A Joint MLE Approach to Large-Scale Structured Latent Attribute Analysis

Yuqi Gu a and Gongjun Xu b

a Department of Statistics, Columbia University, New York; b Department of Statistics, University of Michigan, Ann Arbor, MI

1. Introduction

1.1. A Modern Family of Fine-Grained Discrete Latent Variable Models

Structured latent attribute models (SLAMs) are discrete latent variable models that have attracted substantial attention in various applications, including cognitive diagnosis in educational assessments (Junker and Sijtsma 2001; Henson, Templin, and Willse 2009; de la Torre 2011), psychiatric diagnosis of mental disorders (Templin and Henson 2006; de la Torre, van der Ark, and Rossi 2018), and epidemiological studies of disease etiology (Wu, Deloria-Knoll, and Zeger 2017; O’Brien et al. 2019). A SLAM assumes multiple binary latent attributes explain observed variables in a highly structured fashion. In particular, for each subject $i$ a SLAM models the $J$-dimensional latent observations $r_i = (r_{i1}, \ldots, r_{iJ})$ using a $K$-dimensional latent attribute profile $a_i = (a_{i1}, \ldots, a_{iK}) \in \{0, 1\}^K$. In many applications, each attribute $a_{ik} = 1$ or 0 carries substantive meanings; for example, mastery/deficiency of some skill in an educational test, or presence/absence of some pathogen in epidemiological diagnosis. An important “structured” feature of a SLAM comes from a binary loading matrix, the Q-matrix (Tatsuoka 1983). The $J \times K$ matrix $Q = (q_{jk})$ encodes how the observed variables depend on the latent attributes, where $q_{jk} = 1$ or 0 means whether or not the $j$th observed variable depends on the $k$th latent attribute. By modeling the latent variables as multidimensional binary and incorporating structural constraints in the Q-matrix, SLAMs provide a powerful framework to infer subjects’ fine-grained latent traits, and to perform clustering based on the inferred latent profiles.

Since the latent variables are discrete, a SLAM can be viewed as a mixture model, where each subject’s latent attribute profile $a_i$ is a random variable following a categorical distribution with $|\{0, 1\}^K| = 2^K$ components. Over the past two decades when latent attribute models have attracted a great surge of interest, this perspective of treating subjects’ latent attributes as random effects is usually taken in the literature of modeling (von Davier 2008; Henson, Templin, and Willse 2009; de la Torre 2011), estimation (Chen et al. 2015; Xu and Shang 2018; Culpepper 2019; Gu and Xu 2019a), and study of model identifiability (Xu 2017; Fang, Liu, and Ying 2019; Gu and Xu 2020; Chen, Culpepper, and Liang 2020a). Taking this perspective, estimation is usually performed by maximizing the marginal likelihood. The corresponding estimators can be obtained via an EM algorithm for mixture models. But an obstacle to adopting such an approach in large-scale and high-dimensional data is that the number of latent patterns $2^K$ grows exponentially with the number of attributes $K$. This quickly becomes computationally cumbersome as $K$ grows large, which is commonly seen in modern large-scale assessment data. For example, the Trends in International Mathematics and Science Study (TIMSS) 2003 8th grade dataset available in the R package CDM involves $K = 13$ skill attributes, which gives rise to $2^{13} = 8192$ binary skill patterns.

1.2. The Joint MLE Approach

On the other hand, the joint maximum likelihood estimation (joint MLE) approach treats the subjects’ latent attributes $\{a_i : 1 \leq i \leq N\}$ as fixed effects and directly incorporates them...
into the likelihood as unknown parameters. This approach would naturally avoid the need to model the joint distribution of the exponentially many latent attribute configurations. For traditional problems, joint MLE was usually inconsistent when the sample size goes to infinity (large $N$) but the number of observed variables is fixed (fixed $J$) (Neyman and Scott 1948). But in modern large-scale educational assessments, data are collected in an ever-increasing scope involving many student test-takers (large $N$) and many test items (large $J$). For example, the TIMSS, a series of international assessments of the mathematics and science knowledge, involve students in over 50 countries and have nearly 800 assessment items in total (Mullis et al. 2016). This scope of data provides new opportunities and requires new methods and understanding of latent variable modeling.

The joint MLE’s unique feature of directly incorporating subjects’ latent attributes $a_i$’s as parameters to estimate has important and useful practical implications. In the applications of SLAMs to cognitive diagnosis (von Davier and Lee 2019), estimating each student’s latent skill profiles $a_i$ is of great interest as this can provide useful diagnosis of a student’s strengths and weaknesses to facilitate better follow-up instructions. However, most statistical developments of SLAMs (Chen et al. 2015; Xu 2017; Xu and Shang 2018; Gu and Xu 2019a) focused on the random-effect versions which marginalize out the $a_i$’s in the likelihood and focus on estimating other quantities, so their identifiability and estimation results do not apply to $\{a_i\}$. The important questions of what conditions can guarantee the $\{a_i\}$ is consistently estimable and how to estimate this for large-scale data remain unaddressed. The joint MLE approach considered in this work directly targets at estimating the unknown $\{a_i\}$ and $Q$, and we will use this framework to address the aforementioned questions.

Recently, for structured latent factor analysis with continuous latent variables, Chen, Li, and Zhang (2019) and Chen, Li, and Zhang (2020b) studied the joint MLE approach and established identifiability and estimability of continuous latent factors in the double asymptotic regime when $N$ and $J$ both go to infinity. However, SLAMs form a different landscape with all the latent variables being discrete. Establishing theory for statistical estimability and consistency for discrete latent variables in full generality requires different arguments from those in Chen, Li, and Zhang (2019, 2020b). In addition, new computational methods need to be developed to address the unique challenge of estimation with a large number of discrete latent attributes.

1.3. Our Contributions

We investigate the joint MLE approach to large-scale structured latent attribute analysis, and make the following theoretical and methodological contributions.

1. We consider the triple-asymptotic regime where all of the $N$, $J$, and $K$ can grow to infinity, for the first time in the literature of SLAMs. In this scenario, we establish the estimability and consistency of both the binary factor loadings in the $Q$-matrix and the latent attribute profiles of the subjects $\{a_{ijk}\}$. We also derive finite-sample error bounds for the considered estimators.

2. We propose a scalable approximate algorithm to compute the joint MLE for two-parameter SLAMs (defined in Example 1). We also propose an efficient two-step estimation procedure for general multi-parameter SLAMs (defined in Example 2). This two-step procedure is inspired by investigating a common and interesting type of model oversimplification of SLAMs. When misspecifying a general multi-parameter SLAM to the two-parameter submodel, we show the oversimplified joint MLE can consistently recover part, or even all, of the latent structure under certain conditions.

The rest of the article is organized as follows. Section 2 introduces the setup of SLAMs and discusses its connections with other latent variable models. Section 3 defines the joint MLE and studies its statistical properties. Section 4 proposes scalable algorithms for computing the joint MLE. Section 5 provides simulation studies and Section 6 applies our method to a dataset from the TIMSS 2011 Austrian assessment. Section 7 gives a discussion. Technical proofs and additional discussion on computation are included in the supplementary material.

2. Setup of SLAMs

2.1. General Formulation and Concrete Examples

In this article, we focus on SLAMs for multivariate binary data, which are ubiquitously encountered in educational assessments (correct/wrong answers), social science survey responses (yes/no responses), or biomedical and epidemiological diagnostic tests (positive/negative results). For $N$ subjects and $J$ variables, collect the observed data in an $N \times J$ binary matrix $R = (r_{ij})$, where $r_{ij} = 1$ or 0 denotes whether the $i$th subject gives a positive response to the $j$th variable. Suppose there are $K$ binary latent attributes, then the $J \times K$ binary loading matrix $Q = (q_{jk})$ encodes how the $J$ observed variables depend on the $K$ latent attributes. The $N \times K$ binary matrix $A = (a_{ijk})$ that stores the latent attribute profiles for the $N$ subjects. Both $Q$ and $A$ have binary entries, where $q_{jk} = 1$ or 0 represents whether the $j$th test item depends on the $k$th latent attribute, and $a_{ijk} = 1$ or 0 represents whether the $i$th individual possesses the $k$th attribute. Generally, a SLAM is a probabilistic model with discrete structures $Q, A$, and additional continuous parameters to specify the generative process of the response data $R$.

Denote the additional continuous parameters needed to complete the model specification by $\Theta = \{\theta_1, \ldots, \theta_J\}$. Each observed variable $j$ has its continuous parameter vector which we generically denote by $\theta_j$, whose form depends on the specific model and will be made concrete in Examples 1 and 2. Each observed $r_{ij}$ follows a Bernoulli distribution with parameter $f(a_{ij}, q_{ij}, \theta_j)$ as a function of $a_{ij}, q_{ij}$, and $\theta_j$. Given the subjects’ latent attribute matrix $A$, binary loading matrix $Q$, and parameters $\Theta$, the observed responses are assumed to be conditionally independent. In summary, a SLAM postulates the following statistical model,

$$
(r_{ij}|A, Q, \Theta) \sim \text{Bernoulli}(f(a_{ij}, q_{ij}, \theta_j)),
$$

$$
\mathbb{P}(R|A, Q, \Theta) = \prod_{i=1}^{N} \prod_{j=1}^{J} f(a_{ij}, q_{ij}, \theta_j)^{r_{ij}} \left(1 - f(a_{ij}, q_{ij}, \theta_j)\right)^{1-r_{ij}}.
$$
A visualization of a SLAM as taking $\mathbf{A}$ and $\mathbf{Q}$ as input in a probabilistic model and then generating the data $\mathbf{R}$. The $K$ equals 2 in the figure. The entry $r_{ij}$ follows a Bernoulli distribution with parameter $f(a_{ij}, q_{ij}, \theta_j)$, which is a function of the $i$th row of $\mathbf{A}$ (denoted by $a_{ij}$), the $j$th row of $\mathbf{Q}$ (denoted by $q_{ij}$), and continuous parameters $\theta_j$.

Figure 1 gives a visualization of a SLAM, making clear how the unknown binary matrices $\mathbf{A}$ and $\mathbf{Q}$ underlie the data generating process. In this article, we treat both $\mathbf{A}$ and $\mathbf{Q}$ as unknown fixed parameters and consider the large-scale scenarios where the number of subjects $N$, the number of observed variables $J$, and the number of latent attributes $K$ all can go to infinity, that is, a triple-asymptotic regime.

We next review two main types of SLAMs widely adopted in the cognitive diagnostic modeling literature: the two-parameter models and the multiparameter models.

