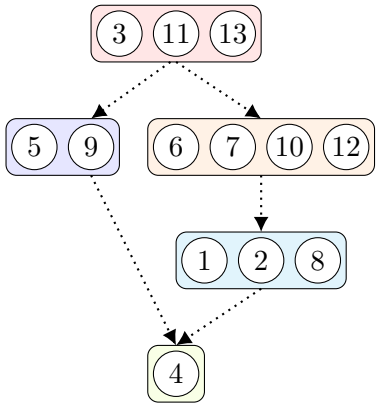
For the Fraction Subtraction data, in addition to the attribute structure chosen by EBIC shown in Figure 7(b), we also present those sets of attribute patterns selected by different Υ's in the solution path. The four sets of patterns and their corresponding attribute structures are presented in Figure 8. As shown in Figure 8(a)–(d), the latent patterns selected by a smaller Υ always form a subset of those patterns selected by a larger Υ. Also, the attribute structures selected by different Υ's share some commonalities. Among the second row of Figure 8, plot (h) is equivalent to the attribute structure in Figure 7(b).

Analysis of TIMSS Data. We also apply the proposed method to the TIMSS 2003 8th grade data. The dataset contains $N = 757$ students' responses to $J = 23$ test items, and the Q-matrix is of size 23×13. Under the two-parameter SLAM, the Q-matrix gives $|\{\vee_{j \in S} \boldsymbol{q}_j : S \subseteq \{1, \ldots, J\}\}| = 1625$ equivalence classes. Figure 9 shows the results of fitting the two-parameter SLAM with $\beta = 0.01$. The fractional power Υ selected by EBIC is 0.84 and the corresponding number of equivalence classes is 5. The smallest EBIC value in Figure 9(a) is 1.96×10^4. We remark that we also fit the general multi-parameter all-effect SLAM to the dataset, while the smallest EBIC given by the multi-parameter model is 7.38×10^4, which is much larger than the best EBIC given by the two-parameter SLAM. So we next focus on the results given by the two-parameter SLAM.

Figure 9(b) plots the attribute structure given by the selected 5 knowledge states. The 13 attributes are separated into five groups $G_1 = \{3, 11, 13\}$, $G_2 = \{5, 9\}$, $G_3 = \{6, 7, 10, 12\}$ and $G_4 = \{1, 2, 8\}$ and $G_5 = \{4\}$, such that the attributes in the same group play the same role in clustering the student population into the five knowledge states. The prerequisite relationships among groups of attributes is also shown in Figure 9(b). Attribute (α_3) *compute fluently with multi-digit numbers and find common factors and multiples*, attribute (α_{11}) *compare two fractions with different numerators and different denominators*, attribute (α_{13}) *use equivalent fraction as a strategy to add and subtract fractions*, are the most basic skills



(a) EBIC values and support sizes versus Υ



(b) attribute structure selected by EBIC

Figure 9: Results of TIMSS 2003 8th Grade Data analyzed using two-parameter SLAM.

in the attribute hierarchy and serve as the prerequisites for all the remaining attributes. Indeed, these three are basic algorithmic operations needed to solve the mathematical problems in the TIMSS test. In addition to the structure selected by EBIC presented in Figure 9(b), other attribute structures corresponding to different $\Upsilon \in [0.7, 0.9]$ are presented in Figure 14 in Appendix B.

Existing works in the literature analyzing the fraction subtraction data and the TIMSS data either make the assumption that all possible configurations of latent attribute patterns exist in the population or pre-specify the attribute structure based on domain experts' judgements (Su et al., 2013). To our knowledge, there has not been a systematic approach to selecting a potentially small set of latent patterns from a high-dimensional space. For the two real datasets, we also find that the EBIC values of the existing EM algorithm are much larger than the proposed method, as indicated in Figures 7 and 9 when Υ close to 1; thus the proposed method provides a better fit of the two datasets.

7. Discussion

In this paper we propose a penalized likelihood method to learn the attribute patterns in the structured latent attribute models, a special family of discrete latent variable models. We allow the number of latent patterns to go to infinity and perform pattern selection by penalizing the proportion parameters of the latent attribute patterns. The theory of pattern selection consistency is established for the proposed regularized MLE. The nice form of the penalty term facilitates the computation. Two algorithms are developed to solve the optimization problem, one being a modification of the EM algorithm, and the other being a variational EM algorithm that results from an alternative Bayesian formulation of the objective function. The simulation study and real data analysis show the proposed methods have good pattern selection performance.

This work assumes the design matrix Q is prespecified and correct. In practice, if there is reason to suspect that the Q-matrix could be misspecified, then one needs to simultaneously

estimate the Q-matrix and learn the attribute patterns from data. Given fixed number of attribute patterns, previous works including Xu and Shang (2018) and Chen et al. (2018) used the likelihood based methods and the Bayesian methods, respectively, to estimate Q. It is also desirable to develop methods to jointly estimate Q and learn attribute patterns with the existence of large number of attributes. We would like to point out that the identifiability results developed in this work (in Section 3) directly apply to this case, and can guarantee both the design matrix Q and the set of significant attribute patterns are learnable from data.

The learnability theory developed in this paper guarantees one can reliably learn a SLAM with an arbitrary set of attribute patterns from data. As mentioned earlier, SLAMs can be expressed as higher-order probability tensors with special structures. Also, SLAMs share similarities with the restricted Boltzmann machines and the deep Boltzmann machines in terms of the bipartite graph structure among the latent and observed multivariate binary variables. Current techniques for proving identifiability of SLAMs could be adapted to develop theory for uniqueness of structured tensor decompositions and learnability of some more complicated latent variable models. We leave these directions for future study.

Acknowledgement

The authors are grateful to the editors and two reviewers for their helpful and constructive comments. This research was partially supported by National Science Foundation grants SES1659328 and DMS-1712717, and Institute of Education Sciences grant R305D160010. This research was also supported in part through the High Performance Computing clusters provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

Appendix

Appendix A: Technical Proofs

We introduce a useful notation, the T-matrix, before proving the identifiability theory. We consider a marginal probability matrix $T(\Gamma^{\mathcal{A}}, \boldsymbol{\Theta}^{\mathcal{A}})$ of size $2^J \times |\mathcal{A}|$ as follows. When it causes no confusion, we also write $T(\Gamma^{\mathcal{A}}, \boldsymbol{\Theta}^{\mathcal{A}})$ simply as $T(\Gamma, \boldsymbol{\Theta})$. Rows of $T(\Gamma, \boldsymbol{\Theta})$ are indexed by the 2^J possible response patterns $\boldsymbol{r} = (r_1, \ldots, r_J)^\top \in \{0,1\}^J$ and columns of $T(\Gamma, \boldsymbol{\Theta})$ are indexed by latent attribute patterns $\boldsymbol{\alpha} \in \mathcal{A}$, while the $(\boldsymbol{r}, \boldsymbol{\alpha})$th entry of $T(\Gamma, \boldsymbol{\Theta})$, denoted by $T_{\boldsymbol{r},\boldsymbol{\alpha}}(\Gamma, \boldsymbol{\Theta})$, represents the marginal probability that subjects in latent class $\boldsymbol{\alpha}$ provide positive responses to the set of items $\{j : r_j = 1\}$, namely $T_{\boldsymbol{r},\boldsymbol{\alpha}}(\Gamma, \boldsymbol{\Theta}) = P(\boldsymbol{R} \succeq \boldsymbol{r} \mid \boldsymbol{\Theta}, \boldsymbol{\alpha}) = \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{r_j}$. Denote the $\boldsymbol{\alpha}$th column vector and the $\boldsymbol{r}$th row vector of the T-matrix by $T_{\boldsymbol{\cdot},\boldsymbol{\alpha}}(\Gamma, \boldsymbol{\Theta})$ and $T_{\boldsymbol{r},\boldsymbol{\cdot}}(\Gamma, \boldsymbol{\Theta})$ respectively. Let $\boldsymbol{e}_j$ denote the J-dimensional unit vector with the jth element being one and all the other elements being zero, then any response pattern $\boldsymbol{r}$ can be written as a sum of some $\boldsymbol{e}$-vectors, namely $\boldsymbol{r} = \sum_{j:r_j=1} \boldsymbol{e}_j$. The $\boldsymbol{r}$th element of the 2^J-dimensional vector $T(\Gamma, \boldsymbol{\Theta})\boldsymbol{p}$ is $\{T(\Gamma, \boldsymbol{\Theta})\boldsymbol{p}\}_{\boldsymbol{r}} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\Gamma, \boldsymbol{\Theta})\boldsymbol{p} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} T_{\boldsymbol{r},\boldsymbol{\alpha}}(\Gamma, \boldsymbol{\Theta}) p_{\boldsymbol{\alpha}} = P(\boldsymbol{R} \succeq \boldsymbol{r} \mid \Gamma, \boldsymbol{\Theta})$. The T-matrix have some nice algebraic properties that will be useful in later proofs. We state them in the following lemma, the proof of which is similar to that of Proposition 3 in Xu (2017) and hence is omitted.

Lemma 19 *Under a SLAM with constraint matrix* Γ, $(\Gamma, \boldsymbol{\Theta}, \boldsymbol{p})$ *are jointly identifiable if and only if for any* $(\Gamma, \boldsymbol{\Theta}, \boldsymbol{p})$ *and* $(\bar{\Gamma}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$,

$$T(\Gamma, \boldsymbol{\Theta})\boldsymbol{p} = T(\bar{\Gamma}, \bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}} \tag{25}$$

implies $(\Gamma, \boldsymbol{\Theta}, \boldsymbol{p}) = (\bar{\Gamma}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$. *For any* $\boldsymbol{\theta}^* = (\theta_1, \ldots, \theta_J)^\top \in \mathbb{R}^J$, *there exists an invertible matrix* $D(\boldsymbol{\theta}^*)$ *only depending on* $\boldsymbol{\theta}^*$, *such that*

$$T(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top) = D(\boldsymbol{\theta}^*) T(\Gamma, \boldsymbol{\Theta}),$$

where $\mathbf{1}^\top$ *denotes an all-one vector and* $\boldsymbol{\theta}^* \mathbf{1}^\top$ *is a matrix of same size as* $\boldsymbol{\Theta}$.

Proof of Theorem 2 and Corollary 3. We aim to prove that if $\Gamma := \Gamma^{\mathcal{A}_0}$ of size $J \times L_0$ ($L_0 = |\mathcal{A}_0|$) satisfies Conditions A and B, then for any binary matrix $\bar{\Gamma}$ also of size $J \times L_0$, which can be viewed as a constraint matrix imposing restrictions on the parameter space of the $J \times L_0$ item parameter matrix $\bar{\boldsymbol{\Theta}}$, and for any L_0-dimensional vector $\bar{\boldsymbol{p}} := (\bar{p}_1, \ldots, \bar{p}_{L_0})$ with $\bar{p}_l \geq 0$ and $\sum_{l=1}^{L_0} \bar{p}_l = 1$, which can be viewed as a population proportion vector giving proportions of the L_0 latent classes, if

$$T(\Gamma, \boldsymbol{\Theta})\boldsymbol{p} = T(\bar{\Gamma}, \bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}} \tag{26}$$

holds, then $(\Gamma, \boldsymbol{\Theta}, \boldsymbol{p}) = (\bar{\Gamma}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ up to a label swapping of the latent classes. If this is proved, then combining Condition C that any column vector of $\Gamma^{\mathcal{A}_0}$ is different from any column vector of $\Gamma^{\mathcal{A}_0^c}$, we would have the conclusion that the identified $\Gamma^{\mathcal{A}_0}$ uniquely maps to the true set of attribute patterns $\mathcal{A}_0$.

We add a remark here that given (26), the columns of the $\bar{\Gamma}$ do not necessarily have the interpretation of representing some K-dimensional binary attribute patterns; instead, these columns just correspond to L_0 latent classes. And after we obtain $(\Gamma, \boldsymbol{\Theta}, \boldsymbol{p}) = (\bar{\Gamma}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ up to a label swapping, we would have the conclusion that $\bar{\Gamma}$ equals Γ up to column permutation; Then with Condition C, the $\bar{\Gamma}$ would have the interpretation of being the constraint matrix for the attribute patterns in $\mathcal{A}_0$. Because of this, in the following proof, we sometimes will also ignore the interpretation of the columns of the true $\Gamma^{\mathcal{A}_0}$, and simply denote the columns of it by the column index integer l, i.e., $\Gamma^{\mathcal{A}_0}$ has columns $\Gamma^{\mathcal{A}_0}_{\bullet,l}$ for $l = 1, \ldots, L_0$.

