Bridging Parametric and Nonparametric Methods in Cognitive Diagnosis *

Chenchen Ma[1], Jimmy de la Torre[2], and Gongjun Xu[1]

[1] University of Michigan

[2] University of Hong Kong

Abstract

A number of parametric and nonparametric methods for estimating cognitive diagnosis models (CDMs) have been developed and applied in a wide range of contexts. However, in the literature, a wide chasm exists between these two families of methods, and their relationship to each other is not well understood. In this paper, we propose a unified estimation framework to bridge the divide between parametric and nonparametric methods in cognitive diagnosis to better understand their relationship. We also develop iterative joint estimation algorithms and establish consistency properties within the proposed framework. Lastly, we present comprehensive simulation results to compare different methods, and provide practical recommendations on the appropriate use of the proposed framework in various CDM contexts.

1 Introduction

Cognitive diagnosis models (CDMs), also known as diagnostic classification models, are typically used in conjunction with diagnostic assessments to determine fine-grained classifications of subjects' latent attribute patterns based on their observed responses to specifically-designed diagnostic items. In educational assessments, the latent attributes can represent the mastery or lack of target skills (de la Torre, 2011; Junker and Sijtsma, 2001). Students' skill profiles, which are inferred from the their responses to test items, are used for subsequent learning interventions. In psychiatric diagnosis, the latent attributes can be construed as the presence or absence of some underlying mental disorders (de la

*This research is partially supported by NSF CAREER SES-1846747, DMS-1712717, SES-1659328.

Torre, van der Ark, and Rossi, 2018; Templin and Henson, 2006). Patients' responses to questionnaire items serve as the basis for identifying their mental disorder statuses, which in turn determines the appropriate treatments.

Several parametric models for cognitive diagnosis have been developed and widely applied in practice. Popular examples include the deterministic input, noisy "and" gate (DINA) model (Junker and Sijtsma, 2001), the deterministic input, noisy "or" gate (DINO) model (Templin and Henson, 2006), the reduced reparameterized unified model (Reduced RUM; Hartz, 2002), the general diagnostic model (GDM; von Davier, 2005), the log-linear CDM (LCDM; Henson et al., 2009), and the generalized DINA model (GDINA; de la Torre, 2011). To estimate these parametric models, estimators maximizing the marginal likelihood or joint likelihood functions have been employed (e.g., Chiu et al., 2016; de la Torre, 2009).

Parametric CDMs, such as the DINA or DINO model, invoke certain parametric assumptions about the item response functions. As pointed out in Chiu and Douglas (2013), such assumptions may raise validity concerns about the assumed model and the underlying process. As an alternative, some researchers have explored nonparametric methods for assigning subjects to latent groups without relying on parametric model assumptions. For example, Chiu and Douglas (2013) proposed the nonparametric classification (NPC) method, where a subject is classified to its closet latent group by comparing the observed responses with ideal responses either from the DINA or DINO model. Its generalization, the general NPC (GNPC) method proposed by Chiu et al. (2018), uses the weighted average of ideal responses from the DINA and DINO models to accommodate more general settings. Consistency results for the NPC and the GNPC methods were established by Wang and Douglas (2015) and Chiu and Köhn (2019a), respectively. Simulation results show that, compared to parametric methods, nonparametric methods tend to perform better when the sample sizes are not sufficiently large to provide reliable maximum likelihood estimates.

Even though the aforementioned parametric and nonparametric methods have been used in many CDM applications, the relationship between these two families of methods have not been explicitly discussed in the literature. Although seemingly divergent from the surface, these frameworks are in fact closely related. In this paper, we propose a unified estimation framework for cognitive diagnosis that subsumes both parametric and nonparametric methods. In the proposed framework, we use a general loss function to measure the distance between a subject's responses and the centroid of a latent class. By using different loss functions, the method can assume different parametric and nonparametric

forms. Under the general framework, we further develop a unified iterative joint estimation algorithm, as well as establish the consistency properties of the corresponding estimators. Finally, we conduct comprehensive simulation studies to compare different parametric and nonparametric methods under a variety of settings, and provide relevant practical recommendations accordingly.