**Example 1 (Two-Parameter SLAMs).** For each item $j$, a two-parameter SLAM compactly uses two distinct Bernoulli parameters to model $r_{ij}$, with $\theta_j = (\theta_j^+, \theta_j^-)$. There are two different types of two-parameter SLAMs, the deterministic input noisy output “And” (DINA) model proposed in Junker and Sijtsma (2001), and the deterministic input noisy output “Or” (DINO) model proposed in Templin and Henson (2006). Under DINA and DINO models, the Bernoulli parameter $f(a_{ij}, q_{ij}, \theta_j)$ in Equation (1) takes the following specific forms,

\[
\begin{align*}
    f^{\text{DINA}}(a_{ij}, q_{ij}, \theta_j) &= \frac{\theta_j^+}{\theta_j^-} \quad \text{if } a_{ij} = 1, \\
    f^{\text{DINO}}(a_{ij}, q_{ij}, \theta_j) &= \frac{\theta_j^+}{\theta_j^-} \quad \text{if } a_{ij} = 1, \quad q_{ij} = 1.
\end{align*}
\]

DINA is often used in educational testing with latent skills as attributes, and DINO often in psychiatric diagnosis with mental disorders as attributes (de la Torre, van der Ark, and Rossi 2018).

Chen et al. (2015) established duality between the DINA and DINO models with $P(r_{ij} = 1 | a_{ij} = \alpha, q_{ij}, \theta_j, \text{DINO}) = 1 - P(r_{ij} = 1 | a_{ij} = \alpha, q_{ij}, \theta_j, \text{DINA})$ for any $\alpha \in \{0, 1\}^K$, where $1_K$ is a K-dimensional all-one vector. Thanks to this duality, identifiability and estimation results developed under DINA easily carry over to the DINO case. So without loss of generality, next we focus on the DINA model when studying two-parameter SLAMs.

**Example 2 (Multiparameter SLAMs).** Unlike a two-parameter model, a multiparameter SLAM models each observed variable $j$ using potentially more than two Bernoulli parameters. The $f(a_{ij}, q_{ij}, \theta_j)$ in Equation (1) now takes the form

\[
\begin{align*}
    f^{\text{mult}}(a_{ij}, q_{ij}, \theta_j) &= f(\mu_{j,0} + \sum_{k=1}^{K} \mu_{j,k}(q_{jk} a_{jk})) \\
    &+ \sum_{1 \leq k_1 < k_2 \leq K} \mu_{j,k_1 k_2}(q_{jk_1} a_{jk_1})(q_{jk_2} a_{jk_2}) \\
    &+ \cdots + \mu_{j,12...K} \prod_{k=1}^{K}(q_{jk} a_{jk}),
\end{align*}
\]

where different link functions $f(\cdot)$ lead to different specific models; when $f(\cdot)$ is the identity, Equation (3) gives the Generalized DINA model (GDINA, de la Torre 2011); when $f(\cdot)$ is the logistic function, Equation (3) gives the Log-linear Cognitive Diagnosis Models (LCDMs, Henson, Templin, and Willse 2009); see also the General Diagnostic Models (GDMs, van Dijker 2008). Note that in (3), not all the $\mu$-coefficients are meaningful and need to be incorporated into the model; for example, if $q_{jk} = 0$ then $\mu_{j,k}$ is not needed and if $q_{jk_1} q_{jk_2} = 0$ then $\mu_{j,k_1 k_2}$ is not needed, etc. In multiparameter SLAMs, the continuous parameter vector $\theta_j$ is the collection of all the meaningful $\mu_{j,k}$-coefficients. Multi-parameter models under Equation (3) are quite general, as they incorporate all the possible main and interaction effects of the meaningful latent attributes.

Examples 1 and 2 imply that the two-parameter model can be viewed as a submodel of the multiparameter model. To see this, just set all the $\mu$-coefficients in Equation (3) to zero except $\mu_{j,0}$ and the highest order term $\mu_{j,12...K} := \mu_{j,k|k=1}$, then $\theta_j^- = f(\mu_{j,0})$ and $\theta_j^+ = f(\mu_{j,0} + \mu_{j,12...K})$ correspond to the two parameters for variable $j$ defined in Example 1.

### 2.2. Connections Between SLAMs and Other Latent Variable Models

We briefly review the family of latent variable models and locate SLAMs within this context. Latent variable models can be categorized into four types according to the nature of the observed and the latent variables. When the observed and latent
variables are both continuous, the factor analysis (Anderson and Rubin 1956) has been widely used. When the observed variables are discrete but the latent variables are continuous, the item response theory (IRT) models (Embretson and Reise 2009) are typical modeling choices. On the other hand, to model continuous observed data using a discrete latent variable, researchers have employed mixture models such as the Gaussian mixtures (Reynolds, Quatieri, and Dunn 2000). Finally, when both the observed variables and the latent one are discrete, the latent class model has been a popular modeling tool since decades ago (Lazarsfeld and Henry 1968).

SLAMs can be viewed as a modern generalization of latent class models (LCMs), in that both adopt discrete latent structure to model discrete data. Despite this similarity, the following two key characteristics distinguish SLAMs from traditional LCMs: (a) the discrete latent constructs in SLAM are multidimensional instead of unidimensional as in LCMs; and (b) a SLAM models dependence of the observed variables on the latent ones by a binary loading matrix. Figure 2 provides graphical model representations of LCMs and SLAMs that highlight their connections and differences. Both LCMs and SLAMs assume the multivariate categorical observations \( r_i = (r_{i1}, \ldots, r_{ij}) \) are conditionally independent given the latent part. When modeling the observed \( r_i \), an LCM in Figure 2(a) adopts a unidimensional latent variable \( z_i \in \{1, \ldots, C\} \) while a SLAM in Figure 2(b) adopts a \( K \)-dimensional binary latent vector \( a_i = (a_{i1}, \ldots, a_{ik}) \in \{0,1\}^K \). Therefore, an LCM does not necessarily distinguish the \( C \) latent classes by definition, while a SLAM naturally defines \( 2^K \) distinct latent classes, each as a pattern detailing the statuses of \( K \) fine-grained traits.

Additionally and perhaps more importantly, a SLAM has the key \( 1 \times K \) binary loading matrix \( Q = (q_{jk}) \), where \( q_{jk} = 1 \) means observed \( r_{ij} \) depends on the latent \( a_{ik} \) and \( q_{jk} = 0 \) otherwise. Such dependence encoded in the \( Q \)-matrix can represent practitioners’ prior knowledge, facilitate dimension reduction, and enhance interpretability. In summary, SLAMs enable uncovering hidden fine-grained scientific information, providing model-based clustering of subjects, and facilitating better intervention. These advantages distinguish SLAMs from traditional models such as IRT models or LCMs, and make SLAMs and their variants suitable for a variety of modern applications, including not only education and psychology (Chen et al. 2015; Xu and Shang 2018; Gu and Xu 2019a), but also epidemiology (O’Brien et al. 2019) and biomedicine (Ni, Müller, and Ji 2020; Chen, Zeng, and Wang 2021).

3. Joint MLE and Its Statistical Properties

3.1. Definition of Joint MLE

We next formally introduce the joint maximum likelihood estimator. Under the general setup in Equation (2), the log of the joint likelihood of \((A, Q, \Theta)\) is

\[
\ell^m(A, Q, \Theta | R) = \sum_{j=1}^{N} \sum_{i=1}^{J} \left[ r_{ij} \log(f^m(a_i, q_j, \theta_j)) + (1 - r_{ij}) \log(1 - f^m(a_i, q_j, \theta_j)) \right],
\]

where the superscript "m" denotes a specific model, for example, a two-parameter or multiparameter model reviewed in Examples 1 and 2. The joint MLE approach has an important feature that the subjects’ latent attributes \( A = (a_{ik}) \) are incorporated as unknown parameters to estimate. This is different from the marginal MLE which marginalize out the \( a_{ik} \)'s and focus on estimating other quantities. Indeed, in the applications of SLAMs...
to cognitive diagnostic modeling (von Davier and Lee 2019), inferring the students’ latent skill profiles is of great interest as they can provide useful diagnosis of a student’s strengths and weaknesses for navigating better follow-up instructions. However, most statistical developments of SLAMs (Chen et al. 2015; Xu 2017; Xu and Shang 2018; Gu and Xu 2019a) focused on the random-effect versions, so their results typically do not apply to the underlying A. The important questions of what conditions can guarantee the A is consistently estimable and how to estimate it for large-scale data remain unaddressed. To this end, the joint MLE approach considered in this work directly targets at estimating the unknown quantities A and Q and provides a natural basis for addressing these questions.

Given the general log-likelihood in Equation (4), define the joint MLE under a specific SLAM as

$$\hat{A}, \hat{Q}, \hat{\Theta} = \operatorname{arg max}_{(A, Q, \Theta)} \ell_m (R; A, Q, \Theta)$$

subject to fitting a K-attribute specified SLAM with

$$\sum_{k=1}^{K} \hat{q}_{j,k} \leq B_j,$$

where B_j’s are prespecified upper bounds depending on the model, imposed for theoretical identifiability reasons. An interesting study Bonhomme and Manresa (2015) also considered the fixed-effect estimation of discrete latent heterogeneity, motivated by panel data in econometrics. In the regime where the number of subjects N and number of time points T both go to infinity, Bonhomme and Manresa (2015) considered continuous data and unidimensional discrete heterogeneity. Different from that, in this work when N and T go to infinity, we consider multivariate categorical data and multidimensional discrete latent features. Therefore, the least-square estimation criterion used in Bonhomme and Manresa (2015) is not applicable here, and we need to seek estimators based on the specific likelihood functions.

We next make several important remarks about the nuances and differences between estimating two-parameter and multi-parameter SLAMs, in terms of both identifiability and computation.

Remark 1 (Solve (5) under a two-parameter model). Based on the setup in Example 1, the two-parameter log-likelihood in Equation (5) can be written in the following explicit form:

$$\ell_{\text{two}} (Q, A, \Theta | R) = \sum_{i=1}^{N} \sum_{j=1}^{J} r_{ij} \left( \prod_{k=1}^{K} a_{i,k}^{\hat{q}_{j,k}} \log \hat{\theta}_{j}^{+} \right) + (1 - \prod_{k=1}^{K} a_{i,k}^{\hat{q}_{j,k}}) \log \hat{\theta}_{j}^{-} \right) + (1 - r_{ij}) \left( \prod_{k=1}^{K} a_{i,k}^{\hat{q}_{j,k}} \log(1 - \hat{\theta}_{j}^{+}) \right) + (1 - \prod_{k=1}^{K} a_{i,k}^{\hat{q}_{j,k}}) \log(1 - \hat{\theta}_{j}^{-}) \right] \right).$$

Under the two-parameter likelihood, when solving Equation (5) for $$(A, Q, \Theta)_{\text{two}}$$, we impose a natural constraint $\hat{\theta}_{j}^{+} > \hat{\theta}_{j}^{-}$ to ensure identifiability (Junker and Sijsma 2001; Gu and Xu 2019b). A careful inspection of the special combinatorial form under the two-parameter DNA model (7) reveals that the upper bound $B_j$ in the optimization problem (5) can be taken as $B_1 = \cdots = B_j = B_{\text{two}} = \infty$. That is, there is essentially no need to constrain the number of “1’s in the estimation of Q. To solve Equation (5) under the two-parameter likelihood, we propose a scalable approximate EM-flavor algorithm. This algorithm treats the unknown discrete structures $Q$, A) as missing data to impute in an approximate E (Expectation) step which is based on a few Gibbs samples, and treats continuous parameters $\Theta = (\theta_j^+\theta_j^- : 1 \leq j \leq J)$ as model parameters to update in an M (Maximization) step; see Section 4.1 for details.