For notational simplicity, we denote $\Gamma^{(S_i, \mathcal{A}_0)}$ by Γ^i for $i = 1, 2$ and $\Gamma^{((S_1 \cup S_2)^c, \mathcal{A}_0)}$ by Γ^3. We also denote item parameter matrix $\boldsymbol{\Theta}^{(S_1, \mathcal{A}_0)}$, $\boldsymbol{\Theta}^{(S_2, \mathcal{A}_0)}$ and $\boldsymbol{\Theta}^{((S_1 \cup S_2)^c, \mathcal{A}_0)}$ by $\boldsymbol{\Theta}^1$, $\boldsymbol{\Theta}^2$ and $\boldsymbol{\Theta}^3$, respectively. So each $\boldsymbol{\Theta}^i$ has the same size as Γ^i and respects the constraints specified by Γ^i. Without loss of generality, suppose Γ takes the form $\Gamma^\top = [(\Gamma^1)^\top, (\Gamma^2)^\top, (\Gamma^3)^\top]$, where each Γ^i is of size $J_i \times L_0$ and $J_1 + J_2 + J_3 = J$. For any item j, by the definition of SLAM we have all those $\boldsymbol{\alpha}$ with $\Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}} = 1$ have the same highest value of item parameter. For simplicity, we denote this value of the item parameter by $\theta_{j,H}$, where "H" stands for "highest" level item parameter for item j.

We first show $T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)$ and $T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)$ both have full column rank L_0, and that $\bar{p}_l > 0$ for all $l \in \{1, \ldots, L_0\}$. By Proposition 3 in Gu and Xu (2019b), Condition A ensures that $T(\Gamma^1, \boldsymbol{\Theta}^1)$ of size 2^{J_1} and $T(\Gamma^2, \boldsymbol{\Theta}^2)$ of size 2^{J_2} both have full column rank L_0, since Γ^1 and Γ^2 are both separable. Moreover, in the proof of that conclusion, an invertible square matrix W_1 of size $2^{J_1} \times 2^{J_1}$ as well as L_0 response patterns $\boldsymbol{r}_1, \ \ldots, \ \boldsymbol{r}_{L_0} \in \{0, 1\}^L$ were constructed

such that the row vectors in the transformed $W_1 \cdot T(\Gamma^1, \boldsymbol{\Theta}^1)$, which are indexed by the chosen $\boldsymbol{r}_1, \ \ldots, \ \boldsymbol{r}_{L_0}$, form a $L_0 \times L_0$ lower triangular matrix with nonzero diagonal elements. In other words, in the $2^{J_1} \times L_0$ rectangular matrix $W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)$, there is a $L_0 \times L_0$ submatrix that is lower triangular and full-rank. For notational simplicity, we denote this submatrix by $\{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}$. Similarly, there exists W_2 and $\boldsymbol{r}'_1, \ \ldots, \ \boldsymbol{r}'_{L_0} \in \{0,1\}^{L_0}$ such that there is a $L \times L$ full-rank submatrix of $W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)$ with rows indexed by $\boldsymbol{r}'_1, \ \ldots, \ \boldsymbol{r}'_{L_0}$, which we denote by $\{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}$.

Based on the above constructions, there exist two invertible square matrices U_1 and U_2 such that $U_1 \cdot \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}} = I_{L_0}$ and $U_2 \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}} = I_{L_0}$. Denote the C row vectors of U_1 by $\{\boldsymbol{u}_l^\top, \ l \in [L_0]\}$, then we have that for any $l \in [L_0]$,

$$\boldsymbol{u}_l^\top \cdot \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}} = (\boldsymbol{0}, \underbrace{1}_{\text{column } l}, \boldsymbol{0}). \tag{27}$$

Next we prove by contradiction that $\{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}$ and $\{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}$ must also be invertible. We focus on $\{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}$ and conclusion for the other is the same. If $\{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}$ does not have full rank, then $U_2 \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}$ also does not have full rank, so there exists a nonzero vector $\boldsymbol{x} = (x_1, \ldots, x_{L_0})$ such that

$$\boldsymbol{x}^\top \cdot U_2 \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}} = \boldsymbol{0}.$$

Note that $\boldsymbol{x}^\top \cdot U_2 \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}} = \boldsymbol{x}$ from the previous construction of W_2. Since $\boldsymbol{x} \neq \boldsymbol{0}$, suppose without loss of generality that $x_l \neq 0$ for some l, then we have

$$\begin{aligned}
&[\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{x}^\top \cdot U_2 \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \boldsymbol{p} = x_l p_l \neq 0, \\
&[\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{x}^\top \cdot U_2 \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \bar{\boldsymbol{p}} = 0,
\end{aligned}$$

which contradicts (26). Here $\boldsymbol{a} \odot \boldsymbol{b}$ denotes the elementwise product of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ of the same length. Therefore $\{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}$ must have full rank C, and so as $\{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}$.

Based on the above conclusion, we next show that $\bar{p}_l > 0$ for any $l \in [L_0]$. Suppose this is not true and $\bar{p}_l = 0$ for some l, then there exists a nonzero vector $\boldsymbol{y} = (y_1, \ldots, y_{L_0})^\top$ such that

$$\boldsymbol{y}^\top \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}} = (\boldsymbol{0}, \underbrace{1}_{\text{column } l}, \boldsymbol{0}).$$

Since $\{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}$ has full rank and $\boldsymbol{y} \neq \boldsymbol{0}$, we have $\boldsymbol{y}^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}} \neq \boldsymbol{0}$. Without loss of generality, suppose the $l^\star$-th column of this product vector is nonzero and denote the nonzero value by $b_{l^\star}$, then using the $\boldsymbol{u}$-vectors constructed previously in (27), we have

$$\begin{aligned}
&[\boldsymbol{u}_{\boldsymbol{\alpha}^\star}^\top \cdot \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{y}^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \boldsymbol{p} = b_{l^\star} p_{l^\star} \neq 0, \\
&[\boldsymbol{u}_{\boldsymbol{\alpha}^\star}^\top \cdot \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{y}^\top \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \bar{\boldsymbol{p}} = 0,
\end{aligned}$$

which contradicts (26). This shows that $\bar{p}_l > 0$ must hold for all $l \in [L_0]$.

We next show that for any $j \in (S_1 \cup S_2)^c$ and any $l \in \{1, \ldots, L_0\}$, $\theta_{j,l} = \theta_{j,\sigma(l)}$, where $\sigma(\cdot)$ is a permutation map from $\{1, \ldots, L_0\}$ to $\{1, \ldots, L_0\}$. There must exist a permutation map $\sigma : \{1, \ldots, L\} \to \{1, \ldots, L\}$ such that for each $l \in [L_0]$,

$$\bar{f}_{\sigma(l)} := [\boldsymbol{u}_l^\top \cdot \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}]_{\sigma(l)} \neq 0.$$

This is because otherwise there would exist $l \in [L_0]$ such that $\{U_1 \cdot T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\cdot,l}$ equals the zero vector, which contradicts the fact that both U_1 and $\{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}$ are invertible matrices. Given the permutation σ, there exists a $L_0 \times L_0$ invertible matrix V with row vectors denoted by $\{\boldsymbol{v}_l,\ l \in [L_0]\}$ such that for each $\boldsymbol{\alpha} \in \mathcal{A}$,

$$\boldsymbol{v}_l^\top \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}} = (\boldsymbol{0}, \underbrace{1}_{\text{column } \sigma(l)}, \boldsymbol{0}). \tag{28}$$

Then we have

$$[\boldsymbol{u}_l^\top \cdot \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \boldsymbol{p} = f_l p_l, \tag{29}$$

$$[\boldsymbol{u}_l^\top \cdot \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \bar{\boldsymbol{p}} = \bar{f}_{\sigma(l)} \bar{p}_{\sigma(l)} \neq 0, \tag{30}$$

where $f_l = [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}]_l$. Now we have $f_l p_l = \bar{f}_{\sigma(l)} \bar{p}_{\sigma(l)} \neq 0$. Next further consider an arbitrary item $j \in (S_1 \cup S_2)^c$. Equation (26) indicates that

$$\theta_{j,l} = \frac{T_{\boldsymbol{e}_j,\cdot}(\Gamma, \boldsymbol{\Theta}) \odot (29)}{(29)} = \frac{T_{\boldsymbol{e}_j,\cdot}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}}) \odot (30)}{(30)} = \bar{\theta}_{j,\sigma(l)}.$$

We next show that for any $j \in S_1 \cup S_2$ and any $l \in \{1, \ldots, L_0\}$ such that $\Gamma_{j,l} = 1$, $\theta_{j,l} = \theta_{j,H} = \bar{\theta}_{j,\sigma(l)} = \bar{\theta}_{j,H}$. We introduce a lemma before proceeding with the proof.

Lemma 20 *Under the assumptions of Theorem 2, the vectors* $\{\boldsymbol{v}_l,\ l \in \mathcal{A}_0\}$ *constructed in* (28) *satisfy that*

$$\{\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}\}_{l'} = 0, \quad \forall \boldsymbol{\alpha}_{l'} \precneqq_{S_1} \boldsymbol{\alpha}_l \text{ under } \Gamma^{\mathcal{A}_0}.$$

Proof of Lemma 20 If $\{\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}\}_{l'} = z_{l'} \neq 0$, then similar to (29) and (30) we have

$$[\boldsymbol{u}_{l'}^\top \cdot \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \boldsymbol{p} = z_{l'} p_{l'} \neq 0,$$

$$[\boldsymbol{u}_{l'}^\top \cdot \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \bar{\boldsymbol{p}} = \bar{f}_{\sigma(l)} \bar{p}_{\sigma(l)},$$

and further we have $\theta_{j,l'} = \bar{\theta}_{j,\sigma(l)} = \theta_{j,l}$ for $j \in (S_1 \cup S_2)^c$, which contradicts condition (C2). This completes the proof of the lemma. □

We proceed with the proof. For any $l \in [L_0]$, define $\boldsymbol{\theta}^* = \sum_{h \in S_1 : \Gamma_{h,l} = 1} \theta_{h,1} \boldsymbol{e}_h$. With $\boldsymbol{\theta}^*$, the row vector corresponding to $\boldsymbol{r}^* = \sum_{h \in S_1 : \Gamma_{h,l} = 0} \boldsymbol{e}_h$ in the transformed T-matrix satisfies that

$$b_l := T_{\boldsymbol{r}^*,l}(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \boldsymbol{1}^\top) \neq 0; \tag{31}$$

$$T_{\boldsymbol{r}^*,l'}(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \mathbf{1}^\top) = 0, \quad \forall \boldsymbol{\alpha}_{l'} \not\preceq_{S_1} \boldsymbol{\alpha}_l \text{ under } \Gamma^{\mathcal{A}_0}.$$

The proof of Step 2 as well as Lemma 20 ensures

$$\begin{aligned} f_l =& [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}]_l \neq 0; \\ & [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}]_{l'} = 0, \quad \forall \boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l \text{ under } \Gamma^{\mathcal{A}_0}. \end{aligned} \tag{32}$$

Consider any $j \in S_1 \cup S_2$ such that $\Gamma_{j,l} = 1$, then obviously $\boldsymbol{e}_j$ is not included in the sum in the previously defined response pattern $\boldsymbol{r}^*$, because $\boldsymbol{r}^*$ only contains those items that $\boldsymbol{\alpha}_l$ is not capable of, i.e., those j s.t. $\Gamma^{\mathcal{A}_0}_{j,l} = 0$. The above two equations (31) and (32) indicate

$$T_{\boldsymbol{r}^*,\cdot}(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}] = \Big(\mathbf{0}^\top, \underbrace{b_l \cdot f_l}_{\text{column } l}, \mathbf{0}^\top\Big), \tag{33}$$

$$T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}] = \Big(\mathbf{0}^\top, \underbrace{\theta_{j,H} \cdot b_l \cdot f_l}_{\text{column } l}, \mathbf{0}^\top\Big). \tag{34}$$

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

$$T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{v}_l^\top \cdot T(\bar{\boldsymbol{\Theta}}^2)\} = \Big(\mathbf{0}^\top, \underbrace{\prod_{h \in S_1 : \Gamma_{h,l} = 0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H})}_{\text{column } \sigma(l)}, \mathbf{0}^\top\Big), \tag{35}$$

$$T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{v}_l^\top \cdot T(\bar{\boldsymbol{\Theta}}^2)\} = \Big(\mathbf{0}^\top, \underbrace{\bar{\theta}_{j,H} \cdot \prod_{h \in S_1 : \Gamma_{h,l} = 0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H})}_{\text{column } \sigma(l)}, \mathbf{0}^\top\Big). \tag{36}$$

Equation (26) implies $(33) \cdot \boldsymbol{p} = (35) \cdot \bar{\boldsymbol{p}}$. By (26), the above four equations give that

$$\theta_{j,H} = \theta_{j,l} = \frac{(34) \cdot \boldsymbol{p}}{(33) \cdot \boldsymbol{p}} = \frac{(36) \cdot \bar{\boldsymbol{p}}}{(35) \cdot \bar{\boldsymbol{p}}} = \bar{\theta}_{j,\sigma(l)} = \bar{\theta}_{j,H}, \quad \forall j \in S_2.$$

Note that the above equality $\theta_{j,H} = \bar{\theta}_{j,H}$ holds for any l and any item j such that $\Gamma_{j,l} = 1$. Therefore we have shown $\theta_{j,H} = \bar{\theta}_{j,H}$ holds for any $j \in S_1 \cup S_2$.