Remark 2 (Solve Equation (5) under a multiparameter model). Under a multiparameter SLAM, denote the true binary loading matrix by $Q_{\text{true}} = (q_{j,k}^{\text{true}})$, and we take the upper bound in (6) to be $B_j = B_j^{\text{mult}} = \sum_{k=1}^{K} q_{j,k}^{\text{true}}$. Under a multiparameter model (3), Q captures the sparsity structure of the underlying continuous parameters $\mu_j$, so the constraint $\sum_{k=1}^{K} \hat{q}_{j,k} \leq B_j = \sum_{k=1}^{K} q_{j,k}^{\text{true}}$ in Equation (6) resembles the $L_0$ constraint on regression coefficients in regression problems for variable selection. Theoretically, such a constraint is necessary to ensure $q_{j,k}$’s are identifiable under a multiparameter SLAM. To see this, consider a toy example with $q_{j,k}^{\text{true}} = (1, 0)$, then the multiparameter model with an identity link in Example 2 gives

$$P(r_{ij} = 1 | a_i, q_{j,k}^{\text{true}}, \mu_j) = \mu_{j,0} + \mu_{j,1} q_{j,k}^{\text{true}}, \mu_{j,1}, \mu_{j,2}, \mu_{j,12}) \text{ where } \mu_{j,2} = 0,$$

while with an alternative $\hat{q}_{j} = (1, 1)$ and $\hat{\mu}_j = (\hat{\mu}_{j,0}, \hat{\mu}_{j,1}, \hat{\mu}_{j,2}, \hat{\mu}_{j,12})$ where $\hat{\mu}_{j,2} = 0$,

$$P(r_{ij} = 1 | a_i, \hat{q}_{j}, \hat{\mu}_j) = \mu_{j,0} + \mu_{j,1} \hat{q}_{j,1} a_{i,1} + \mu_{j,2} \hat{q}_{j,1} a_{i,2} + \hat{\mu}_{j,12} (\hat{q}_{j,1} a_{i,1}) (\hat{q}_{j,2} a_{i,2})$$

$$= \mu_{j,0} + \mu_{j,1} a_{i,1}.$$

if $\hat{\mu}_{j,0} = \mu_{j,0}$ and $\hat{\mu}_{j,1} = \mu_{j,1}$.

This example illustrates that despite $q_{j,k}^{\text{true}} = (1, 0) \neq \hat{q}_{j} = (1, 1)$, the distribution of $r_{ij}$ given the two are identical, indicating non-identifiability. Therefore, theoretically, we need to constrain the number of “1’s in Q for identifiability when the model is multi-parameter.

Although the constraint $\sum_{k=1}^{K} \hat{q}_{j,k} \leq \sum_{k=1}^{K} q_{j,k}^{\text{true}}$ is needed for theoretical identifiability under multiparameter models as stated above, practically, the constrained optimization problem (6) can be replaced by an unconstrained one by imposing an appropriate penalty. Indeed, our estimation method for multiparameter SLAMs does not assume knowledge of the true values of $\sum_{k=1}^{K} q_{j,k}^{\text{true}}$, but rather adopts marginal screening and variable selection approaches to directly estimate the entries of Q in a second regression stage, following a first stage of approximate estimation of latent attributes in A; see Section 4.2 for details.
3.2. Theoretical Properties of the Joint MLE

From now on, we consider the model sequence indexed by \((N, J, K)\), where each of \(N\), \(J\), and \(K\) can go to infinity. Thus, far we have treated \(\Theta\) as a generic notation for continuous parameters in any specific SLAM. For technical convenience, we next fix the notation of \(\Theta\) as a \(J \times 2^K\) matrix \(\Theta = (\theta_{j,\alpha})\), where

\[
\theta_{j,\alpha} = P(ri_j = 1|a_i = \alpha, q_j, \text{ specific model}) \tag{8}
\]

for \(j \in [J]\) and \(\alpha \in \{0, 1\}^K\). The expressions of \(\theta_{j,\alpha}\) under specific two- or multiparameter models can be easily derived based on Examples 1–2. The following assumptions are made on the true parameters \((\Theta_{\text{true}}, Q_{\text{true}}, A_{\text{true}})\) that generate the data:

**Assumption 1.** There exists a finite number \(d \geq 2\) such that

\[
\frac{1}{d} \leq \min_{1 \leq j \leq J} \theta_{j,\alpha}^{\text{true}} \leq \max_{1 \leq j \leq J} \theta_{j,\alpha}^{\text{true}} \leq 1 - \frac{1}{J}. \tag{9}
\]

**Assumption 2.** For two-parameter SLAMs, suppose \(\theta_{j,\alpha}^{+,\text{true}} > \theta_{j,\alpha}^{-,\text{true}}\) for each \(j\) and that there exists \(\{\beta_j\} \subseteq (0, \infty)\) such that

\[
\min_{1 \leq j \leq J} \left( \theta_{j,\alpha}^{+,\text{true}} - \theta_{j,\alpha}^{-,\text{true}} \right)^2 \geq \beta_j, \tag{10}
\]

For multiparameter SLAMs, there exists \(\{\beta_j\} \subseteq (0, \infty)\) such that

\[
\min_{1 \leq j \leq J} \left\{ \min_{\alpha \circ q_j \neq \hat{q}_j} \left( \theta_{j,\alpha}^{+,\text{true}} - \theta_{j,\alpha}^{-,\text{true}} \right)^2 \right\} \geq \beta_j, \tag{11}
\]

where \(\alpha \circ q_j = (\alpha_1q_j, \ldots, \alpha_Kq_j)\) denotes element-wise product of binary vectors \(\alpha\) and \(q_j\).

**Assumption 3.** There exist \(\{\delta_j\}\), \(\{p_N\} \subseteq (0, \infty)\) and a constant \(\epsilon > 0\) such that

\[
\min_{1 \leq k \leq K} \frac{1}{N} \sum_{j=1}^{J} I(q_j^{\text{true}} = \epsilon_k) \geq \delta_j; \tag{12}
\]

\[
\min_{\alpha \in \{0, 1\}^K} \frac{1}{N} \sum_{i=1}^{N} I(\alpha_i^{\text{true}} = \alpha) \geq p_N \geq \frac{\epsilon}{2^K}. \tag{13}
\]

Also assume \(\sum_{k=1}^{K} q_{j,k}^{\text{true}} \leq K_0\) for a constant \(K_0\).

Note that by writing all the lower bounds in the above assumptions as depending on a subscript \(J\) or \(N\), we indeed allow them to go to zero as \(J\) and \(N\) go to infinity. This type of assumptions distinguish the current theoretical investigation from all the previous works on SLAMs (e.g., Chen et al. 2015; Xu and Shang 2018; Gu and Xu 2019a). As to be shown in the following theorems, as long as the rate at which these \(\beta_j, \delta_j, \text{ and } p_N\) go to zero satisfy some mild requirements, consistency of joint MLE can be ensured.

More detailed discussions on the assumptions are in order. **Assumption 1** is a very mild condition on the Bernoulli parameters \(\theta_{j,\alpha}\’s\). **Assumption 2** lower bounds the gap of Bernoulli parameters for different latent classes, under the two- and multiparameter SLAM, respectively. Such a gap \(\beta_j\) measures how separated the latent classes are and hence quantifies how strong the signals are. This assumption has its counterpart in the finite-\(J\) regime; for example, Xu and Shang (2018) and Gu and Xu (2019b, 2020) imposed \(\beta_j > 0\) when studying identifiability. Instead, here we allow \(\beta_j \to 0\) and establish estimability and consistency. **Assumption 3** is about the discrete structures \(Q\) and \(A\), where Equation (12) resembles a requirement that \(Q\) should contain an identity submatrix \(I_K\) in the studies with finite \(J\) (Chen et al. 2015; Xu and Shang 2018). Here with \(J \to \infty\), a finite number of submatrices \(I_K\) in \(Q\) may not suffice for estimability and consistency, and Equation (12) requires \(Q\) to contain an increasing number of \(I_K\’s\) as \(J\) grows. In the literature, Wang and Douglas (2015) made a similar assumption on \(Q\) when establishing consistency of a nonparametric estimator for cognitive diagnostic models, and Chen, Li, and Zhang (2020b) also imposed a similar requirement on the loading matrix when studying continuous latent factor models. Note that this requirement (12) implies the matrix \(Q\) does not contain any all-zero column. Theoretically, if \(Q\) contains some all-zero column, then the model is not identifiable. This is because dropping this all-zero column of \(Q\) and reducing the number of latent attributes by one will give the same distribution of the observables. As for another requirement (13) in **Assumption 3**, it implies the \(2^K\) latent patterns do not exhibit too uneven proportions in the sample. A resemblance for this requirement in random-effect SLAMs is \(p_\alpha > 0\) for all \(\alpha \in \{0, 1\}^K\), where \(p_\alpha\) denotes the population proportion of latent pattern \(\alpha\).

Denote \(M = (NJ)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{J} P(ri_j = 1|\text{true model})\), the average positive response rate in the sample. The following main theorem establishes the consistency and bounds the rate of convergence of joint MLE in recovering the latent structure.

**Theorem 1** (Consistency of joint MLE under either two- or multiparameter model). Consider either a two-parameter or a multiparameter SLAM with \((Q, A)\) obtained from solving Equation (5). When \(N, J \to \infty\), suppose \(\sqrt{J} = O(\sqrt{MN^{1-c}})\) for some small constant \(c \in (0, 1)\) and \(K = o(MJ \log J))\). Under Assumptions 1–3, the following two conclusions hold:

(a) There is

\[
\frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \left( \frac{P(ri_j = 1|Q^{\text{true}}, A^{\text{true}}, \Theta^{\text{true}})}{\hat{Q} \hat{A} \Theta^{\text{true}}} - P(ri_j = 1|Q, A, \Theta^{\text{true}}) \right)^2 \geq o_p \left( \frac{\gamma_j}{\beta_j} \right), \tag{14}
\]

where for a small positive constant \(\epsilon > 0\),

\[
\gamma_j = \frac{(\log J)^{1+c}}{J} \cdot \sqrt{M \log (2^K)}. \tag{15}
\]

(b) Up to a permutation of the \(K\) latent attributes, there is

\[
\frac{1}{J} \sum_{j=1}^{J} I(q_j^{\text{true}} \neq \hat{q}_j) = o_p \left( \frac{\gamma_j}{\beta_j} \right), \tag{16}
\]

\[
\frac{1}{N} \sum_{i=1}^{N} I(\alpha_i^{\text{true}} \neq \alpha_i) = o_p \left( \frac{\gamma_j}{\beta_j} \right). \tag{16}
\]
Remark 3. For large-scale continuous latent factor analysis, Chen, Li, and Zhang (2019, 2020b) exploited the low-rank matrix structure to establish consistency and bound convergence rate of MLE. In a continuous latent factor model, the low-rankness usually exactly captures the intrinsic characteristic of the model; for example, the latent structure is summarized as an inner product term $\Theta_{N \times K}A_{J \times K}^T$ in Chen, Li, and Zhang (2020b) (where $\Theta_{N \times K}$ collect the continuous person-parameters and $A_{J \times K}$ collect the continuous item-parameters), which is a matrix with low rank $K$. However, for discrete latent variable models, especially the complicated SLAMs considered here, the low-rankness is often a too rough and sometimes imprecise summary of the latent structure. This is because discrete latent structure ($A$ and $Q$ here) would induce an unobserved partition of data underlying a probabilistic model, which is not the case when latent variables are continuous and hence requires different analysis.

Remark 4. The proof of part (a) of Theorem 1 uses a similar technique as the profile likelihood approach in the network community detection literature (see, e.g., Choi, Wolfe, and Airoldi 2012; Zhao, Bickel, and Weko 2020). A proof technique of a similar spirit is useful here because the existence of discrete latent variables allows reformulating the maximum likelihood problem (5) as performing certain model-based clustering. Indeed, each vector $q_j$ categorizes the attribute patterns into distinct clusters locally for each observed variable $j$, in different ways under different model assumptions in Examples 1 and 2. Notably, also apparent from Examples 1 and 2 is that the model setup of a SLAM is fundamentally different from a stochastic block model for network data. The unobserved partition in SLAMs are more subtle to deal with than other simpler discrete latent variable models (including the network community models) due to the parameter constraints imposed by the $Q$-matrix. The overall proof procedure used to establish Theorem 1 needs to take into account such unique parameter constraints.

The $\gamma_j$ in part (a) of Theorem 1 bounds the error of recovering the average positive response probability under the estimated $(\hat{Q}, \hat{A})$. Part (b) further separately bounds the errors of the estimators for $\hat{Q}$ and $\hat{A}$, respectively. The derived rates in Theorem 1(b) imply that the sequences $\{\beta_j\}$, $\{p_{N,j}\}$, and $\{\delta_j\}$ are allowed to go to zero while still guaranteeing consistency, as long as $\gamma_j/(\beta_j \cdot p_{N,j}) \to 0$ and $\gamma_j/(\beta_j \cdot \delta_j) \to 0$. Theorem 1(b) not only ensures the asymptotic consistency of joint MLE, but also offers insight into the accuracy of estimating $Q$ and $A$ with finite samples and finite $J$. In particular, if $\beta_j$ and $\delta_j$ are constants and $K$ is finite, then the finite sample error bounds in Equation (16) become $O(\log J)^{1+\epsilon} \cdot J^{-1/2}$.

### 4. Scalable Estimation Algorithms

This section presents algorithms for computing the joint MLE for two-parameter and multiparameter SLAMs. Recall that a two-parameter SLAM can be viewed as a submodel for a multiparameter one. The succinct form of two-parameter models allows for developing a scalable approximate estimation approach, and the next Section 4.1 proposes an algorithm specifically tailored for two-parameter models. Then Section 4.2 builds on this algorithm and further provides an estimation approach for the more general multiparameter models.

#### 4.1. Estimation Under the Two-Parameter Model

EM algorithms (Dempster, Laird, and Rubin 1977) are popular methods for latent variable model estimation. For SLAMs, a traditional EM algorithm for computing the marginal MLE under a random-effect model assumes each $a_i$ follows a categorical distribution with $|\{0,1\}^K| = 2^K$ components. In this setup, the E step updates the probabilities of each $a_i$ being each possible pattern in $|0,1|^K$. The cardinality of this space $|\{0,1\}^K| = 2^K$ grows exponentially with $K$, so evaluating all the $a_i$ and $q_j$’s probabilities of being all the possible configurations has complexity $O((N + J)2^K)$ in each EM iteration. This incurs high computational cost for moderate to large $K$. On the other hand, here we consider the joint MLE for fixed-effect SLAMs and treat subjects’ latent attributes $a_{i,k}$’s and also $q_{j,k}$’s as parameters. This formulation requires different estimation procedures from the traditional EM for computing the marginal MLE. We next propose a new EM-flavor algorithm with a stochastic component well suited to the considered scenario. The new algorithm directly targets at estimating the individual $a_{i,k}$’s and $q_{j,k}$’s, and further uses a stochastic step in order to scale to high-dimensional data. In particular, our new algorithm draws a few Gibbs samples of the entries of the discrete $Q$ and $A$ in an approximate E step to achieve scalability.

The details of the algorithm are as follows. The entries of $Q$, $A$ are treated as missing data to be imputed in an approximate E step, and the continuous $\Theta = (\theta^+,\theta^-)$ are treated as model parameters to be updated in an M step. In the approximate E step, we propose to take an approximation by drawing a few (denote the number by $C$) Gibbs samples of entries of $A = (a_{i,k})$ (along the direction of updating subjects’ patterns), and then draw $C$ Gibbs samples of entries of $Q = (q_{j,k})$ (along the direction of updating variables’ loadings). Under the two-parameter log-likelihood in Equation (7), given the current iterates of parameters $(\theta^+,\theta^-)$, the conditional distributions of each $a_{i,k}$ and $q_{j,k}$ from which we draw the Gibbs samples are

$$P(a_{i,k} = 1|\cdot) = \sigma \left( - \frac{1}{N} \sum_{i=1}^{N} \sum_{m=1}^{K} \sum_{n=1}^{2^K} a_{i,m}^{q_{j,m}} r_{i,j} \log \frac{\theta^+}{\theta^-} + (1 - r_{i,j}) \log \frac{1 - \theta^+}{1 - \theta^-} \right),$$

$$P(q_{j,k} = 1|\cdot) = \sigma \left( \frac{1}{C} \sum_{i=1}^{C} \left( a_{i,k} \prod_{m=1}^{K} a_{i,m}^{q_{j,m}} r_{i,j} \log \frac{\theta^+}{\theta^-} + (1 - r_{i,j}) \log \frac{1 - \theta^+}{1 - \theta^-} \right) \right),$$

where $\sigma(x) = \exp(x)/(1 + \exp(x))$ denotes the sigmoid function. In approximate E step in the $n$th iteration,
after drawing \(C\) Gibbs samples \(A^{(t),1}, \ldots, A^{(t),C}\) and \(Q^{(t)}, \ldots, Q^{(t),C}\), we take a stochastic approximation of \(A\) in the following manner,

\[
A^{\text{ave},(t)} \leftarrow (1 - t^{-1}) A^{\text{ave},(t-1)} + t^{-1} \sum_{c=1}^{C} A^{(t),c}/C, \quad (17)
\]

where \(A^{\text{ave},(t-1)}\) denotes the \(A\)-matrix averaged from all the previous iterations up to iteration \(t-1\). The update (17) uses a similar idea to the stochastic approximation EM (SAEM) algorithm in Deloyer, Lavielle, and Moulines (1999). For \(Q\), we define \(Q^{(t)} = I(\sum_{c=1}^{C} Q^{(t),c}/C > 1/2)\); that is, the average \(Q\) obtained from the \(C\) Gibbs samples is rounded element-wise to the nearest integer (0 or 1) to give \(Q^{(t)}\). Then in the M step, fixing the current \(A^{\text{ave},(t)}\) and \(Q^{(t)}\), we can update the item parameters \(\theta^+\) and \(\theta^-\) in closed forms under the two-parameter model. We call such an algorithm EM with Alternating Direction Gibbs EM (ADG-EM) as each E step iteratively draws Gibbs samples of discrete latent structures \(A\) and \(Q\). Preliminary simulations show that drawing \(C < 10\) Gibbs samples in each E step usually suffices for good performance. The steps of ADG-EM are summarized in Algorithm 1. The Supplementary Material includes simulation studies assessing the convergence behavior of this algorithm.

We make a remark on the stochastic approximation step of the proposed algorithm. As briefly mentioned before, computation can be challenging for data with large \(N\), \(J\), and \(K\), because \(Q\) and \(A\) will be huge matrices with complex dependencies. But if we think of entries of \(Q\) and \(A\) individually in a Bayesian fashion, then each entry follows a Bernoulli distribution \(a \text{ posteriori}\) and indeed has an analytic posterior that is easy to sample from. With this thinking, our stochastic approximation procedure relies on a few Gibbs steps to achieve scalability. Such a procedure is specifically motivated by the multidimensional binary nature of the latent structures.

Our Algorithm 1 applies the stochastic approximation to updating \(A\) but not to that of \(Q\) in each iteration; that is, the update \(Q = I(Q^{\text{sum}}/C > 1/2)\) element-wisely in Algorithm 1 does not depend on the iteration number \(t\), in contrast to (17). We find through simulations that this algorithm has good estimation accuracy in various cases including when \(N\) and \(J\) are both very large. But one could similarly apply the stochastic approximation to both \(Q\) and \(A\); we present this modified version as Algorithm S.1 in the supplementary material.

In terms of computational complexity, Algorithm 1 has \(O((N + J)K)\) complexity in each iterative step thanks to the approximation based on a small number (\(C < 10\)) of Gibbs samples, in contrast to the \(O((N + J)2^K)\) complexity of the regularized EM algorithms in Chen et al. (2015) and Xu and Shang (2018) that evaluate the probabilities of all the \(2^K\) configurations of the binary latent patterns. This reduction to linear complexity in \(K\) greatly reduces the computational cost of estimating a SLAM for large-scale data. To our knowledge, this is among the first estimation algorithms for SLAMs or cognitive diagnostic models that have linear complexity in \(K\) and enjoy good estimation accuracy; see the simulation studies in Section 5 for details of performance.

### 4.2. Estimation Under the MultiParameter Model

The multiparameter model in Example 2 involves potentially many more parameters than the two-parameter model, since all the main effects and interaction effects of latent attributes possibly enter the likelihood. This complicated form poses a greater challenge to computation, especially for large-scale and high-dimensional scenarios considered here. Fortunately, the two-parameter DINA model is a submodel for multiparameter SLAMs in an interesting way such that under a same \(Q\)-matrix, the main term in the former exactly captures the highest-order interaction term of the active attributes in the latter (see the discussion after Example 2). Therefore, when the key interest is in recovering the discrete latent structures in \(Q\) and \(A\), such a relationship inspires the following question: can one maximize the two-parameter likelihood to obtain any meaningful approximate estimator when data indeed come from a multiparameter model?