We next show that for any $j \in S_1 \cup S_2$ and any $l \in \{1, \ldots, L_0\}$ such that $\Gamma_{j,l} = 0$, $\theta_{j,l} = \bar{\theta}_{j,\sigma(l)}$, and show $p_l = \bar{p}_{\sigma(l)}$ for any $l \in \{1, \ldots, L_0\}$. We use an induction method to show for any $l \in [L_0]$,

$$\forall j \in S_1 \cup S_2, \quad \theta_{j,l} = \bar{\theta}_{j,\sigma(l)}, \quad p_l = \bar{p}_{\sigma(l)}. \tag{37}$$

We first introduce the *lexicographic order* between two binary vectors of the same length. For two vectors $\boldsymbol{a} = (a_1, \ldots, a_L)$ and $\boldsymbol{b} = (b_1, \ldots, b_L)$, we say $\boldsymbol{a}$ has smaller lexicographic order than $\boldsymbol{b}$ and denote by $\boldsymbol{a} \prec_{\text{lex}} \boldsymbol{b}$, if either $a_1 < b_1$, or $a_l < b_l$ for some integer $l \leq L$ and $a_m = b_m$ for all $m = 1, \ldots, l-1$. By Condition A, $\Gamma^{(S_i, \mathcal{A}_0)}$ has distinct column vectors

for $i = 1, 2$, so without loss of generality, we can assume the columns of it are sorted in an increasing lexicographic order, i.e.,

$$\Gamma^{(S_1,\mathcal{A}_0)}_{\bullet,1} \prec_{\text{lex}} \cdots \prec_{\text{lex}} \Gamma^{(S_1,\mathcal{A}_0)}_{\bullet,L_0}. \tag{38}$$

Firstly, we prove (37) hold for $l = 1$, where from (38) we have $\Gamma^{(S_1,\mathcal{A}_0)}_{\bullet,1}$ has the smallest lexicographical order among the column vectors of $\Gamma^{(S_1,\mathcal{A}_0)}$. We claim that $\Gamma^{(S_2,\mathcal{A}_0)}_{\bullet,1}$ has the smallest lexicographical order among the column vectors of $\Gamma^{(S_2,\mathcal{A}_0)}$, because otherwise " $\succeq_{S_1}$=$\succeq_{S_2}$" under $\mathcal{A}_0$ will not hold. For $l = 1$ we define

$$\boldsymbol{\theta}^* = \sum_{h\in S_1:\Gamma_{h,1}=0} \theta_{h,H}\boldsymbol{e}_h,$$

and consider the row vector of the transformed T-matrix $T(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ corresponding to $\boldsymbol{r} = \sum_{h\in S_1:\Gamma_{h,1}=0} \boldsymbol{e}_h$ has only one potentially nonzero element in the first column, i.e.,

$$T_{\boldsymbol{r},\bullet}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(\prod_{h\in S_1:\Gamma_{h,1}=0} (\theta_{h,1} - \theta_{h,H}),\ 0, \ldots, 0 \Big)$$

Then similarly for parameters $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ we have

$$T_{\boldsymbol{r},\bullet}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(0, \ldots, 0, \underbrace{\prod_{h\in S_1:\Gamma_{h,1}=0} (\bar{\theta}_{h,\sigma(1)} - \theta_{h,H})}_{\text{column } \sigma(1)},\ 0, \ldots, 0 \Big)$$

and

$$\prod_{h\in S_1:\Gamma_{h,1}=0} (\theta_{h,1} - \bar{\theta}_{h,H}) \neq 0, \qquad \prod_{h\in S_1:\Gamma_{h,1}=0} (\bar{\theta}_{h,1} - \theta_{h,H}) \neq 0.$$

Now consider $\theta_{j,1}$ for any $j \in S_2$ and $\Gamma_{j,1} = 0$. The row vectors of $T(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ and $T(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ corresponding to the response pattern $\boldsymbol{r} + \boldsymbol{e}_j$ are

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(\prod_{h\in S_1:\Gamma_{h,1}=0} (\theta_{h,1} - \theta_{h,H}) \cdot \theta_{j,1},\ 0, \ldots, 0 \Big), \tag{39}$$

and

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(0, \ldots, 0, \underbrace{\prod_{h\in S_1:\Gamma_{h,1}=0} (\bar{\theta}_{h,\sigma(1)} - \theta_{h,H}) \cdot \bar{\theta}_{j,\sigma(1)}}_{\text{column } \sigma(1)},\ 0, \ldots, 0 \Big), \tag{40}$$

respectively. The only potentially nonzero term in the first column of (39) is indeed nonzero, because we have $\theta_{h,1} < \theta_{h,H}$ for $h \in S_1$, $\Gamma_{h,1} = 0$. Now Equation (25) implies that

$$\theta_{j,1} = \frac{T_{\boldsymbol{r}+\boldsymbol{e}_j,\bullet}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p}}{T_{\boldsymbol{r},\bullet}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p}} = \frac{T_{\boldsymbol{r}+\boldsymbol{e}_j,\bullet}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}}}{T_{\boldsymbol{r},\bullet}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}}} = \bar{\theta}_{j,\sigma(1)},$$

for any $j \in S_2$ and $\Gamma_{j,1} = 0$. Similarly we can obtain $\theta_{j,1} = \bar{\theta}_{j,\sigma(1)}$ for any $j \in S_1$ and $\Gamma_{j,\sigma(1)} = 0$.

After obtaining these $\bar{\theta}_{j,\sigma(1)} = \theta_{j,1}$ for $j \in (S_1 \cup S_2)$ and $\Gamma_{j,1} = 0$, the previous equations (39) and (40) just become the following,

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(\prod_{h \in S_1 : \Gamma_{h,1}=0} (\theta_{h,1} - \theta_{h,H}) \cdot \theta_{j,1},\ 0, \ldots, 0 \Big), \tag{41}$$

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(0, \ldots, 0, \underbrace{\prod_{h \in S_1 : \Gamma_{h,1}=0} (\theta_{h,\sigma(1)} - \theta_{h,H}) \cdot \theta_{j,\sigma(1)}}_{\text{column } \sigma(1)},\ 0, \ldots, 0 \Big). \tag{42}$$

Therefore $(41) \cdot \boldsymbol{p} = (42) \cdot \bar{\boldsymbol{p}}$ just gives $p_1 = \bar{p}_{\sigma(1)}$.

Now as the inductive hypothesis, we assume for an $l \in [L_0]$,

$$\forall \boldsymbol{\alpha}_{l'} \text{ s.t. } \boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l, \quad \forall j \in S_1 \cup S_2, \quad \theta_{j,l'} = \bar{\theta}_{j,\sigma(l')}, \quad p_{l'} = \bar{p}_{\sigma(l')}.$$

Recall that $\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l$ if and only if $\boldsymbol{\alpha}_{l'} \preceq_{S_2} \boldsymbol{\alpha}_l$ under $\mathcal{A}_0$. Define $\boldsymbol{\theta}^*$ as

$$\boldsymbol{\theta}^* = \sum_{h \in S_1 : \Gamma_{h,l}=0} \theta_{h,H} \boldsymbol{e}_h + \sum_{h \in S_1 : \Gamma_{h,l}=1} \theta_{h,l} \boldsymbol{e}_h,$$

then for $\boldsymbol{r}^* := \sum_{h \in S_1} \boldsymbol{e}_h$ we have

$$\begin{aligned} T_{\boldsymbol{r}^*,\cdot}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p} = &\sum_{\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l} t_{\boldsymbol{r}^*,l'} \cdot p_{l'} \\ &+ \prod_{h \in S_1 : \Gamma_{h,l}=0} (\theta_{h,l} - \theta_{h,H}) \prod_{h \in S_1 : \Gamma_{h,l}=1} (\theta_{h,l} - \theta_{h,1}) \cdot p_l, \end{aligned} \tag{43}$$

$$\begin{aligned} T_{\boldsymbol{r}^*,\cdot}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}} = &\sum_{\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l} \bar{t}_{\boldsymbol{r}^*,\sigma(l')} \cdot \bar{p}_{\sigma(l')} \\ &+ \prod_{h \in S_1 : \Gamma_{h,l}=0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H}) \prod_{h \in S_1 : \Gamma_{h,l}=1} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,1}) \cdot \bar{p}_{\sigma(l)}, \end{aligned} \tag{44}$$

where the notations $t_{\boldsymbol{r}^*,l'}$ and $\bar{t}_{\boldsymbol{r}^*,l'}$ are defined as

$$\begin{aligned} t_{\boldsymbol{r}^*,l'} &= \prod_{h \in S_1 : \Gamma_{h,l}=0} (\theta_{h,l'} - \theta_{h,H}) \prod_{h \in S_1 : \Gamma_{h,l}=1} (\theta_{h,l} - \theta_{h,1}), \\ \bar{t}_{\boldsymbol{r}^*,l'} &= \prod_{h \in S_1 : \Gamma_{h,l}=0} (\bar{\theta}_{h,\sigma(l')} - \theta_{h,H}) \prod_{h \in S_1 : \Gamma_{h,l}=1} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,1}). \end{aligned}$$

Note that by induction assumption we have $\theta_{h,l'} = \bar{\theta}_{h,\sigma(l')}$ for any l' such that $\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l$ under $\mathcal{A}_0$. This implies $t_{\boldsymbol{r}^*,l'} = \bar{t}_{\boldsymbol{r}^*,\sigma(l')}$ and further implies

$$\sum_{\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l} t_{\boldsymbol{r}^*,l'} \cdot p_{l'} = \sum_{\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l} \bar{t}_{\boldsymbol{r}^*,\sigma(l')} \cdot \bar{p}_{\sigma(l')}.$$

So (43) = (44) gives

$$\prod_{h\in S_1:\Gamma_{h,l}=0}(\theta_{h,l}-\theta_{h,H})\prod_{h\in S_1:\Gamma_{h,l}=1}(\theta_{h,l}-\theta_{h,1})\cdot p_l \tag{45}$$
$$=\prod_{h\in S_1:\Gamma_{h,l}=0}(\bar\theta_{h,\sigma(l)}-\theta_{h,H})\prod_{h\in S_1:\Gamma_{h,l}=1}(\bar\theta_{h,\sigma(l)}-\theta_{h,1})\cdot \bar p_{\sigma(l)},$$

and the two terms on both hand sides of the above equation are nonzero. Now consider any $j\notin S_1$ and similarly $T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\Gamma,\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p}=T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar\Gamma,\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}}$ yields

$$\theta_{j,l}\cdot\prod_{h\in S_1:\Gamma_{h,l}=0}(\theta_{h,l}-\theta_{h,H})\prod_{h\in S_1:\Gamma_{h,l}=1}(\theta_{h,\boldsymbol{\alpha}}-\theta_{h,1})\cdot p_l \tag{46}$$
$$=\bar\theta_{j,\sigma(l)}\cdot\prod_{h\in S_1:\Gamma_{h,l}=0}(\bar\theta_{h,\sigma(l)}-\theta_{h,H})\prod_{h\in S_1:\Gamma_{h,l}=1}(\bar\theta_{h,\sigma(l)}-\theta_{h,1})\cdot\bar p_{\sigma(l)}.$$

Taking the ratio of the above two equations (46) and (45) gives $\theta_{j,l}=\bar\theta_{j,\sigma(l)},\quad\forall j\notin S_1$. Redefining $\boldsymbol{r}^*:=\sum_{h\in S_2}\boldsymbol{e}_h$ similarly as above we have $\theta_{j,l}=\bar\theta_{j,\sigma(l)}$ for any $j\in S_1$. Plug $\theta_{j,l}=\bar\theta_{j,\sigma(l)}$ for all $j\in S_1$ into (45), then we have $p_l=\bar p_{\sigma(l)}$. Now we have shown (37) hold for this particular l. Then the induction argument gives

$$\forall l\in[L_0],\quad \forall j\in S_1\cup S_2,\quad \theta_{j,l}=\bar\theta_{j,\sigma(l)},\quad p_l=\bar p_{\sigma(l)}.$$

Now we have shown for any item j and latent class index l, $\theta_{j,l}=\bar\theta_{j,\sigma(l)}$, which we denote by $\bar{\boldsymbol{\Theta}}=\sigma(\boldsymbol{\Theta})$. We claim that this result also indicates that the permutation σ is unique. This is because $U_1\cdot\{W_1T(\Gamma^1,\boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}=I_L$ implies that

$$U_1\cdot\{W_1T(\bar\Gamma^1,\bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}=U_1\cdot\{W_1T(\Gamma^1,\boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}\cdot\sigma(I_L)=\sigma(I_L),$$

which means given U_1 constructed from $(\Gamma,\boldsymbol{\Theta})$, the form of $U_1\cdot\{W_1T(\bar\Gamma^1,\bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}$ explicitly and uniquely determines σ. Now we have shown $\bar\Gamma=\Gamma=\Gamma^{\mathcal{A}_0}$ and $(\bar{\boldsymbol{\Theta}},\bar{\boldsymbol{p}})=(\boldsymbol{\Theta},\boldsymbol{p})$ must hold up to the column permutation σ.