On the practical side, the two-parameter DINA is indeed a very popular model employed by practitioners and likely the most widely used model in analyzing diagnostic assessment data.
in education (see, e.g., Chen et al. 2015; Culpepper 2015; Chen et al. 2018), though the multiparameter models are more general and flexible alternatives (Henson, Templin, and Wilse 2009; de la Torre 2011). Such practices are mainly due to the computational simplicity and nice interpretability of the two-parameter model, yet the risk of over-simplification exists. Motivated by the computational need and the scientific practice stated above, we next first study the property of the oversimplified joint MLE, obtained from maximizing the two-parameter likelihood when the true data-generating model is instead multiparameter. Later, we will show that such theoretical property inspires the development of a scalable two-step estimation procedure for multiparameter models.

4.3. Property of the Oversimplified Joint MLE Under the Two-Parameter Likelihood

Next, we consider the situation when a multiparameter SLAM is oversimplified to a two-parameter SLAM. We first provide conditions that guarantee a oversimplified joint MLE is consistent in estimating part of model structure. This provides a basis for subsequent second-stage estimation. We also establish that under certain stronger conditions, the oversimplified two-parameter joint MLE directly give consistent estimation of rows of Q and A. Together, these theoretical results will inspire the development of valid and efficient estimation methods for multiparameter SLAMs in the later half of this section.

We introduce some notation. Under a multiparameter SLAM, denote $p_{ij}^\text{true} = \mathbb{P}(r_{ij} = 1 | \text{true model})$. Given any $(Q, A)$, we define the two-parameter approximation by

$$
\hat{p}_{ij}(Q, A) = \begin{cases} 
\frac{\sum_{m=1}^{N} 1(a_m \geq q_j)}{\sum_{m=1}^{N} 1(a_m \geq q_j)} & \text{if } a_i \geq q_i; \\
\frac{\sum_{m=1}^{N} 1(a_m \geq q_j)}{\sum_{m=1}^{N} 1(a_m \neq q_j)} & \text{if } a_i \neq q_i.
\end{cases}
$$

The $\hat{p}_{ij}(Q, A)$ is determined by an arbitrary specification of the discrete latent structure $(Q, A)$ and also the true continuous parameters $p_{ij}^\text{true}$ (which further depends on $(Q^\text{true}, A^\text{true})$ and $\Theta^\text{true}$). As implied by the definition in Equation (18), the $\hat{p}_{ij}(Q, A)$ is indeed a two-parameter approximation, because for each set $j$, the set of probabilities $\{\hat{p}_{ij}(Q, A) : 1 \leq i \leq N\}$ only take two possible values, depending on whether or not $a_i \geq q_i$.

We first provide conditions sufficient for consistency of part of the discrete latent structures given by a oversimplified MLE. These conditions would imply a two-stage estimation procedure to be described in Section 4.2. Denote $D(p || q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))$, the Kullback–Leibler divergence of a Bernoulli distribution with parameter $p$ that with parameter $q$. Define the following function of $(Q, A)$,

$$
f_j(Q, A) = \sum_{i=1}^{N} D(\hat{p}_{ij} | p_{ij}^\text{true}(Q, A)).
$$

To interpret, for item $j$ the $f_j(Q, A)$ characterizes the KL divergences from the true parameters $p_{ij}^\text{true}$ to the two-parameter approximation induced by the discrete structure $(Q, A)$. We first consider the following assumption to replace the previous Assumption 2 on the true parameters under the multiparameter model. For two numbers $a$ and $b$, denote the maximum of them by $a \vee b$. Recall that the $\mathcal{E}_0$ defined in Assumption 4 is the set of variables that depend on some single latent attribute. Define $\mathcal{E}_0 = \{j \in [J] : q_j = e_k \text{ for some } k \in [K]\}$. Under Assumption 4, we have the following theorem.

**Assumption 4.** Define $\mathcal{E}_0 = \{j \in [J] : q_j = e_k \text{ for some } k \in [K]\}$. The true data-generating multiparameter SLAM satisfies

$$
\min_{j \in \mathcal{E}_0} \left( q_j - q_j^\text{true} \right)^2 > \zeta;
$$

$$
\sum_{j \notin \mathcal{E}_0} f_j(\mathbb{Z}^\text{true}) = \min_{\mathbb{Z}=(Q,A)} \sum_{j \notin \mathcal{E}_0} f_j(\mathbb{Z}) + o(N \cdot \eta),
$$

for some $\{\zeta_j\} \subseteq (0, 1)$ and some bounded sequence $\{\eta_j\} \subseteq [0, \infty)$.

In the special case where $\eta_j = 0$, Equation (20) implies that the oracle two-parameter approximation is the best possible two-parameter approximation in the sense of minimizing the KL-divergence. In general cases when $\{\eta_j\} \subseteq [0, \infty)$ is a bounded sequence, Equation (20) weakens requiring the oracle two-parameter approximation to be close to the best. This Equation (20) in Assumption 4 imposes a quite mild requirement on the data-generating true parameters.

**Theorem 2.** Suppose the data $\mathbb{Z}^\text{true}$ come from a multiparameter SLAM but the estimators $\hat{Z}^\text{true} = (Q^\text{true}, A^\text{true})$ are obtained through maximizing the oversimplified two-parameter likelihood (5). Suppose Assumptions 1, 3, and 4 hold. With $\sqrt{J} = O((\sqrt{MN}^{-1})^c)$ for a small $c > 0$ and $\gamma$ defined in Equation (15), there is

$$
\frac{1}{J} \sum_{j \in \mathcal{E}_0} I(\hat{q}_j^\text{true} \neq q_j) = o_p \left( \frac{\eta_j \vee \gamma}{\zeta_j \cdot \delta_j} \right),
$$

$$
\frac{1}{N} \sum_{i=1}^{N} I(\hat{a}_i^\text{true} \neq a_i) = o_p \left( \frac{\eta_j \vee \gamma}{\zeta_j \cdot \delta_j} \right),
$$

up to a permutation of the $K$ attributes. The joint MLE under an oversimplified two-parameter submodel is consistent in recovering $A^\text{true}$ and the single-attribute rows in $Q^\text{true}$.

Theorem 2 has the following useful practical implication. After a first step of maximizing the oversimplified two-parameter likelihood to estimate $A^\text{true}$ and the single-attribute rows in $Q^\text{true}$, a "regression"-type second step can be used to further estimate the remaining multi-attribute rows in $Q^\text{true}$ based on the first stage estimator $\hat{A}$. In Section 4.2, we provide a practical estimation procedure following this rationale.

In practice, when the true parameters are more similar to the two-parameter submodel than Assumption 4, the oversimplified joint MLE can even directly gives the consistency of all row vectors of $Q$ and $A$. The following assumption and theorem formalize this intuition.
Assumption 5 (True Parameters More Similar to a Two-Parameter Model). As $N, J \to \infty$, the true data-generating multiparameter SLAM satisfies

$$\min_{j \in [J]} \min_{\alpha \neq \beta} \sum_{j \in S_0} f_j(Q_{\alpha j} \neq Q_{\beta j}) = o(NJ \cdot \eta_j^2),$$

for some $\{\Delta_j\}, \{\eta_j\} \subseteq (0, \infty)$, where $f_j(Q_{\alpha j}, A_{\beta j})$ is as defined in Equation (19).

Theorem 3 (True Parameters More Similar to a Two-Parameter Model). Suppose the data $R_{\text{true}}$ come from a multiparameter SLAM but the estimators $\hat{Q}_{2\alpha} = (Q_{2\alpha}, \hat{A}_{2\alpha})$ are obtained through maximizing the oversimplified two-parameter likelihood (5). Under Assumptions 1, 3, and 5, as $N, J \to \infty$, with $\sqrt{J} = O(\sqrt{MN}^{1-c})$ for a small $c > 0$ and $\gamma_j$ defined in (15),

$$\frac{1}{J} \sum_{j=1}^{J} I(\hat{Q}_{2\alpha} \neq Q_{\alpha j}) = o_p(J \cdot \eta_j^2),$$

up to a permutation of the $K$ latent attributes. In this case, the joint MLE under a oversimplified two-parameter submodel is consistent in recovering rows of $Q_{\alpha j}$ and $A_{\beta j}$.

The implication of Theorem 3 is that when the true parameters are similar enough to a two-parameter model, directly maximizing the oversimplified two-parameter likelihood suffices in recovering all the discrete latent structures $Q$ and $A$. Thanks to Theorems 2 and 3, the scalable estimation algorithm for the two-parameter model proposed in Section 4.1 can serve as a useful approximation for computing joint MLE under a multiparameter model. In particular, the different scenarios characterized by Assumption 4 (referred to as multiparameter model with weaker two-parameter signal from now on) and Assumption 5 (referred to as multiparameter model with stronger two-parameter signal) inspire two ways of performing estimation. Since the conditions in Theorem 2 are weaker than those in Theorem 3, we next focus on the more general case of weaker two-parameter signal and present a two-stage estimation procedure. We also provide the one-stage estimation results corresponding to the stronger-two-parameter-signal case in the supplementary material.

4.4. Two-Stage Estimation for Multiparameter SLAMs Corresponding to Theorem 2

When the multiparameter model satisfies Assumption 4, Theorem 2 offers a useful insight that directly maximizing the oversimplified two-parameter likelihood can lead to consistent estimators of $A$ and those single-attribute row vectors in $Q$. Such theoretical guarantee about $A$ via a oversimplified MLE inspires the following two-stage estimation procedure. After using Algorithm 1 to obtain $\hat{Q}$ and $\hat{A}$, we fix $\hat{A}$ as some surrogate "covariates" in order to re-estimate the matrix $Q$ through a second regression step. Specifically, a multiparameter SLAM in Example 2 has the following reparameterization,

$$\theta_j, a_i = f \left( \sum_{S \subseteq \{1, \ldots, K\}} \mu_{jS} \prod_{k \in S} \hat{a}_{j,k} \right), \quad \mu_{jS} \neq 0 \text{ only if } q_{jS} = 1,$$

where $\mu_{jS}$ is the coefficient for the interaction effect of the attributes in $S$, and $q_{jS} = (q_{j,k} : k \in S)$. Therefore, the sparsity structure of vector $\mu_j = (\mu_{jS} : S \subseteq \{0, 1\}^K)$ in the reparameterization (22) encode the information of $q_j$. Now if $a_i$’s are treated as known instead of latent, for each $j$ we can use a penalized logistic regression to find the nonzero regression coefficients $\mu_{jS}$’s. Then those nonzero $\mu_{jS}$’s define the set $\mathcal{K}_j$ and hence determine the vector $q_j$. This is the basic rationale for our second regression stage.