As stated in the beginning of the proof, combining Condition C that any column in $\Gamma^{\mathcal{A}_0}$ is different from any column in $\Gamma^{\mathcal{A}_0^c}$, the identification of $\Gamma^{\mathcal{A}_0}$ uniquely identifies the set of true patterns $\mathcal{A}_0$. The proof of both Theorem 2 and Corollary 3 is complete. □

Proof of Theorem 6 and Corollary 7. The following proofs of Theorem 6 and Corollary 7 use a similar proof idea as that of Allman et al. (2009); see also proofs of Theorems 4.2 and 4.3 in Gu and Xu (2019b).

Proof of Theorem 6. We need to introduce the definition of *algebraic variety*, a concept in algebraic geometry. An algebraic variety $\mathcal{V}$ is defined as the simulateneous zero-set of a finite collection of multivariate polynomials $\{f_i\}_{i=1}^n\subseteq\mathbb{R}[x_1,x_2,\ldots,x_d]$, $\mathcal{V}=\mathcal{V}(f_1,\ldots,f_n)=\{\boldsymbol{x}\in\mathbb{R}^d\mid f_i(\boldsymbol{x})=0,\ 1\le i\le n.\}$ An algebraic variety $\mathcal{V}$ is all of $\mathbb{R}^d$ only when all the polynomials defining it are zero polynomials; otherwise, $\mathcal{V}$ is called a *proper subvariety* and is of dimension less than d, hence necessarily of Lebesgue measure zero in $\mathbb{R}^d$. The same

argument holds when $\mathbb{R}^d$ is replaced by the parameter space $\Omega \subseteq \mathbb{R}^d$ that has full dimension in $\mathbb{R}^d$. For the structured latent attribute model, we consider the following parameter space,

$$\Omega = \Big\{(\boldsymbol{\Theta}, \boldsymbol{p}) : \forall\, j, \max_{\boldsymbol{\alpha}:\Gamma_{j,\boldsymbol{\alpha}}=1} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\Gamma_{j,\boldsymbol{\alpha}}=1} \theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'},\ \forall\, \Gamma_{j,\boldsymbol{\alpha}'} = 0\Big\}.$$

On Ω, altering some entries of zero to one in the Γ-matrix is equivalent to impose more affine constraints on the parameters and force them to be in a subset Ω^* of Ω. Condition $A^\star$ guarantees that, there exists a Ω^* such that Condition A holds for model parameters belonging to this Ω^*, the proof of Theorem 2 gives that the matrix $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ has full column rank C for $i = 1, 2$ for $(\boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)}, \boldsymbol{p}^{\mathcal{A}_0}) \in \Omega^*$. Note that the statement that $2^{|S_i|} \times C$ matrix $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ has full column rank is equivalent to the statement that the map sending $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ to all its $\binom{2^{|S_i|}}{C}$ possible $C \times C$ minors $A^i_1, A^i_2, \ldots, A^i_{2^{|S_i|}}$ yields at least one nonzero minor, where $A^i_1, A^i_2, \ldots, A^i_{2^{|S_i|}}$ are all polynomials of the item parameters $\boldsymbol{\Theta}_{S_i}$. Define

$$\mathcal{V} = \bigcup_{i=1,2} \Big\{ \bigcap_{l=1}^{2^{|S_i|}} \{(\boldsymbol{\Theta}, \boldsymbol{p}) \in \Omega : A^i_l(\boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)}) = 0\}\Big\},$$

then $\mathcal{V}$ is a algebraic variety defined by polynomials of the model parameters. Moreover, $\mathcal{V}$ is a proper subvariety of Ω, since the fact $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ has full column rank C for $i = 1, 2$ for one particular set of $(\boldsymbol{\Theta}, \boldsymbol{p}) \in \Omega^*$ ensures that there exists one particular set of model parameters that give nonzero values when plugged into the polynomials defining $\mathcal{V}$. This indicates that the polynomials defining $\mathcal{V}$ are not all zero polynomials on Ω. Then restricting parameters to Ω^* and proceeding in the same steps as the proof of Theorem 2 proves the conclusion of the proposition.

Proof of Corollary 7. Consider a Q-matrix in the form of (13). We denote $S_1 = \{1, \ldots, K\}$, $S_2 = \{K+1, \ldots, 2K\}$ and $S_3 = \{2K+1, \ldots, J\}$, which are item sets corresponding to Q_1, Q_2 and Q', respectively. According to the proof of Theorem 4.3 in Gu and Xu (2019b), since the two submatrices Q_1 and Q_2 have all the diagonal elements equal to one, the $2^K \times 2^K$ T-matrices $T(\Gamma^{(S_1,\text{all})}, \boldsymbol{\Theta}^{(S_1,\text{all})})$ and $T(\Gamma^{(S_2,\text{all})}, \boldsymbol{\Theta}^{(S_2,\text{all})})$ are generically full-rank. Furthermore, the matrix $T(\Gamma^{(S_3,\text{all})}, \boldsymbol{\Theta}^{(S_3,\text{all})}) \cdot \text{Diag}(\boldsymbol{p}^{\text{all}})$ has Kruskal rank at least two. This means generically, any two columns of $T(\Gamma^{(S_3,\text{all})}, \boldsymbol{\Theta}^{(S_3,\text{all})}) \cdot \text{Diag}(\boldsymbol{p}^{\text{all}})$ are linearly independent.

Now consider an arbitrary set of attribute patterns $\mathcal{A}_0 \subseteq \{0, 1\}$, we have the conclusion that $T(\Gamma^{(S_1,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_1,\mathcal{A}_0)})$ and $T(\Gamma^{(S_2,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_2,\mathcal{A}_0)})$ have full column rank generically. This is because for $i = 1, 2$, the $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ is just a submatrix of $T(\Gamma^{(S_i,\text{all})}, \boldsymbol{\Theta}^{(S_i,\text{all})})$ whose columns are a subset of different column vectors of the latter matrix. Therefore columns of $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ must be linearly independent, and hence the matrix must have full column rank generically. Also, the columns of $T(\Gamma^{(S_3,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_3,\mathcal{A}_0)}) \cdot \text{Diag}(\boldsymbol{p}^{\mathcal{A}_0})$ can also be considered as a subset of different columns of $T(\Gamma^{(S_3,\text{all})}, \boldsymbol{\Theta}^{(S_3,\text{all})}) \cdot \text{Diag}(\boldsymbol{p}^{\text{all}})$ up to a resealing of the columns. Therefore the former matrix must have any two different columns linearly independent generically and hence has Kruskal rank at least two. Now by Kruskal's conditions for unique tensor decomposition, a probability distribution of $\mathbf{R}$ with $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$, $i = 1, 2, 3$ having the above properties uniquely determines

$T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ and also $\boldsymbol{p}^{\mathcal{A}_0}$ generically. Therefore $(\Gamma^{\mathcal{A}_0}, \boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ are generically identifiable. Then combined with Condition C, we have the conclusion that $\mathcal{A}_0$ is generically identifiable. This completes the proof of the corollary. □

Proof of Corollary 10. Under our definition of $\mathcal{A}^{\text{rep}}$ and also Condition C, this matrix must have distinct column vectors, and each of its column corresponds to an equivalence class. We define $\boldsymbol{\Theta}^{\text{rep}}$ to be item parameters corresponding to the representative patterns in $\mathcal{A}^{\text{rep}}$. We further define the proportion parameters of the equivalence classes $\boldsymbol{\nu}^{\text{rep}} = (\nu_{[\boldsymbol{\alpha}_{\ell_1}]}, \ldots, \nu_{[\boldsymbol{\alpha}_{\ell_m}]})$, where $\nu_{[\boldsymbol{\alpha}_{\ell_i}]} > 0$ and $\sum_{i=1}^{m} \nu_{[\boldsymbol{\alpha}_{\ell_i}]} = 1$. Note that each $\nu_{[\boldsymbol{\alpha}_{\ell_i}]}$ is a sum of population proportions of the attribute patterns that are in the same equivalence class of $\boldsymbol{\alpha}_{\ell_i}$. Since $\Gamma^{\mathcal{A}^{\text{rep}}}$ also satisfies Conditions A and B by the assumption of the corollary. So Theorem 2 gives that $\mathcal{A}^{\text{rep}}$ is identifiable. □

Proof of Theorem 13 and Proposition 15. We use $L = |\mathcal{A}_{\text{input}}|$ to denote the number of attribute patterns as input given to the penalized likelihood method, then $L = 2^K$ if there is no screening stage as preprocessing. We denote the true proportion parameters by $\boldsymbol{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}_{\text{input}})$, where $p_{\boldsymbol{\alpha}} \geq 0$ for $\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}$ and $\sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} p_{\boldsymbol{\alpha}} = 1$. Denote the number of true attribute patterns by $|\mathcal{A}_0|$. We now consider the following log likelihood with penalty parameter λ_N for some $\gamma > 0$,

$$\begin{aligned}\ell^{\lambda_N}(\boldsymbol{p}, \boldsymbol{\Theta}) =& \frac{1}{N} \sum_{i=1}^{N} \log \Big\{ \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{R_{i,j}} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big\} \\ &+ \frac{\lambda_N}{N} \underbrace{\sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} \Big[\log p_{\boldsymbol{\alpha}} \cdot I(p_{\boldsymbol{\alpha}} > \rho_N) + \log \rho_N \cdot I(p_{\boldsymbol{\alpha}} \leq \rho_N) \Big]}_{\log_{\rho_N}(\boldsymbol{p})}.\end{aligned}$$

For a given λ_N, denote the estimated support of the proportion parameters $\widehat{\boldsymbol{p}}$ by $\widehat{\mathcal{A}}$, namely $\widehat{\mathcal{A}} = \{1 \leq l \leq L : \widehat{p}_{\boldsymbol{\alpha}_l} > \rho_N\}$. We denote the true and the estimated $|\mathcal{A}_{\text{input}}|$-dimensional proportions by $\boldsymbol{p}_{\text{full}}^{\mathcal{A}_0} = (p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}_{\text{input}} : p_{\boldsymbol{\alpha}} > 0 \text{ if and only if } \boldsymbol{\alpha} \in \mathcal{A}_0)$ and $\widehat{\boldsymbol{p}}_{\text{full}}^{\widehat{\mathcal{A}}} = (\widehat{p}_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}_{\text{input}} : \widehat{p}_{\boldsymbol{\alpha}} > \rho_N \text{ if and only if } \boldsymbol{\alpha} \in \widehat{\mathcal{A}})$. Denote the oracle MLE obtained assuming $\mathcal{A}_0$ is known by $\widehat{\boldsymbol{\Theta}}^0 := \widehat{\boldsymbol{\Theta}}^{\mathcal{A}_0}$ and $\widehat{\boldsymbol{p}}^0 := \widehat{\boldsymbol{p}}^{\mathcal{A}_0}$, and denote $\widehat{\boldsymbol{\eta}}^0 = (\widehat{\boldsymbol{\Theta}}^0, \widehat{\boldsymbol{p}}^0)$. Note that for $\widehat{\mathcal{A}} \neq \mathcal{A}_0$ the event $\{\ell^{\lambda_N}(\widehat{\boldsymbol{\eta}}^{\widehat{\mathcal{A}}}) > \ell^{\lambda_N}(\widehat{\boldsymbol{\eta}}^0)\}$ implies the following event

$$\begin{aligned}&\frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{\sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} \widehat{p}_{\boldsymbol{\alpha}} \prod_j \widehat{\theta}_{j,\boldsymbol{\alpha}}^{R_{i,j}} (1 - \widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}}}{\sum_{\boldsymbol{\alpha} \in \mathcal{A}_0} \widehat{p}_{\boldsymbol{\alpha}}^0 \prod_j (\widehat{\theta}_{j,\boldsymbol{\alpha}}^0)^{R_{i,j}} (1 - \widehat{\theta}_{j,\boldsymbol{\alpha}}^0)^{1-R_{i,j}}} \right] \\ &> \frac{|\lambda_N|}{N} \Big\{ \log_{\rho_N}(\widehat{\boldsymbol{p}}_{\text{full}}^{\widehat{\mathcal{A}}}) - \log_{\rho_N}(\widehat{\boldsymbol{p}}_{\text{full}}^{\mathcal{A}_0}) \Big\}.\end{aligned} \tag{47}$$

In the case of $|\widehat{\mathcal{A}}| > |\mathcal{A}_0|$ (which we call the overfitted case), the right hand side (RHS) of (47) regarding the difference between the penalty terms has order $O(N^{-1}|\lambda_N| \cdot |\mathcal{A}_0| \cdot |\log \rho_N|)$. In this overfitted case, we now consider the left hand side (LHS) of (47),

$$\text{LHS of (47)} = \frac{1}{N} \sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} \widehat{p}_{\boldsymbol{\alpha}} \prod_j \widehat{\theta}_{j,\boldsymbol{\alpha}}^{R_{i,j}} (1 - \widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \right]$$

$$- \frac{1}{N}\sum_{i=1}^{N} \log \left[\sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} \widehat{p}^{\,0}_{\boldsymbol{\alpha}} \prod_j (\widehat{\theta}^{0}_{j,\boldsymbol{\alpha}})^{R_{i,j}} (1-\widehat{\theta}^{0}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \right] \equiv I_1 - I_0,$$

where the I_1 part can be written as

$$\begin{aligned} I_1 &= \frac{1}{N}\sum_{i=1}^{N} \log \Big[\sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}},\\ \widehat{p}_{\boldsymbol{\alpha}}>\rho_N}} \widehat{p}_{\boldsymbol{\alpha}} \prod_j \widehat{\theta}^{R_{i,j}}_{j,\boldsymbol{\alpha}} (1-\widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big] + O(|\mathcal{A}_{\text{input}}|\rho_N) \\ &= \frac{1}{N}\sum_{i=1}^{N} \log \Big[\sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}},\\ \widehat{p}_{\boldsymbol{\alpha}}>\rho_N}} \widehat{p}_{\boldsymbol{\alpha}} \prod_j \widehat{\theta}^{R_{i,j}}_{j,\boldsymbol{\alpha}} (1-\widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big] + O(N^{-\delta}), \end{aligned} \tag{48}$$

where the last equality follows from the assumption $|\mathcal{A}_{\text{input}}| \cdot \rho_N = O(N^{-\delta})$ in the theorem. So we further have the LHS of (47) equal to

$$\begin{aligned} I_1 - I_0 = \frac{1}{N}\sum_{i=1}^{N} &\log \Big[\sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}},\\ \widehat{p}_{\boldsymbol{\alpha}}>\rho_N}} \widehat{p}_{\boldsymbol{\alpha}} \prod_j \widehat{\theta}^{R_{i,j}}_{j,\boldsymbol{\alpha}} (1-\widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big] \\ &- \frac{1}{N}\sum_{i=1}^{N} \log \Big[\sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} \widehat{p}^{\,0}_{\boldsymbol{\alpha}} \prod_j (\widehat{\theta}^{0}_{j,\boldsymbol{\alpha}})^{R_{i,j}} (1-\widehat{\theta}^{0}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big] + O(N^{-\delta}). \end{aligned}$$

Note that other than the last term $O(N^{-\delta})$ in the above display, the difference of the first two terms also has order $O_p(N^{-\delta})$ from assumption (19), so LHS of (47)$= I_1 - I_0 = O_p(N^{-\delta})$. In order to have selection consistency in the overfitted case, we need the event described in (47) to happen with probability tending to zero, so the $|\lambda_N|$ needs to be sufficiently large such that

$$N^{-\delta} \lesssim O\Big(N^{-1}|\lambda_N| \cdot |\mathcal{A}_0| \cdot |\log \rho_N|\Big). \tag{49}$$

Note that by (18), we have $\rho_N \asymp N^{-d}$ for some $d > 0$. So if $\delta < 1$, i.e., if the convergence rate is slower than the $\sqrt{N}$ rate, then λ_N must go to negative infinity as N goes to infinity since $\delta < 1$. Specifically, we obtain the following lower bound of the magnitude of the penalty parameter λ_N,

$$|\lambda_N| \gtrsim N^{1-\delta}/|\log \rho_N|$$

would suffice for (49) to hold.

We now prove the conclusion of Proposition 15. A further implication of the above discussion is that, with $\rho_N \asymp N^{-d}$ as assumed in (18), just imposing a proper Dirichlet prior with a positive hyperparameter would fail to select the true model consistently. In particular, with a proper Dirichlet prior density with hyperparameter $\beta = \lambda_N + 1 \in (0, 1)$, Equation (49) instead becomes $N^{1-\delta} = o(\log N)$. However, when $0 < \delta < 1$, $N^{1-\delta}/\log N \to \infty$. So (49) fails to hold, and one can not have consistent selection in the overfitted case. So if we denote the set of attribute patterns estimated by maximizing (17) by $\widehat{\mathcal{A}}^{\lambda}$. Then for any $\{\lambda_N\} \subseteq [-1, 0)$, $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda} = \mathcal{A}_0) \not\to 1$ as $N \to \infty$. This proves Proposition 15.

Now we consider the random set $\{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}} : \widehat{p}_{\boldsymbol{\alpha}} > \rho_N\} =: \widehat{\mathcal{A}}$ appearing in I_1 in (48). With probability tending to one, the cardinality of this set is smaller than $|\mathcal{A}_0|$. This is

because if $|\widehat{\mathcal{A}}| > |\mathcal{A}_0|$, the log-penalty term corresponding to $\widehat{\mathcal{A}}$ would be smaller than that corresponding to $\mathcal{A}_0$ by $N^{-1}|\lambda_N| \cdot |\log \rho_N|$ which has order at least $N^{-\delta}$. Recall that the right hand side of (47) has order $O_P(N^{-\delta})$, which means when $|\widehat{\mathcal{A}}| > |\mathcal{A}_0|$ the extent that the log-penalty part favors the a smaller model $\mathcal{A}_0$ would dominate the extent that the likelihood part favors a larger model $\widehat{\mathcal{A}}$ in the proposed penalized likelihood. Therefore any larger model $\widehat{\mathcal{A}}$ with $|\widehat{\mathcal{A}}| \geq |\mathcal{A}_0|$ would be favored over $\mathcal{A}_0$ with probability tending to zero. Therefore we have the conclusion that $\mathbb{P}(\widehat{\mathcal{A}} \neq \mathcal{A}_0) \not\to 0$ could only happen for $|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|$. So in the following discussion we will focus on the case where $|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|$ and prove consistency in this case. Namely, we aim to bound

$$\mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|,\, \widehat{\mathcal{A}} \neq \mathcal{A}_0} [\ell^{\lambda_N}(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell^{\lambda_N}(\boldsymbol{\eta}^{\mathcal{A}_0})] > 0 \Big). \tag{50}$$

Next, we consider the upper bound of the magnitude of the penalty term. In order to have selection consistency in the case of $|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|$ and $\widehat{\mathcal{A}} \neq \mathcal{A}_0$, the log-penalty term can not be too large such that the extent that the penalty part favors a smaller model does not dominate the extent that the likelihood part favors the true model. We follow a similar argument to Shen et al. (2012). Specifically, considering the term $-\epsilon_N^2 \to 0$ in the large deviation inequality (53) below; for a small constant $t > \epsilon_N$, we need that the difference of the penalty part of the true and any alternative smaller model to be less than t^2, i.e.,

$$|\lambda_N| \cdot |\mathcal{A}_0| \cdot |\log \rho_N|/N \lesssim t^2, \tag{51}$$

Equation (51) would hold if

$$|\lambda_N| = o(N/|\log \rho_N|). \tag{52}$$

We next show that such λ_N can guarantee selection consistency. So we have a sample-size dependent λ_N that penalizes the overfitted mixture and constrains the support size of the proportion parameters to be less than the true support size $|\mathcal{A}_0|$. As said, with such λ_N it suffices to consider the case $|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|$.

In order to bound this mis-selection probability, we need to introduce the notion of bracketing Hellinger metric entropy $H(t, \mathcal{B}_{\mathcal{A}})$. Let $h(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0})$ denote the Hellinger distance between the probability mass functions of $\mathbf{R}$ indexed by $\boldsymbol{\eta}^{\mathcal{A}}$ and $\boldsymbol{\eta}^{\mathcal{A}_0}$, i.e.,

$$h(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0}) = \Big(\sum_{\boldsymbol{r} \in \{0,1\}^J} \Big[\mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}})^{\frac{1}{2}} - \mathbb{P}(\mathbf{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})^{\frac{1}{2}} \Big] \Big)^{\frac{1}{2}}.$$

Consider the local parameter space $\mathcal{B}_{\mathcal{A}} = \{\boldsymbol{\eta}^{\mathcal{A}} = (\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) : |\mathcal{A}| \leq |\mathcal{A}_0|,\, h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0}) \leq 2\epsilon_N^2\}$, the $H(t, \mathcal{B}_{\mathcal{A}})$ is defined as the logarithm of the cardinality of the t-bracketing of $\mathcal{B}_{\mathcal{A}}$ of the smallest size. More specifically, following the definition in Shen et al. (2012), consider a bracket covering $S(t, m) = \{f_1^l, f_1^u, \ldots, f_m^l, f_m^u\}$ satisfying that $\max_{1 \leq j \leq m} \|f_j^u - f_j^l\|_2 \leq t$ and for any $f \in \mathcal{B}_{\mathcal{A}}$ there is some j such that $f_j^l \leq f \leq f_j^u$ almost surely. Then $H(t, \mathcal{B}_{\mathcal{A}})$ is $\log(\min\{m : S(t, m)\})$. The $H(t, \mathcal{B}_{\mathcal{A}})$ measures the complexity of the local parameter space. The next lemma gives an upper bound for the bracketing Hellinger metric entropy $H(t, \mathcal{B}_{\mathcal{A}})$ for $|\mathcal{A}| \leq |\mathcal{A}_0|$.

Lemma 21 *Denote $N_{[]}(t, \mathcal{B}_{\mathcal{A}}) = \exp(H(t, \mathcal{B}_{\mathcal{A}}))$. For the considered structured latent attribute model, denote the item parameter space of the ℓ-th attribute pattern by $\mathcal{F}_\ell$. For $|\mathcal{A}| \leq |\mathcal{A}_0|$ and any $2^{-4}\epsilon < t < \epsilon$, there is $H(t, \mathcal{B}_{\mathcal{A}}) \lesssim |\mathcal{A}_0| \log |\mathcal{A}_{\text{input}}| \log(2\epsilon/t)$.*

By the assumption of the theorem there is $\log|\mathcal{A}_{\text{input}}|/N \to 0$, so if we take

$$\epsilon_N = \sqrt{1/N|\mathcal{A}_0|\log|\mathcal{A}_{\text{input}}|},$$

there is $\epsilon_N = o(1)$. We next verify the entropy integral condition in Theorem 1 of Wong and Shen (1995) is satisfied with this ϵ_N, in order to obtain a large deviation inequality to bound the mis-selection probability. With Lemma 21, the integral of bracketing Hellinger metric entropy in the interval $[2^{-8}\epsilon_N^2, \sqrt{2}\epsilon_N]$ satisfies the following inequality

$$\begin{aligned}\int_{2^{-8}\epsilon_N^2}^{\sqrt{2}\epsilon_N} H^{1/2}(t,\mathcal{B}_{\mathcal{A}})dt &\le \int_{2^{-8}\epsilon_N^2}^{\sqrt{2}\epsilon_N} \sqrt{|\mathcal{A}_0|\log|\mathcal{A}_{\text{input}}|\log(2\epsilon_N/t)}dt \\ &= \sqrt{|\mathcal{A}_0|\log|\mathcal{A}_{\text{input}}|}\int_{\sqrt{\log\sqrt{2}}}^{\sqrt{\log\frac{2^9}{\epsilon_N}}} 4\epsilon_N u^2 e^{-u^2}du \\ &= \sqrt{|\mathcal{A}_0|\log|\mathcal{A}_{\text{input}}|}\cdot 2\epsilon_N \underbrace{\int_{\log\sqrt{2}}^{\log\frac{2^9}{\epsilon_N}} \sqrt{u}e^{-u}du}_{\text{bounded as } \epsilon_N\to 0} \lesssim \sqrt{N}\epsilon_N^2.\end{aligned}$$

So the entropy integral condition in Theorem 1 in Wong and Shen (1995) is satisfied and the large deviation inequality there holds. In particular, we have

$$\begin{aligned}&\mathbb{P}\Big(\sup_{h^2(\widehat{\boldsymbol{\eta}}^{\widehat{\mathcal{A}}},\boldsymbol{\eta}^{\mathcal{A}_0})\ge\epsilon_N^2}\Big[\frac{1}{N}\ell(\widehat{\boldsymbol{\eta}}^{\widehat{\mathcal{A}}}) - \frac{1}{N}\ell(\widehat{\boldsymbol{\eta}}^{\mathcal{A}_0})\Big] > -\epsilon_N^2\Big) \\ \le\ &\mathbb{P}\Big(\sup_{h^2(\widehat{\boldsymbol{\eta}}^{\widehat{\mathcal{A}}},\boldsymbol{\eta}^{\mathcal{A}_0})\ge\epsilon_N^2}\Big[\frac{1}{N}\ell(\widehat{\boldsymbol{\eta}}^{\widehat{\mathcal{A}}}) - \frac{1}{N}\ell(\boldsymbol{\eta}^{\mathcal{A}_0})\Big] > -\epsilon_N^2\Big) \le \exp(-N\epsilon_N^2). \qquad (53)\end{aligned}$$

where $\boldsymbol{\eta}^{\mathcal{A}_0} = (\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ denote the true parameters. Indeed, Theorem 1 in Wong and Shen (1995) guarantees the inequality (53) holds with ϵ_N replaced by any $t > \epsilon_N = \sqrt{|\mathcal{A}_0|\log|\mathcal{A}_{\text{input}}|/N}$. This large deviation inequality will be used later to bound the mis-selection probability in the case of $|\mathcal{A}| \le |\mathcal{A}_0|$.