More specifically, in this second regression stage, for each item $j$, the parameter vector $\mu_j = (\mu_{jS} : S \subseteq [K])$ involves all the possible interaction effects of the $K$ binary attributes. So $\mu_j$ has dimension $2^K$, which can be huge given a moderate number of latent attributes. This is in analogy to the high-dimensional regression problem for a generalized linear model with link function $f^{-1}$. When $2^K$ is huge, we recommend using the independence screening approach (Fan and Lv 2008) to select candidate interactions of the attributes and then performing the variable selection only on the set of candidate interactions of attributes. The all-effect marginal screening method is as follows. For an arbitrary subset $S \subseteq [K]$ of latent attributes, viewing the interaction term $\prod_{k \in S} \hat{a}_{j,k}$ as a "feature," we define its maximum marginal likelihood estimator $\hat{\mu}_{jS}^M = (\hat{\mu}_{jS0}, \hat{\mu}_{jS1})$ based on the logistic regression as

$$\hat{\mu}_{jS}^M = \arg \min_{\mu_{jS}} \frac{1}{N} \sum_{i=1}^{N} \left[ r_{ij} \left( \mu_{jS0} + \mu_{jS1} \prod_{k \in S} \hat{a}_{j,k} \right) + \log \left( 1 + \exp \left( \mu_{jS0} + \mu_{jS1} \prod_{k \in S} \hat{a}_{j,k} \right) \right) \right];$$

here using logistic regression is appropriate for marginal screening because the responses $\{r_{1j}, \ldots, r_{Nj}\}$ are binary. Then we select the following set $\mathcal{N}_{j, \text{Scree}}$ of candidate interactions,

$$\mathcal{N}_{j, \text{Scree}} = \left\{ S \subseteq [0, 1]^K : |\hat{\mu}_{jS}^M| > \tau_{\text{thres}} \right\},$$

where $\tau_{\text{thres}} > 0$ is a prespecified threshold. An even faster screening method is the main-effect marginal screening, which only screens the marginal main effects of the $K$ attributes for each item. That is, for $\tau_{\text{thres}} > 0$, define

$$\hat{\mathcal{K}}_{j, \text{main}} = \left\{ k \in \{1, \ldots, K\} : |\hat{\mu}_{j,(k)}^M| > \tau_{\text{thres}} \right\},$$

where

Remark 5. In practice, one can bypass the issue of the selection of the threshold $\tau_{\text{thres}}$ in Equation (24) in the following way. That is, we can arrange the absolute values of the $K$ marginal main effects $|\hat{\mu}_{j,(1)}^M|, \ldots, |\hat{\mu}_{j,(K)}^M|$ from the largest to the smallest, denoted by $|\hat{\mu}_{j,(1)}^M| > |\hat{\mu}_{j,(2)}^M| > \cdots > |\hat{\mu}_{j,(K)}^M|$. From this ranking, we then
select the first $K$ attributes as the candidate ones for which the gap $|\hat{\mu}_{j,k}^M| - |\hat{\mu}_{j,k}^M|_{K+1}$ is the largest. In the simulation studies, we find that main-effect marginal screening coupled with this selection strategy usually suffices for good performance.

Finally, for each item $j$, given the set of candidate terms $\hat{\mathcal{M}}_{j,m\text{Scree}}$ (or $\hat{\mathcal{M}}_{j,\text{Scree}}$), we use an $L_1$-penalized logistic regression treating these candidate terms as predictors to arrive at a final set of selected terms $\hat{\mathcal{M}}_{j,\text{pen}}$, which is a subset of the power set of $[K]$. The tuning parameter of the $L_1$ penalty is chosen by 5-fold cross-validation. Based on this, the vector $\hat{q}_j$ can be determined. The following example illustrates the two-stage estimation procedure.

**Example 3 (Estimating a MultiParameter Model).** We generate data with $(N, J, K) = (2400, 1200, 3)$ under the multiparameter model with $f(\cdot)$ being the identity link (GDINA; de la Torre 2011). The true $\mathbf{Q}$ has half of the row vectors loading on some single attribute, one fourth loading on two attributes, and the remaining one fourth loading on all three attributes; these are visualized in Figure 3. Define $K_j = \{k \in [K] : q_{jk} = 1\}$ to be the set of active attributes for variable $j$ and specify the $\mu$-parameters in (3) as

$$
\theta_{j,k} = \mu_{j,\emptyset} = 0.2,
\theta_{j,1,k} = \sum_{S \subseteq K_j \setminus j} \mu_{j,S} = 0.8;
\mu_{j,S} = \frac{\theta_{j,1,k} - \theta_{j,k}}{2|K_j| - 1} \quad \text{for any } S \subseteq K_j, S \neq \emptyset,
$$

where $|K_j| = 1, 2, 3$. This setting of the item parameters are the same as the simulation settings in Xu and Shang (2018), that is, for each item all the main-effect and interaction-effect parameters are equal. The results for $(N, J, K) = (2400, 1200, 3)$ are presented in the upper panel of Figure 3. In this scenario, the first-stage $\hat{\mathbf{Q}}^{1st}$ differs from $\mathbf{Q}^{\text{true}}$ by 39 entries, out of the $J \times K = 3600$ entries. The first stage $\hat{\mathbf{A}}$ exactly equals $\mathbf{A}_0$. Treating $\hat{\mathbf{A}}$ as known and fixed in the second stage estimation leads to a second-stage estimator $\hat{\mathbf{Q}}^{2nd}$ which exactly equals $\mathbf{Q}^{\text{true}}$. In the bottom panel of Figure 3, we show the estimation results for a simulated dataset with $(N, J, K) = (3000, 2000, 10)$ and the findings are similar.

## 5. Simulation Studies

### 5.1. Simulations under the Two-Parameter Model

In this simulation study, we generate data under the two-parameter DINA model and examine Algorithm 1’s performance under $(N, J) = (100, 1000)$, $(1000, 1000)$, and $(2000, 2000)$, and $K = 7, 10, 15 (2^7 = 128, 2^{10} = 1024, 2^{15} = 32768)$. In each simulation setting, the true $\mathbf{Q}$ vertically stacks $J/(2K)$ copies of $\mathbf{I}_K$, $J/(4K)$ copies of $\mathbf{Q}^{(2)}_{\text{block}} = (q_{jk}^{(2)})$, and another $J/(4K)$ copies of $\mathbf{Q}^{(3)}_{\text{block}} = (q_{jk}^{(3)})$; here the entries of $\mathbf{I}_K$ are in the locations $q_{jk}^{(2)} = 1$ for $k \in [K], q_{jk}^{(2)} = 1$ for $k \in \{1, \ldots, K - 1\}$ and $q_{jk}^{(2)} = 1$ for $k \in [K], q_{jk}^{(3)} = 1$ for $k \in \{1, \ldots, K - 1\}$ and $q_{jk}^{(3)} = 1$ for $k \in \{1, \ldots, K - 2\}$ and $q_{jk}^{(3)} = 1$ for $k \in \{1, \ldots, K - 1\}$ and $q_{jk}^{(3)} = 1$ for $k \in [K], q_{jk}^{(3)} = 1$ for $k \in [K]$.

The true parameters are set to $1 - \theta_j^+ = \theta_j^- = 0.2$ for each $j$. In each setting, 200 independent replications are carried out. The estimation accuracies are presented in Table 1. The column labeled as “$\hat{\mathbf{A}} = \mathbf{A}^{\text{true}}$” records the number of replications out
Table 1. Two-parameter model (DINA) estimation results.

| $2^k$ | $J$ | $N$ | $\hat{A}$ | $\hat{\theta}$ |
|-------|-----|-----|------------|----------------|
|       |     |     | $\hat{A} = A_{\text{true}}$ | $\hat{\theta}_{ij} = \theta_{ij}$ |
| $2^7$ | 100 | 1000| 0/200 | 0.910 | 0.986 | 200/200 | 1.000 | 1.000 |
|       | 1000 | 1000 | 188/200 | 0.940 | 0.970 | 188/200 | 0.941 | 0.975 |
|       | 2000 | 2000 | 200/200 | 1.000 | 1.000 | 200/200 | 1.000 | 1.000 |
| $2^{10}$ | 100 | 1000 | 0/200 | 0.678 | 0.949 | 185/200 | 0.965 | 0.987 |
|       | 1000 | 1000 | 189/200 | 0.955 | 0.980 | 191/200 | 0.956 | 0.993 |
|       | 2000 | 2000 | 200/200 | 1.000 | 1.000 | 200/200 | 1.000 | 1.000 |
| $2^{15}$ | 200 | 2000 | 0/200 | 0.709 | 0.956 | 166/200 | 0.935 | 0.980 |
|       | 1000 | 1000 | 138/200 | 0.985 | 0.993 | 194/200 | 0.985 | 0.995 |
|       | 2000 | 2000 | 200/200 | 1.000 | 1.000 | 200/200 | 1.000 | 1.000 |

Table 2. Two-stage estimation for multiparameter model (GDINA) under the weaker two-parameter signal with $\mu_{j,k} = (\theta_{j,k} - \theta_{j,k}) / (2^{[K_j / 2]} - 1)$ where $|K_j| = 1, 2, 3$.

| $2^k$ | $J$ | $N$ | $\hat{A}$ | $\hat{\theta}$ |
|-------|-----|-----|------------|----------------|
|       |     |     | $\hat{A} = A_{\text{true}}$ | $\hat{\theta}_{ij} = \theta_{ij}$ |
| $2^7$ | 100 | 1000 | 0.858 | 0.978 | 0.971 | 0.996 | 0.991 | 0.998 |
|       | 1000 | 1000 | 1.000 | 1.000 | 0.916 | 0.988 | 0.999 | 0.999 |
|       | 2000 | 2000 | 1.000 | 1.000 | 0.955 | 0.994 | 1.000 | 1.000 |
| $2^{10}$ | 100 | 1000 | 0.560 | 0.941 | 0.943 | 0.993 | 0.971 | 0.995 |
|       | 1000 | 1000 | 1.000 | 1.000 | 0.910 | 0.991 | 0.986 | 0.998 |
|       | 2000 | 2000 | 1.000 | 1.000 | 0.950 | 0.995 | 1.000 | 1.000 |
| $2^{15}$ | 200 | 2000 | 0.546 | 0.942 | 0.960 | 0.996 | 1.000 | 1.000 |
|       | 1000 | 1000 | 1.000 | 1.000 | 0.904 | 0.994 | 0.983 | 0.999 |
|       | 2000 | 2000 | 1.000 | 1.000 | 0.941 | 0.996 | 1.000 | 1.000 |

Table 3. One-stage estimation for multiparameter model (GDINA) under the stronger two-parameter signal with $\mu_{j,k} = (\theta_{j,k} - \theta_{j,k}) / 2$.