We next further look at the Hellinger distance between $\boldsymbol{\eta}^0 := \boldsymbol{\eta}^{\mathcal{A}_0}$ and $\boldsymbol{\eta}^{\mathcal{A}}$ for $|\mathcal{A}| \le |\mathcal{A}_0|$, and investigate how the distance between a set of true patterns $\mathcal{A}_0$ and an alternative set relate to identifiability of $\mathcal{A}_0$.

$$\begin{aligned}&\frac{h^2(\boldsymbol{\eta}^{\mathcal{A}},\boldsymbol{\eta}^{\mathcal{A}_0})}{\max(|\mathcal{A}_0\setminus\mathcal{A}|,1)} \\ \asymp\ &[\max(|\mathcal{A}_0\setminus\mathcal{A}|,1)]^{-1}\sum_{\boldsymbol{r}\in\{0,1\}^J}\Big[\Big(\sum_{\boldsymbol{\alpha}\in\mathcal{A}}\mathbb{P}(\mathbf{R}=\boldsymbol{r}\mid\boldsymbol{\Theta}^{\mathcal{A}},\mathbf{A}=\boldsymbol{\alpha})p_{\boldsymbol{\alpha}}^{\mathcal{A}}\Big)^{1/2} - \\ &\qquad\Big(\sum_{\boldsymbol{\alpha}\in\mathcal{A}_0}\mathbb{P}(\mathbf{R}=\boldsymbol{r}\mid\boldsymbol{\Theta}^{\mathcal{A}_0},\mathbf{A}=\boldsymbol{\alpha})p_{\boldsymbol{\alpha}}^{\mathcal{A}_0}\Big)^{1/2}\Big]^2 \\ \asymp\ &[\max(|\mathcal{A}_0\setminus\mathcal{A}|,1)]^{-1}\sum_{\boldsymbol{r}\in\{0,1\}^J}\Big(\sum_{\boldsymbol{\alpha}\in\mathcal{A}}T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\Theta}^{\mathcal{A}})\,p_{\boldsymbol{\alpha}}^{\mathcal{A}} - \sum_{\boldsymbol{\alpha}\in\mathcal{A}_0}T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\Theta}^{\mathcal{A}_0})\,p_{\boldsymbol{\alpha}}^{\mathcal{A}_0}\Big)^2\end{aligned}$$

$$= [\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)]^{-1} \|T(\Gamma^{\mathcal{A}}, \boldsymbol{\Theta}^{\mathcal{A}})\boldsymbol{p}^{\mathcal{A}} - T(\Gamma^{\mathcal{A}_0}, \boldsymbol{\Theta}^{\mathcal{A}_0})\boldsymbol{p}^{\mathcal{A}_0}\|_2^2$$

To proceed with the proof, we need to use Theorem 2 to establish an identifiability argument. Theorem 2 and Corollary 3 state that if the true constraint matrix $\Gamma^{\mathcal{A}_0}$ satisfies conditions A, B and C, then $(\Gamma^{\mathcal{A}_0},\ \boldsymbol{\Theta}^{\mathcal{A}_0},\ \boldsymbol{p}^{\mathcal{A}_0})$ are jointly identifiable. This implies that given the set of true attribute patterns $\mathcal{A}_0$, for any other set $\mathcal{A} \neq \mathcal{A}_0$, $|\mathcal{A}| \leq |\mathcal{A}_0|$, and model parameters defined by $\mathcal{A}$ must lead to different $T(\boldsymbol{\Theta}^{\mathcal{A}})\boldsymbol{p}^{\mathcal{A}}$ that is different from $T(\boldsymbol{\Theta}^{\mathcal{A}_0})\boldsymbol{p}^{\mathcal{A}_0}$. Moreover, consider the parameter space $\mathcal{B} = \{(\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) :\ |\mathcal{A}| \leq |\mathcal{A}_0|,\ p_{\boldsymbol{\alpha}} > \rho_N\ \forall \boldsymbol{\alpha} \in \mathcal{A}\}$. Then $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0}) \in \mathcal{B}$ and for any $(\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) \in \mathcal{B}$ with $\mathcal{A} \neq \mathcal{A}_0$, **either** some elements in $\boldsymbol{\Theta}^{\mathcal{A}}$ differs from those in $\boldsymbol{\Theta}^{\mathcal{A}_0}$ by a nonzero constant, **or** some elements in $\boldsymbol{p}^{\mathcal{A}}$ differs from those in $\boldsymbol{p}^{\mathcal{A}_0}$ by a nonzero constant. Since $T^{\mathcal{A}}(\boldsymbol{\Theta})\boldsymbol{p}^{\mathcal{A}}$ is a continuous vector-valued function of the model parameters, we must have $[\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)]^{-1}\|T^{\mathcal{A}}(\boldsymbol{\Theta})\boldsymbol{p}^{\mathcal{A}} - T^{\mathcal{A}_0}(\boldsymbol{\Theta})\boldsymbol{p}^{\mathcal{A}_0}\|_2^2 \geq C_0$ for some $C_0 > 0$. By the conditions of the theorem $\epsilon_N^2 = o(1)$, so we have obtained for some small constant $t > \epsilon_N$,

$$C_{\min}(\boldsymbol{\eta}^0) \equiv \inf_{\boldsymbol{\eta}^{\mathcal{A}}:\, \mathcal{A} \neq \mathcal{A}_0, |\mathcal{A}| \leq |\mathcal{A}_0|} \left\{ \frac{h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0})}{\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)} \right\} \geq C_0 \gtrsim t^2 > \epsilon_N^2. \tag{54}$$

Finally, with the λ_N of the previously specified order, we use the large deviation inequality (53) and also the (54) to bound the false selection probability (50). The following argument uses a similar proof idea as that of Theorem 1 in Shen et al. (2012) which establishes finite sample mis-selection error bound of the L_0-constrained maximum likelihood estimation. Consider $|\widehat{\mathcal{A}} \cap \mathcal{A}_0| = m \leq |\mathcal{A}_0| - 1$, by (54) we have $h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0}) \geq (|\mathcal{A}_0| - m) C_{\min}(\boldsymbol{\eta}^0)$. So

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|,\, \widehat{\mathcal{A}} \neq \mathcal{A}_0} \Big[\frac{1}{N} \ell^{\lambda_N}(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \frac{1}{N} \ell^{\lambda_N}(\boldsymbol{\eta}^0) \Big] > 0 \Big) \\
\leq\ & \sum_{m=0}^{|\mathcal{A}_0|-1} \sum_{j=1}^{|\mathcal{A}_0|-m} \mathbb{P}\Big(\sup_{\substack{h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0}) \geq \\ (|\mathcal{A}_0| - m) C_{\min}(\boldsymbol{\eta}^0)}} \frac{1}{N} \Big[\ell(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell(\boldsymbol{\eta}^0) \Big] > -\frac{|\lambda_N| \cdot |\mathcal{A}_0| \cdot |\log \rho_N|}{N} \Big) \\
\leq\ & \sum_{m=0}^{|\mathcal{A}_0|-1} \sum_{j=1}^{|\mathcal{A}_0|-m} \mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}} \cap \mathcal{A}_0| = m} \frac{1}{N} \Big[\ell(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell(\boldsymbol{\eta}^0) \Big] > -t^2 \Big) \quad \text{(by (52))} \\
\leq\ & \sum_{m=0}^{|\mathcal{A}_0|-1} \sum_{j=1}^{|\mathcal{A}_0|-m} \mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}} \cap \mathcal{A}_0| = m} \frac{1}{N} \Big[\ell(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell(\boldsymbol{\eta}^0) \Big] > -(|\mathcal{A}_0| - m) C_{\min}(\boldsymbol{\eta}^0) \Big) \quad \text{(by (54))} \\
\leq\ & \sum_{m=0}^{|\mathcal{A}_0|-1} \binom{|\mathcal{A}_0|}{m} \exp\Big(-c_2 N (|\mathcal{A}_0| - m) C_{\min}(\boldsymbol{\eta}^0) \Big) \sum_{j=1}^{|\mathcal{A}_0|-m} \binom{|\mathcal{A}_{\text{input}}| - |\mathcal{A}_0|}{j} \quad \text{(by (53))} \\
\leq\ & c_3 \exp\Big(-c_2 N C_{\min}(\boldsymbol{\eta}^0) + 2\log(|\mathcal{A}_{\text{input}}| + 1) \Big),
\end{aligned}$$

where the last but one line above uses the large deviation inequality in (53), and c_2, c_3 are some constants. And the last line follows from the calculations in the proof of Theorem 1 in Shen et al. (2012) using some basic inequalities about binomial coefficients. Since $C_{\min}(\boldsymbol{\eta}^0) \geq C_0$, and $\log|\mathcal{A}_{\text{input}}| = o(N)$ by the assumption of the theorem, the right hand

side of the above display goes to zero as $N \to \infty$. Therefore $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} \neq \mathcal{A}_0,\ |\widehat{\mathcal{A}}^{\lambda_N}| \leq |\mathcal{A}_0|) \to 0$ as $N \to \infty$. Combined with the previously shown result $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} \neq \mathcal{A}_0) \not\to 0$ could potentially happen only for $|\widehat{\mathcal{A}}^{\lambda_N}| \leq |\mathcal{A}_0|$, we have the conclusion $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} \neq \mathcal{A}_0) \to 0$ as $N \to \infty$. The proof of the theorem is complete. □

Proof of Theorem 17. Denote $\theta_j^+ = \theta_{j,H}$ and $\theta_j^- = \max_{\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j} \theta_{j,\boldsymbol{\alpha}}$ for each j. Since the screening algorithm is developed for the two-parameter SLAM introduced in Example 2, for each item j there are exactly two estimated item parameters, and we denote them by $\widehat{\theta}_j^+$ and $\widehat{\theta}_j^-$. We claim that it suffices to prove that for any $\boldsymbol{\alpha} \in \mathcal{A}_0$, there exists a response pattern $\boldsymbol{r}^{\boldsymbol{\alpha}} \in \{0,1\}^J$ such that as $K \to \infty$,

$$\mathbb{P}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathbf{A} = \boldsymbol{\alpha} \mid \boldsymbol{\Theta}) > \mathbb{P}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathbf{A} = \widetilde{\boldsymbol{\alpha}} \mid \boldsymbol{\Theta}), \quad \forall \widetilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}. \tag{55}$$

For $\boldsymbol{\alpha} \in \mathcal{A}_0$, define $\boldsymbol{r}^{\boldsymbol{\alpha}} = (r_1^{\boldsymbol{\alpha}}, \ldots, r_J^{\boldsymbol{\alpha}})$ to be $r_j^{\boldsymbol{\alpha}} = I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j) = \prod_k \alpha_k^{q_{j,k}}$. For a general structured latent attribute model, consider the joint distribution of observed response vector $\mathbf{R}$ and latent attribute pattern vector $\mathbf{A}$ is

$$\mathbb{P}(\mathbf{R} = \boldsymbol{r},\ \mathbf{A} = \boldsymbol{\alpha} \mid \boldsymbol{\Theta}) = \exp\Big\{ \sum_{j=1}^{J} \Big[r_j \Big(\prod_k \alpha_k^{q_{j,k}} \log \theta_j^+ + (1 - \prod_k \alpha_k^{q_{j,k}}) \log \theta_{j,\boldsymbol{\alpha}}^- \Big) + (1 - r_j) \Big(\prod_k \alpha_k^{q_{j,k}} \log(1 - \theta_j^+) + (1 - \prod_k \alpha_k^{q_{j,k}}) \log(1 - \theta_{j,\boldsymbol{\alpha}}^-) \Big) \Big] \Big\}.$$

Therefore

$$\mathbb{P}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathbf{A} = \widetilde{\boldsymbol{\alpha}} \mid \boldsymbol{\Theta}) = \exp\Big\{ \sum_{j=1}^{J} \Big[\prod_k \alpha_k^{q_{j,k}} \Big(\prod_k \widetilde{\alpha}_k^{q_{j,k}} \log \theta_j^+ + (1 - \prod_k \widetilde{\alpha}_k^{q_{j,k}}) \log \theta_{j,\widetilde{\boldsymbol{\alpha}}}^- \Big) + (1 - \prod_k \alpha_k^{q_{j,k}}) \Big(\prod_k \widetilde{\alpha}_k^{q_{j,k}} \log(1 - \theta_j^+) + (1 - \prod_k \widetilde{\alpha}_k^{q_{j,k}}) \log(1 - \theta_{j,\widetilde{\boldsymbol{\alpha}}}^-) \Big) \Big] \Big\}.$$

$$\mathbb{P}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathbf{A} = \boldsymbol{\alpha} \mid \boldsymbol{\Theta}) = \exp\Big\{ \sum_{j=1}^{J} \Big[\prod_k \alpha_k^{q_{j,k}} \log \theta_j^+ + (1 - \prod_k \alpha_k^{q_{j,k}}) \log(1 - \theta_{j,\boldsymbol{\alpha}}^-) \Big] \Big\}.$$