| $2^k$ | $J$ | $N$ | $\hat{A}$ | $\hat{\theta}$ |
|-------|-----|-----|------------|----------------|
|       |     |     | $\hat{A} = A_{\text{true}}$ | $\hat{\theta}_{ij} = \theta_{ij}$ |
| $2^7$ | 100 | 1000 | 0.888 | 0.983 | 192/200 | 1.000 | 1.000 |
|       | 1000 | 1000 | 200/200 | 1.000 | 1.000 | 192/200 | 1.000 | 1.000 |
|       | 2000 | 2000 | 200/200 | 1.000 | 1.000 | 192/200 | 1.000 | 1.000 |
| $2^{10}$ | 100 | 1000 | 0.638 | 0.952 | 160/200 | 0.982 | 0.997 |
|       | 1000 | 1000 | 199/200 | 1.000 | 1.000 | 114/200 | 1.000 | 1.000 |
|       | 2000 | 2000 | 200/200 | 1.000 | 1.000 | 114/200 | 1.000 | 1.000 |
| $2^{15}$ | 200 | 2000 | 0.689 | 0.968 | 171/200 | 0.973 | 0.994 |
|       | 1000 | 1000 | 148/200 | 1.000 | 1.000 | 40/200 | 0.998 | 1.000 |
|       | 2000 | 2000 | 200/200 | 1.000 | 1.000 | 170/200 | 1.000 | 1.000 |

of 200 where the algorithm exactly recovers the entire matrix $A$; column “$\hat{\theta}_{ij} = \theta_{ij}$” records the mean accuracy of recovering the $N$ row vectors of $A$ across the replications; column “$\hat{\theta}_{ij,k} = \theta_{ij,k}$” records the mean accuracy of recovering the NK individual entries of $A$. The columns $Q = Q_{\text{true}}$, $\hat{\theta}_{ij} = \theta_{ij}$, $\hat{\theta}_{ij,k} = \theta_{ij,k}$ record similar measures for $Q$. When $N = J = 2000$, for all the considered $K$, both matrix $Q$ and matrix $A$ are exactly recovered in each replication.

5.2. Simulations under the MultiParameter Model

We generate data under parameter settings similar to Example 3 for various $N, J$, and $K$. For $K = 7$ or $K = 10$, we vary $(N, J)$ in $(100, 1000)$, $(1000, 1000)$, and $(2000, 2000)$; for $K = 15$, we vary $(N, J)$ in $(200, 2000)$, $(1000, 1000)$, and $(2000, 2000)$. In each of the considered scenarios, 200 simulation replications are carried out. The estimation results are shown in Tables 2 and 3, respectively. The true parameter settings behind Table 2 correspond to the weaker multiparameter Assumption 4, and they are the same as the simulation settings in Xu and Shang (2018). Table 2 shows that the first stage estimation yields very high accuracy of estimating rows in $A$ (perfect recovery in the considered scenarios), which provides a good basis for proceeding with the second stage of re-estimating rows in $Q$. Indeed, the second-stage estimator $\hat{Q}^{(2)}$ based on the penalized regression approach introduced in Section 4.1 shows desirable improvement over the first-stage estimator $\hat{Q}^{(1)}$. The true parameter settings behind Table 3 correspond to the stronger multiparameter Assumption 5. As $(N, J)$ increase from $(1000, 1000)$ to $(2000, 2000)$, the oversimplified two-parameter MLE improves to almost perfect recovery of the discrete latent structures. This corroborates Theorem 3 that when the true parameters underlying a multiparameter model are more similar to a two-parameter model, the oversimplified MLE from one-stage estimation can itself leads to consistency.
In practice, when fitting a SLAM to real data, if it is not clear whether a two-parameter model or a multiparameter one is more suitable, we recommend performing two-stage estimation as described in Section 4.2 to improve the estimation accuracy of the Q-matrix, as empirically shown in Table 2. Then after performing the two-stage estimation procedure, one can apply some information criterion such as BIC to compare the first-stage estimator under the two-parameter model and second-stage estimator under the multiparameter model in order to reach a final decision. In summary, our simulation studies show that across all the considered scenarios including the challenging case with $2^{15} = 32,768$, the proposed estimators have good accuracy of recovering the $q_j$’s and $a_i$’s.

6. Real Data Analysis

We apply the proposed estimation method to real data from an educational assessment, the TIMSS. This dataset is a subset of the TIMSS 2011 Austrian data for analyzing students’ abilities in mathematical sub-competences and is available in the R package cdm. It includes responses of $N = 1010$ Austrian fourth grade students and $J = 47$ items. Nine ($K = 9$) attributes were specified in George and Robitzsch (2015): (DA) Data and Applying, (DK) Data and Knowing, (DR) Data and Reasoning, (GA) Geometry and Applying, (GK) Geometry and Knowing, (GR) Geometry and Reasoning, (NA) Numbers and Applying, (NK) Numbers and Knowing, (NR) Numbers and Reasoning; a provisional Q-matrix $Q^{\text{orig}}$ of size $47 \times 9$ was also provided.

One structure specific to such large scale assessments is that only a subset of all items in the entire study is presented to each student (George and Robitzsch 2015). This results in many missing values in the $N \times J$ data matrix, and the considered dataset has a missing rate 51.73%. The joint MLE approach can be easily extended to handle the missing data under the ignorable missingness assumption. Under such an assumption, it indeed suffices to replace the log-likelihood function over the $\{r_{ij} : (i,j) \in [N] \times [J]\}$ by that over $\{r_{ij} : (i,j) \in \Omega\}$, where $\Omega \subseteq [N] \times [J]$ is the set of indices in $R$ corresponding to those observed entries. In particular, the original log-likelihood function (7) under the two-parameter model should be replaced by the following objective function,

$$
\ell_{\Omega, \text{two}}(Q, A, \Theta|\mathbf{R}) = \sum_{(i,j) \in \Omega} \left[ r_{ij} \left( \prod_{k} a_{qj,k} \log \theta_{qj}^{+} + \left(1 - \prod_{k} a_{qj,k} \right) \log \theta_{qj}^{-} \right) \right] + (1 - r_{ij}) \left( \prod_{k} a_{qj,k} \log (1 - \theta_{qj}^{+}) + (1 - \prod_{k} a_{qj,k}) \log (1 - \theta_{qj}^{-}) \right).
$$

With missing values in $R$, the previous ADG-EM Algorithm 1 can be replaced by Algorithm S.2 presented in the Supplementary Material.

The original Q-matrix provided in the TIMSS dataset $Q^{\text{orig}}$ has each item measuring only one attribute. This $Q^{\text{orig}}$ gives the interpretation of the attributes and encodes the domain knowledge about the test items. Therefore, we use $Q^{\text{orig}}$ to initialize the proposed algorithm. Moreover, we fix $I_{\text{anchor}} = K = 9$ “anchor” items’ row vectors in $Q^{\text{orig}}$ along the iterations of the algorithm. The anchor items are chosen such that their corresponding row vectors form an identity submatrix $I_K$ of the Q-matrix. By this we hope to fix the interpretation of the K columns as the K provided attributes. The two-parameter DINA model is often used to model and analyze data from educational assessments. In the data analysis, we first perform estimation under the two-parameter DINA model and then also proceed with the second-stage estimation as described in Section 4.2 to estimate Q-matrix under a multiparameter GDINA model. But the two-parameter model gives a smaller BIC value and indicates a better fit. So next we only discuss the results given by the two-parameter model fitting.

In the resulting estimator $Q^{\text{est}}$, there are 10 rows that have more than 1 nonzero entries, which are presented in Table 4 together with their item number and item label. First, for each of these 10 items, the estimated $q$-vector always measures the attribute originally specified in $Q^{\text{orig}}$ (the dark orange entry of “1” in each of the 10 rows in Table 4). This implies that the meaning of the original attributes are preserved in our estimation. In addition, Table 4 reveals extra information that some items depend on certain additional attributes besides the originally specified one (the dark blue entries of “1” in Table 4). For example, items M031379, M031380, M051001 originally are designed to measure attribute (NR) Number and Reasoning, but the estimated $Q^{\text{est}}$ implies they also depend on the attribute (NA) Number and Applying. In particular, the third item M051001 “Soccer tournament” asks: in a soccer tournament, teams get: 3 points for a win, 1 point for a tie, 0 points for a loss. Zedland has 11 points. What is the smallest number of games Zedland could have played? This is a difficult question for fourth graders and targets complicated skills in the content domain “Number”; its difficulty

| Item No. | Item Label               | DA | DK | DR | GA | GK | GR | NA | NK | NR |
|---------|--------------------------|----|----|----|----|----|----|----|----|----|
| M031379 | Trading sports cards     | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  |
| M031380 | Trading cartoon cards    | 0  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  |
| M051001 | Soccer tournament        | 0  | 1  | 0  | 0  | 0  | 0  | 1  | 1  | 1  |
| M051015 | Complete Jay’s shape     | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| M051123 | Lines of symmetry complex figure | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 1  | 0  |
| M041098 | How many cans must Sean buy | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| M041104 | Number between 5 and 6    | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| M041299 | Fraction of the cake eaten | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 1  | 0  |
| M041143 | Identify shapes in the picture | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  |
| M051006 | Cost of ice cream         | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
Table 4 shows that items generally seem to have some clustered dependence on attributes falling in the same cognitive domain or the same content domain: attributes (NA), (NK), (NR) in the content domain “Number” are often measured together, and attributes (GA) and (NA) in the cognitive domain “Applying” are often measured together.

We also examine the estimated $N \times K$ matrix $A^{est}$ for the $N = 1010$ students. Based on $A^{est}$, marginally, students master skills regarding Data (average mastery 50.63%) better than Geometry (average mastery 48.12%) and Number (average mastery 46.50%); and they master skills regarding Applying (average mastery 50.83%) and Knowing (average mastery 49.17%) better than Reasoning (average mastery 45.25%). The “average mastery” above is calculated as follows: for Data, the average mastery is taken to be the average of the three columns of $A^{est}$ corresponding to DA, DK, DR. Figure 4 further shows the pairwise correlations between the nine attributes based on $A^{est}$. It can be seen that the attributes falling in the same content domain Number do show relatively high correlations, where the three pairwise correlations between NA, NK, NR are 0.23, 0.24, 0.20. The correlation between GA and NA is 0.28, also high. This aligns with our earlier observation that the estimated row vectors in $Q^{est}$ also tend to measure these attributes together.

7. Discussion

This article investigates the joint MLE approach to large-scale structured latent attribute analysis from both the theoretical and methodological perspectives. We provide theoretical guarantees for the estimability and consistency of the latent structures in the regime where all of the number of individuals, the number of observed variables, and the number of latent attributes can grow large. The obtained estimation error bounds not only guarantee asymptotic consistency of estimating both the variable loading vectors and subject latent profiles, but also offer insights into their estimation accuracies with finite samples. These consistency results also give practical implications for designing the $Q$-matrix in cognitive diagnostic applications. For computation, we develop a scalable approximate algorithm to find the joint MLE of two-parameter SLAMs and also propose an effective two-stage estimation procedure for multiparameter SLAMs. Simulation studies and real data analysis demonstrate the usefulness of the proposed estimation approaches.