Then for any $\widetilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}$,

$$\begin{aligned} &\log \mathbb{P}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathbf{A} = \boldsymbol{\alpha} \mid \boldsymbol{\Theta}) - \log \mathbb{P}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathbf{A} = \widetilde{\boldsymbol{\alpha}} \mid \boldsymbol{\Theta}) \\ &\geq \min_{j=1,\ldots,J} \{ \log \theta_j^+ - \log \theta_{j,\widetilde{\boldsymbol{\alpha}}}^-,\ \log(1 - \theta_{j,\boldsymbol{\alpha}}^+) - \log(1 - \theta_j^+) \} \geq d > 0. \end{aligned} \tag{56}$$

That the above probability is bounded away from zero follows from the second part of assumption (20). So the claim (55) is proved. We next bound the probability of failure of including all the true patterns in the screening stage. First, since $\mathbf{A}_1, \ldots, \mathbf{A}_N \overset{i.i.d.}{\sim} \text{Multinomial}(N,\ (p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}_0))$, then $|\{i \in [N] : \mathbf{A}_i = \boldsymbol{\alpha}\}|$ denotes the number of subjects in the random sample whose attribute pattern is $\boldsymbol{\alpha}$. By the concentration inequality of the multinomial distribution, for any $\boldsymbol{\alpha} \in \mathcal{A}_0$,

$$\mathbb{P}\Big(\Big| \{i \in [N] : \mathbf{A}_i = \boldsymbol{\alpha}\} \Big| \geq N p_{\boldsymbol{\alpha}} - 2\sqrt{N} t \Big) \geq 1 - 2^{|\mathcal{A}_0|} \exp(-2t^2), \quad \forall t > 0.$$

Because of (20), we have $Np_{\boldsymbol{\alpha}} \geq Nc_0 \to \infty$ for all $\boldsymbol{\alpha} \in \mathcal{A}_0$. Assume that $\widehat{\theta}_j^+ - \widehat{\theta}_j^- > \delta > 0$ for each $j \in [J]$. This constraint can be incorporated into the screening procedure or checked *a posteiriori* after screening. So with probability at least $1 - 2^{|\mathcal{A}_0|}\exp(-2t^2)$ for a suitable t,

$$\begin{aligned}
&\mathbb{P}(\widehat{\mathcal{A}}_{\text{screen}} \not\supseteq \mathcal{A}_0) \leq \sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} \mathbb{P}(\widehat{\mathbf{A}}_i \neq \boldsymbol{\alpha}\ \forall i \in [N] \text{ s.t. } \mathbf{A}_i = \boldsymbol{\alpha}) \\
&\leq \sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} \Big[\mathbb{P}\Big(\mathbf{R}_i = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \exists \widetilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}, \\
&\qquad \widehat{\mathbb{P}}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}}, \mathbf{A} = \boldsymbol{\alpha}) > \widehat{\mathbb{P}}(\mathbf{R} = \boldsymbol{r}^{\boldsymbol{\alpha}}, \mathbf{A} = \widetilde{\boldsymbol{\alpha}})\Big|\mathbf{A}_i = \boldsymbol{\alpha}\Big)\Big]^{N\left(p_{\boldsymbol{\alpha}} - 2t/\sqrt{N}\right)} \to 0,
\end{aligned}$$

as $N \to \infty$. Here $\widehat{\mathbb{P}}$ refers to the probability measure of $\mathbf{R}$ and $\mathbf{A}$ given the estimated item parameters $\widehat{\boldsymbol{\theta}}^+$ and $\widehat{\boldsymbol{\theta}}^-$. This is because the probability inside the bracket in the above expression is strictly less than 1 due to (56); we denote this quantity by C_δ since it depends on δ. Therefore there is

$$\begin{aligned}
\mathbb{P}(\widehat{\mathcal{A}}_{\text{screen}} \not\supseteq \mathcal{A}_0) \leq \sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} C_\delta^{N(p_{\boldsymbol{\alpha}} + o(1))} &= \sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} \exp[-N(p_{\boldsymbol{\alpha}} + o(1))\log(1/C_\delta)] \\
&\leq |\mathcal{A}_0|\exp(-N\beta_{\min}),
\end{aligned}$$

where $\beta_{\min}$ is a positive constant which can be taken as $c_0/2\log(1/C_\delta)$. The last inequality above results from $p_{\boldsymbol{\alpha}} \geq c_0$ for $\boldsymbol{\alpha} \in \mathcal{A}_0$ in (20) and that $C_\delta < 1$. Now we have obtained $\mathbb{P}(\widehat{\mathcal{A}}_{\text{screen}} \supseteq \mathcal{A}_0) \geq 1 - |\mathcal{A}_0|\exp(-N\beta_{\min})$, so the sure screening property holds and the proof is complete. □

Proof of Lemma 21. Following the proof of Theorem 2 in Genovese and Wasserman (2000), the overall bracketing entropy of the mixture distribution over $|\mathcal{A}|$ mixture components (latent attribute patterns) can be bounded by the entropy of the $|\mathcal{A}| - 1$ dimensional simplex multiplied by the product of the entropy of the item parameter space for each mixture component. Since there are a total number of $\binom{|\mathcal{A}_{\text{input}}|}{|\mathcal{A}|}$ possibilities of choosing $|\mathcal{A}|$ components from $|\mathcal{A}_{\text{input}}|$ ones, we have

$$N_{[]}(t, \mathcal{B}_{\mathcal{A}}) \leq \binom{|\mathcal{A}_{\text{input}}|}{|\mathcal{A}|} N_{[]}(t, \mathcal{T}^{|\mathcal{A}|-1}) \prod_{l=1}^{|\mathcal{A}|} N_{[]}(t/3, \mathcal{F}_l).$$

Next, Lemma 2 in Genovese and Wasserman (2000) gives the following bracketing entropy bound for the simplex, $N_{[]}(t, \mathcal{T}^{|\mathcal{A}|-1}) \leq |\mathcal{A}|(2\pi e)^{|\mathcal{A}|/2}/t^{|\mathcal{A}|-1}$. Since we consider the local parameter space around the true parameters (with squared Hellinger distance between the alternative model and the true model not greater than $2\epsilon^2$), the $1/t$ in the above display can be replaced by ϵ/t. Also, $N_{[]}(t/3, \mathcal{F}_l) \leq C_0\epsilon/t$ since the Hellinger distance is bounded by the L_2 distance and the t-bracketing number under the L_2 norm is bounded by $O(\epsilon/t)$. Therefore we have

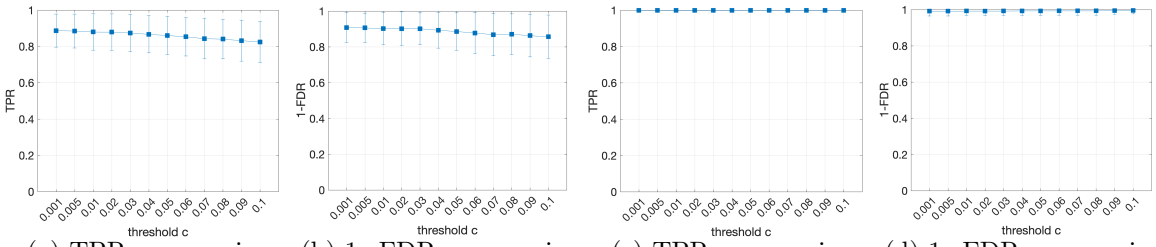
$$\begin{aligned}
H(t, \mathcal{B}_{\mathcal{A}}) &\leq \log\Big\{\binom{|\mathcal{A}_{\text{input}}|}{|\mathcal{A}|} \frac{|\mathcal{A}|(2\pi e)^{|\mathcal{A}|/2}(\epsilon)^{|\mathcal{A}|-1}}{t^{|\mathcal{A}|-1}} \Big(\frac{\epsilon}{t}\Big)^{|\mathcal{A}|}\Big\} \\
&\lesssim |\mathcal{A}|\log|\mathcal{A}_{\text{input}}| + \log|\mathcal{A}| + |\mathcal{A}|\log(\epsilon/t)
\end{aligned}$$

$$\lesssim |\mathcal{A}_0| \log |\mathcal{A}_{\text{input}}| \log(\epsilon/t).$$

where $|\mathcal{A}| \le |\mathcal{A}_0|$ and an elementary inequality $\binom{a}{b} \le a^b$ are used. □

Appendix B: Additional Experimental Results

Impact of the value of the pre-specified c in Algorithm 1. In Algorithm 1, there is a pre-specified constant $c > 0$ when updating the Δ_l's. This constant c should be small, ideally close to zero. In all of our experiments in Section 5, we take $c = 0.01$. Next we examine how the value of c impacts the selection result of Algorithm 1. Since the performance of Algorithm 1 is the focus here, we choose the simulation setting with $K = 10$ such that screening can be omitted. Under sample sizes $N = 150$ and $N = 500$, the plots of the two accuracy measures versus c are presented in Figure 10. We observe that the results of Algorithm 1 are generally not that sensitive to the choice of c, though smaller c gives slightly better results for both accuracy measures under a small sample size $N = 150$. For N as large as 500, for all the values of $c \in \{0.001, 0.005\} \cup \{0.01 \times i : i = 1, 2, \ldots, 10\}$, the two accuracy measures are very close to one and do not have much variation. In practice, we recommend fixing c to a value no greater than 0.01.

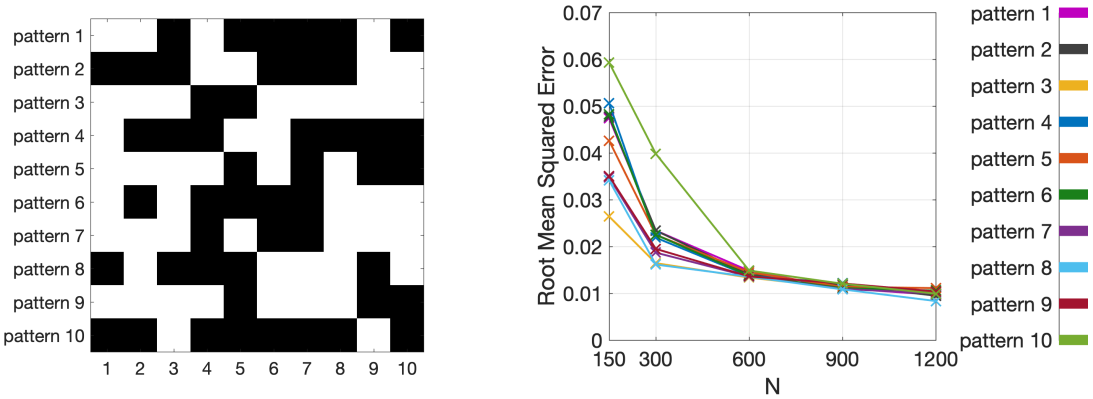


(a) TPR versus c in Algo. 1, $N = 150$ (b) 1−FDR versus c in Algo. 1, $N = 150$ (c) TPR versus c in Algo. 1, $N = 500$ (d) 1−FDR versus c in Algo. 1, $N = 500$

Figure 10: Performance of Algorithm 1 across various values for threshold c. Setting is $K = 10$ and $1 - \theta_j^+ = \theta_j^- = 0.2$. In each scenario 200 runs are carried out, and the error bar is within one standard deviation of the mean accuracy.

Algorithm 1's performance on estimating the actual proportions of patterns. Other than the two accuracy measures for pattern selection presented in Table 2, we also evaluate how well the algorithms perform on estimating the actual proportions of the latent patterns. We use the simulation setting of the two-parameter SLAM with $K = 10$, $|\mathcal{A}_0| = 10$, $Q = (Q_1^\top, Q_2^\top, Q_3^\top)^\top$, with parameters $p_{\boldsymbol{\alpha}} = 0.1$ for $\boldsymbol{\alpha} \in \mathcal{A}$ and $1 - \theta_j^+ = \theta_j^- = 0.2$. This is the same setting as that of Example 7. We vary the sample size $N \in \{150, 300, 600, 900, 1200\}$ and compute the Root Mean Square Errors (RMSEs) of estimating the true proportions of latent patterns. The randomly generated 10 true patterns in $\mathcal{A}_0$ are presented in Figure 11(a), where each row represents a K-dimensional binary pattern. For each N, the RMSE of each proportion $p_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathcal{A}_0$ is computed based on 200 runs; and in each run, we first perform pattern selection by using EBIC to choose $\lambda \in \{-0.2 \times i : i = 1, 2, \ldots, 20\}$ in Algorithm 1 and then estimate the proportions based on the selected set of patterns. The results of RMSEs are presented in Figure 11(b). As can be seen from the figure, under a small sample size $N = 150$, the RMSEs of patterns are rel-

atively diverse. In particular, the largest RMSE is around 0.06 and corresponds to pattern 10, $\boldsymbol{\alpha}_{10} = (0010000010)$, which is the pattern consisting of most "0"s; while the smallest RMSE is less than half of the largest and corresponds to pattern 3, $\boldsymbol{\alpha}_3 = (1110011111)$, which is the pattern consisting of most "1"s. Interestingly, this observation implies for a very small sample size and a sparse Q-matrix (each row having at most three entries of "1"s), those attribute patterns possessing fewer attributes are harder to estimate while those possessing more attributes are easier to estimate. While as N increases, the RMSEs of all the proportions decrease and their difference become not discernible. For $N = 1200$, all the RMSEs are around 0.01.



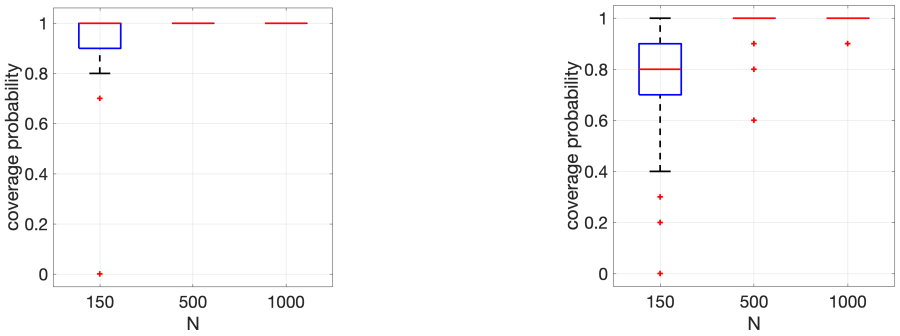
(a) patterns in $\mathcal{A}_0$: white for 1, black for 0

(b) RMSE of $\{p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}_0\}$ versus N

Figure 11: Root Mean Square Errors (RMSEs) for estimating the true proportions of patterns decrease as sample size N increases. Results are based on 200 runs for each N.

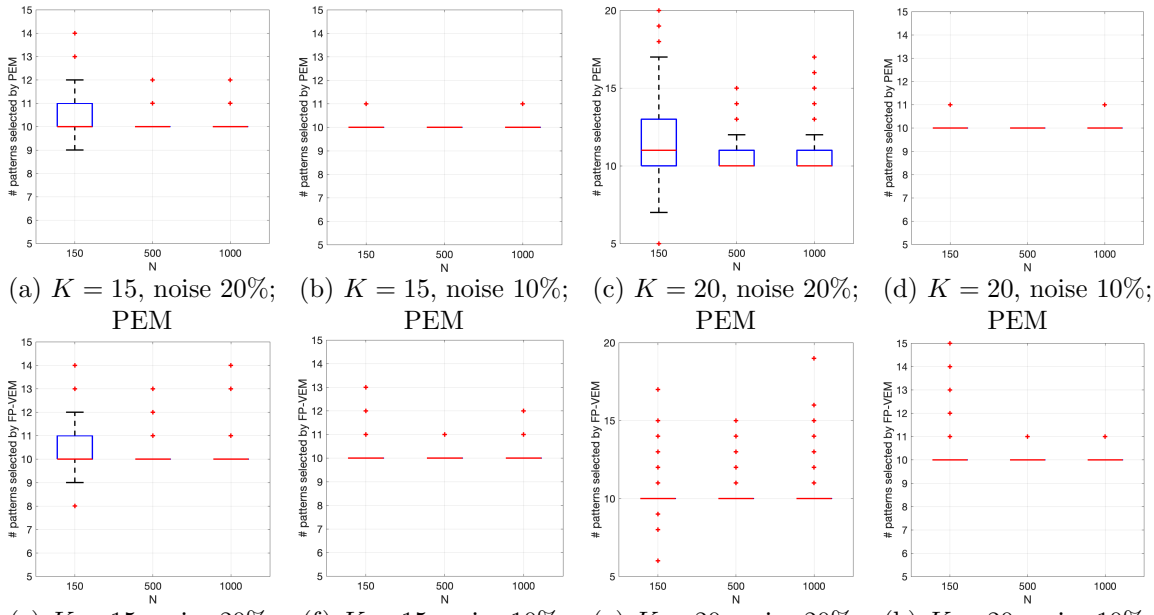
Evaluating the screening procedure under the multi-parameter SLAM. In the multi-parameter setting, we also evaluate the performance of the approximate screening procedure that is developed based on the likelihood of the two-parameter model. The results of the coverage probabilities are presented in Figure 12. The figure shows that despite being an approximate procedure, the screening Algorithm 3 has excellent performance for the multi-parameter SLAM that covers the two-parameter model as a submodel. Specifically, Figure 12 shows that for both $K = 15$ and $K = 20$, the approximate screening procedure almost always has a 100% coverage probability for $N = 500$ and $N = 1000$.

Sizes of the set of finally selected patterns under scenarios in Table 2. We present the results of the number of patterns that are finally selected by the proposed methods, corresponding to simulation scenarios in Table 2. Denote the set of patterns selected by the PEM algorithm and that selected by the FP-VEM algorithm by $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$, respectively. As shown in Figure 13, in the relatively strong signal setting with $1-\theta_j^+ = \theta_j^- = 10\%$, the sizes of $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$ almost always equal 10, the number of true patterns. Combined with the accuracy measures presented in Table 2 in the main text, in most cases these selected 10 patterns are indeed exactly the true ones in $\mathcal{A}_0$. And in the relatively weak signal setting with $1-\theta_j^+ = \theta_j^- = 20\%$, the sizes of $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$ can be slightly larger than $|\mathcal{A}_0|$ but still close to it.



(a) $K = 15$, multi-parameter SLAM (b) $K = 20$, multi-parameter SLAM

Figure 12: Coverage probabilities of the true patterns, from the approximate screening procedure under the multi-parameter SLAM. Boxplots are from 200 runs in each scenario.

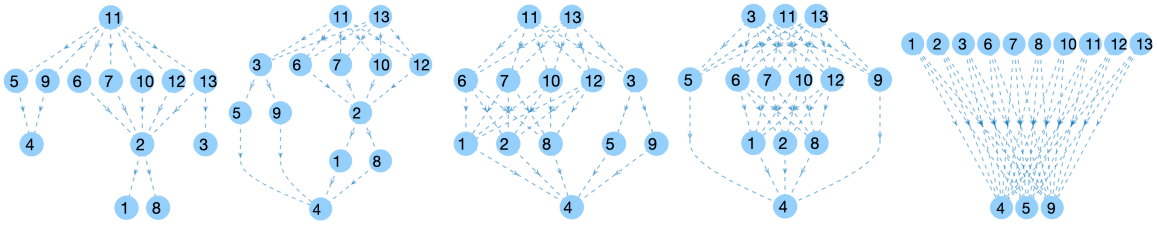


(a) $K = 15$, noise 20%; PEM (b) $K = 15$, noise 10%; PEM (c) $K = 20$, noise 20%; PEM (d) $K = 20$, noise 10%; PEM

(e) $K = 15$, noise 20%; FP-VEM (f) $K = 15$, noise 10%; FP-VEM (g) $K = 20$, noise 20%; FP-VEM (h) $K = 20$, noise 10%; FP-VEM

Figure 13: Sizes of the finally selected patterns $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$ under the two-parameter SLAM. The "noise" refers to the value of $1 - \theta_j^+ = \theta_j^-$. The number of true patterns is $|\mathcal{A}_0| = 10$.

TIMSS Data: Attribute structures corresponding to different Υ's. For the TIMSS data, we obtain those different attribute structures corresponding to different Υ's in the FP-VEM algorithm. The results are presented in Figure 14. Apart from the five structures shown in Figure 14(a)–(e), the two patterns selected when $\Upsilon \in [0.70, 0.74]$ are the all-zero and the all-one patterns, which do not result in any structure among the 13 attributes. Note that the structure in Figure 14(d) is equivalent to the structure selected by EBIC in Figure 9(b).



(a) $\Upsilon = 0.90$ (b) $\Upsilon = 0.88$ (c) $\Upsilon = 0.86$ (d) $\Upsilon \in [0.80, 0.84]$ (e) $\Upsilon \in [0.76, 0.78]$

Figure 14: Different attribute structures corresponding to various Υ's in Algorithm 2. Plot (d) here is equivalent to Figure 9(b), the attribute structure selected by EBIC.

References

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37: 3099–3132, 2009.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Animashree Anandkumar, Daniel Hsu, Majid Janzamin, and Sham Kakade. When are overcomplete topic models identifiable? Uniqueness of tensor tucker decompositions with structured sparsity. *Journal of Machine Learning Research*, 16:2643–2694, 2015.

Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Conference on Learning Theory*, pages 742–778, 2014.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Miguel A Carreira-Perpinán and Steve Renals. Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12(1):141–152, 2000.

Hanfeng Chen, Jiahua Chen, and John D Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(1):19–29, 2001.

Hanfeng Chen, Jiahua Chen, and John D Kalbfleisch. Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):95–115, 2004.

Jiahua Chen. Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, 23:221–233, 1995.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Yinghan Chen, Steven Andrew Culpepper, Yuguo Chen, and Jeffrey Douglas. Bayesian estimation of the DINA Q-matrix. *Psychometrika*, 83(1):89–108, 2018.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *arXiv preprint arXiv:1805.05054*, 2018.

Kyong Mi Choi, Young-Sun Lee, and Yoon Soo Park. What CDM can tell about what students have learned: An analysis of timss eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education*, 11(6), 2015.

Jimmy de la Torre. The generalized DINA model framework. *Psychometrika*, 76:179–199, 2011.

Jimmy de la Torre and Jeffrey A Douglas. Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3):333–353, 2004.

Jimmy de la Torre, L Andries van der Ark, and Gina Rossi. Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 51(4):281–296, 2018.

Lawrence T DeCarlo. On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35(1):8–26, 2011.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Louis V DiBello, William F Stout, and Louis A Roussos. Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cognitively diagnostic assessment*, pages 361–389, 1995.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. Algebraic factor analysis: tetrads, pentads and beyond. *Probability Theory and Related Fields*, 138(3-4):463–493, 2007.

Ivo Düntsch and Günther Gediga. Skills and knowledge structures. *British Journal of Mathematical and Statistical Psychology*, 48(1):9–27, 1995.

Christopher R Genovese and Larry Wasserman. Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28(4):1105–1127, 2000.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Yuqi Gu and Gongjun Xu. The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*, 84(2):468–483, 2019a.

Yuqi Gu and Gongjun Xu. Partial identifiability of restricted latent class models. *The Annals of Statistics*, forthcoming, 2019b.

Mats Gyllenberg, Timo Koski, Edwin Reilink, and Martin Verlaan. Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31:542–548, 1994.

Philippe Heinrich and Jonas Kahn. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844–2870, 2018.

Robert A. Henson, Jonathan L. Templin, and John T. Willse. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74:191–210, 2009.

Nhat Ho and XuanLong Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics*, 44(6):2726–2755, 2016.

C.C. Holmes and S.G. Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

Ariel Jaffe, Roi Weiss, Shai Carmi, Yuval Kluger, and Boaz Nadler. Learning binary latent variable models: A tensor eigenpair approach. *arXiv preprint arXiv:1802.09656*, 2018.

Brian W. Junker and Klaas Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25:258–272, 2001.

Joseph B Kruskal. 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, 1977.

Young-Sun Lee, Yoon Soo Park, and Didem Taylan. A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, 11(2):144–177, 2011.

Jacqueline P Leighton, Mark J Gierl, and Stephen M Hunka. The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41(3):205–237, 2004.

Eric Maris. Estimating multiple classification latent class models. *Psychometrika*, 64(2): 187–212, 1999.

Gunter Maris and Timo M Bechger. Equivalent diagnostic classification models. *Measurement*, 7:41–46, 2009.

Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

André A Rupp, Jonathan Templin, and Robert A Henson. *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, 2010.

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

P Smolensky. Chapter 6: information processing in dynamical systems: foundations of harmony theory. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1, 1986.

Yu-Law Su, KM Choi, WC Lee, T Choi, and M McAninch. Hierarchical cognitive diagnostic analysis for timss 2003 mathematics. *Centre for Advanced Studies in Measurement and Assessment*, 35:1–71, 2013.

Jonathan Templin and Laine Bradshaw. Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79(2): 317–339, 2014.

Jonathan L. Templin and Robert A. Henson. Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11:287–305, 2006.

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

Matthias von Davier. The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, 67(1):49–71, 2014.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, pages 339–362, 1995.

Zhenke Wu, Maria Deloria-Knoll, and Scott L Zeger. Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics*, 18(2):200–213, 2017.

Zhenke Wu, Livia Casciola-Rosen, Antony Rosen, and Scott L Zeger. A Bayesian approach to restricted latent class models for scientifically-structured clustering of multivariate binary outcomes. *arXiv preprint arXiv:1808.08326*, 2018.

Gongjun Xu. Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45:675–707, 2017.

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

Gongjun Xu and Stephanie Zhang. Identifiability of diagnostic classification models. *Psychometrika*, 81:625–649, 2016.

Kazuhiro Yamaguchi and Kensuke Okada. Comparison among cognitive diagnostic models for the timss 2007 fourth grade mathematics assessment. *PLoS ONE*, 12(2):e0188691, 2018.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, forthcoming, 2019.