The developments in this work also open up several possibilities for future research. On the methodological side, based on the established results on consistency and finite sample error bounds for estimating latent structures, an interesting future task is performing statistical inference on SLAMs with a large number of test items and high-dimensional latent attributes. On the computational front, it would be interesting to relate or generalize the idea of the proposed estimation algorithm to other discrete optimization problems; it is also desirable to investigate the algorithm’s theoretical properties in the future.

This article focuses on the discrete latent attribute modeling framework, and we include a particular study of misspecifying a multiparameter SLAM to a two-parameter submodel motivated by computational needs and scientific practices. Besides such possible oversimplification, there could be other types of misspecifications, such as potentially misspecifying the continuous latent variables to be discrete. As for this, we point out that this work does not intend to replace continuous latent factor modeling with the discrete counterpart, but rather to complement the former in suitable applications. In the future, it would be interesting to study consequences of the potential misspecification of continuous latent variables to discrete ones to elucidate their differences and connections.

On a final note, many specific models belonging to the SLAM family were initially proposed in the literature of cognitive diagnostic modeling. But this modeling framework’s unique advantages of capturing fine-grained latent information and providing model-based clustering allow for applications far beyond this discipline. For example, similar modeling approaches have recently been employed in psychiatric evaluation (de la Torre, van der Ark, and Rossi 2018), disease epidemiology diagnosis (O’Brien et al. 2019), electronic health records (Ni, Müller, and van der Ark, and Rossi 2018), disease epidemiology diagnosis (O’Brien et al. 2019), electronic health records (Ni, Müller, and van der Ark, and Rossi 2018), and precision medicine (Chen, Zeng, and Wang 2021). Just like the continuous latent factor analysis is nowadays widely used (Fan et al. 2021; Chen, Li, and Zhang 2020b; Bing et al. 2020) beyond its initial application in psychometrics, we believe the multidimensional discrete latent trait modeling also has great future promise in broader fields and warrants further statistical developments. By introducing and analyzing a principled joint MLE approach here, we hope this work contributes a step toward that farreaching goal.

Acknowledgment

The authors thank the Editor, Associate Editor, and reviewer for helpful and constructive comments.
Funding

This research was supported by NSF CAREER SES-1846747, DMS-1712717, SES-1659328, and also by NIH NIEHS R01ES027498, R01ES028804. This research has also received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation program (grant agreement No 856506).

Supplementary Material

The Supplementary Material contains all the technical proofs of the theoretical results and also includes additional discussion on computation.

ORCID

Yuqi Gu  http://orcid.org/0000-0002-4124-113X

References

Anderson, T. W., and Rubin, H. (1956), “Statistical Inference in Factor Analysis,” in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 5, pp. 111–150. [749]
Bacci, S., and Bartolucci, F. (2016), “Two-Tier Latent Class IRT Models in R,” The R Journal, 8, 139. [749]
Bartolucci, F. (2007), “A Class of Multidimensional IRT Models for Testing Unidimensionality and Clustering Items,” Psychometrika, 72, 141. [749]
Bing, X., Bunea, F., Ning, Y., and Wegkamp, M. (2020), “Adaptive Estimation in Structured Factor Models with Applications to Overlapping Clustering,” Annals of Statistics, 48, 2055–2081. [759]
Bonhomme, S., and Manresa, E. (2015), “Grouped Patterns of Heterogeneity in Panel Data,” Econometrica, 83, 1147–1184. [750]
Chen, Y., Culppeper, S., and Liang, F. (2020a), “A Sparse Latent Class Model for Cognitive Diagnosis,” Psychometrika, 85, 121–153. [746]
Chen, Y., Culppeper, S. A., Chen, Y., and Douglas, J. (2018), “Bayesian Estimation of the DIN A Q-matrix,” Psychometrika, 83, 89–108. [754]
Chen, Y., Li, X., and Zhang, S. (2019), “Joint Maximum Likelihood Estimation for High-Dimensional Exploratory Item Factor Analysis,” Psychometrika, 84, 124–146. [747,752]
Chen, Y., Li, X., and Zhang, S. (2020b), “Large-Scale Latent Factor Analysis for Large-Scale Data: Identifiability, Estimability, and Their Implications,” Journal of the American Statistical Association, 115, 1756–1770. [747,751,752,759]
Chen, Y., Liu, J., Xu, G., and Ying, Z. (2015), “Statistical Analysis of Q-Matrix-Based Diagnostic Classification Models,” Journal of the American Statistical Association, 110, 850–866. [746,747,748,749,750,751,753,754]
Chen, Y., Zeng, D., and Wang, Y. (2021), “Learning Individualized Treatment Rules for Multiple-Domain Latent Outcomes,” Journal of the American Statistical Association, 116, 269–282. [747,759]
Choi, D. S., Wolfe, P. J., and Airolidi, E. M. (2012), “Stochastic Blockmodels With a Growing Number of Classes,” Biometrics, 99, 273–284. [752]
Culpepper, S. A. (2015), “Bayesian Estimation of the DIN A Model With Gibbs Sampling,” Journal of Educational and Behavioral Statistics, 40, 454–476. [754]
Culpepper, S. A. (2019), “Estimating the Cognitive Diagnosis Q-matrix With Expert Knowledge: Application to the Fraction-Subtraction Dataset,” Psychometrika, 84, 333–357. [746]
de la Torre, J. (2011), “The Generalized DINA Model Framework,” Psychometrika, 76, 179–199. [746,747,748,749,750,751,753,754]
de la Torre, J., van der Ark, L. A., and Rossi, G. (2018), “Analysis of Clinical Data From a Cognitive Diagnosis Modeling Framework,” Measurement and Evaluation in Counseling and Development, 51, 281–296. [746,748,759]
Delyon, B., Lavielle, M., and Moulines, E. (1999), “Convergence of a Stochastic Approximation Version of the EM Algorithm,” Annals of Statistics, 94, 94–128. [753]
Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood From Incomplete Data Via the EM Algorithm,” Journal of the Royal Statistical Society, Series B, 39, 1–22. [752]
Embroten, S. E., and Reise, S. P. (2013), Item Response Theory, New York: Psychology Press. [749]
Fan, J., and Lv, J. (2008), “Sure Independence Screening for Ultrahigh Dimensional Feature Space,” Journal of the Royal Statistical Society, Series B, 70, 849–911. [755]
Fan, J., Wang, K., Zhong, Y., and Zhu, Z. (2021), “Robust High-Dimensional Factor Models With Applications to Statistical Machine Learning,” Statistical Science, 36, 303–327. [759]
Fang, G., Liu, J., and Ying, Z. (2019), “On the Identifiability of Diagnostic Classification Models,” Psychometrika, 84, 19–40. [746]
George, A. C. and Robitzsch, A. (2015), “Cognitive Diagnosis Models in R: A Didactic,” The Quantitative Methods for Psychology, 11, 189–205. [758]
Gu, Y., and Xu, G. (2019a), “Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models,” Journal of Machine Learning Research, 20, 1–58. [746,747,749,750,751]
——— (2019b), “The Sufficient and Necessary Condition for the Identifiability and Estimability of the DIN A Model,” Psychometrika, 84(2):468–483. [750,751]
——— (2020), “Partial Identifiability of Restricted Latent Class Models,” Annals of Statistics, 48, 2082–2107. [746,751]
Henson, R. A., Templin, J. L., and Willse, J. T. (2009), “Defining a Family of Cognitive Diagnosis Models Using Log Linear Models With Latent Variables,” Psychometrika, 74, 191–210. [746,747,754]
Junker, B. W., and Sijtsma, K. (2001), “Cognitive Assessment Models With Few Assumptions, and Connections With Nonparametric Item Response Theory,” Applied Psychological Measurement, 25, 258–272. [746,748,750]
Lazarsfeld, P. F., and Henry, N. W. (1968), Latent Structure Analysis, Boston, MA: Houghton Mifflin. [749]
Mullis, I. V. S., Martin, O. M., Goh, S., and Cotter, K. (2016), TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science. Chestnut Hill, MA: Boston College, TIMSS and PIRLS International Study Center. [747]
Neyman, J., and Scott, E. L. (1948), “Consistent Estimates Based on Partially Consistent Observations,” Econometrica: Journal of the Econometric Society, 1–32. [747]
Ni, Y., Müller, P., and Ji, Y. (2020), “Bayesian Double Feature Allocation for Phenotyping With Electronic Health Records,” Journal of the American Statistical Association, 115, 1620–1634. [749,759]
O’Brien, K. L., Baggett, H. C., and Brooks, W. A. (2019), “Causes of Severe Pneumonia Requiring Hospital Admission in Children Without HIV Infection From Africa and Asia: The Perch Multi-Country Case-Control Study,” The Lancet, 394, 757–759. [746,747,759]
Reckase, M. (2009), Multidimensional Item Response Theory, Vol. 150. Springer. [749]
Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000), “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, 10, 19–41. [749]
Tatsuoka, K. K. (1983), “Rule Space: An Approach for Dealing With Misconceptions Based on Item Response Theory,” Journal of Educational Measurement, 20, 345–354. [746]
Templin, J. L., and Henson, R. A. (2006), “Measurement of Psychological Disorders Using Cognitive Diagnosis Models,” Psychological Methods, 11, 287–305. [746,748]
von Davier, M. (2008), “A General Diagnostic Model Applied to Language Testing Data,” British Journal of Mathematical and Statistical Psychology, 61, 287–307. [746,748]
von Davier, M., and Lee, Y.-S. (2019), Handbook of Diagnostic Classification Models. Cham: Springer International Publishing. [747,750]
Wang, S., and Douglas, J. (2015), “Consistency of Nonparametric Classification in Cognitive Diagnosis,” Psychometrika, 80, 85–100. [751]
Wu, Z., Deloria-Knoll, M., and Zeger, S. L. (2017), “Nested Partially Latent Class Models for Dependent Binary Data; Estimating Disease Etiology,” Biostatistics, 18, 200–213. [746]
Xu, G. (2017), “Identifiability of Restricted Latent Class Models With Binary Responses,” Annals of Statistics, 45, 675–707. [746,747,750]
Xu, G., and Shang, Z. (2018), “Identifying Latent Structures in Restricted Latent Class Models,” Journal of the American Statistical Association, 113, 1284–1295. [746,747,749,750,751,753,756,757]
Zhao, Y., Bickel, P., and Weko, C. (2020), “Identifiability and Consistency of Network Inference Using the Hub Model and Variants: A Restricted Class of Bernoulli Mixture Models,” arXiv:2004.09709. [752]