The rest of the paper is organized as follows. Section 2 gives a brief review of both parametric and nonparametric methods in cognitive diagnosis assessment. Section 3 introduces the proposed general estimation framework with several illustrative examples. Section 4 presents the consistency results of the proposed method, and Section 5 presents the simulation results. Finally, Section 6 discusses some future extensions, whereas proofs of the main theorems are reported in the online Appendix.

2 Parametric and Nonparametric Methods

Before introducing our proposed estimation framework, we give a brief review of both parametric and nonparametric methods that are widely used in the CDM literature.

2.1 Parametric Methods

Parametric methods directly model item response functions under certain parametric model assumptions. Most of CDMs are parametric models, where the item response probabilities are modeled as functions of item parameters and the latent attributes of subjects. Specifically, in a CDM with J items and K latent attributes, two types of subject-specific variables are of interest. One is the observed responses to J items $\boldsymbol{x} = (x_1, \ldots, x_J) \in \{0,1\}^J$, and the other is the mastery profile of K latent attributes $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) \in \{0,1\}^K$. There are 2^K possible latent patterns, and we use $\boldsymbol{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K)$ to denote the proportion parameters for the latent attribute patterns of subjects, which satisfies $p_{\boldsymbol{\alpha}} \in [0,1]$ and $\sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} = 1$.

Given a subject's latent attribute pattern $\boldsymbol{\alpha}$, the responses to J items are assumed to be independent and follow Bernoulli distributions with parameters $\theta_{1,\boldsymbol{\alpha}}, \ldots, \theta_{J,\boldsymbol{\alpha}}$. Specifically, $\theta_{j,\boldsymbol{\alpha}} := \mathbb{P}(x_j = 1|\boldsymbol{\alpha})$, which is the probability (item response function) of providing a positive response to item j for latent class $\boldsymbol{\alpha}$. We write $\boldsymbol{\theta} = \left(\theta_{j,\boldsymbol{\alpha}} : j \in [J], \boldsymbol{\alpha} \in \{0,1\}^K\right)$ to denote the item response probability matrix, with $[J]$ denoting the set $\{1, \ldots, J\}$. Then under the local independence assumption, the probability

mass function of a subject's response vector $\boldsymbol{x} = (x_1, \ldots, x_J) \in \{0,1\}^J$ takes the form

$$\mathbb{P}(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{p}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{x_j} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-x_j}. \tag{1}$$

To reflect the dependence between items and the latent attributes of subjects, a structural matrix, the so-called Q-matrix (Tatsuoka, 1983), is used to impose constraints on item parameters. Specifically, $\boldsymbol{Q} \in \{0,1\}^{J \times K}$, where $q_{j,k} = 1$ if item j requires (or depends on) attribute k. The jth row vector of $\boldsymbol{Q}$ denoted by $\boldsymbol{q}_j$ describes the full dependence of item j on K latent attributes. Usually in applications such as cognitive diagnostic assessments, the matrix $\boldsymbol{Q}$ is pre-specified by domain experts (George and Robitzsch, 2015; Junker and Sijtsma, 2001; von Davier, 2005) to reflect some scientific assumptions. The structural matrix $\boldsymbol{Q}$ puts constrains on item parameters in certain ways under different model assumptions. One important common assumption is that the item response function $\theta_{j,\boldsymbol{\alpha}}$ only depends on whether latent attribute pattern $\boldsymbol{\alpha}$ contains the required attributes by item j (i.e., the attributes in the set $\mathcal{K}_j = \{k \in [K] : q_{j,k} = 1\}$ with $[K]$ denoting the set $\{1, \ldots, K\}$). Here we introduce three commonly used parametric models.

Example 1 (DINA)**.** *The DINA (Junker and Sijtsma, 2001) model assumes a conjunctive relationship among attributes, where mastery of all the required attributes for an item is necessary for a subject to be deemed capable of providing a positive response (e.g., correct response, item endorsement), and possessing additional unnecessary attributes does not compensate for the lack of necessary attributes. In the DINA model, the so-called ideal response for each item j and each latent attribute pattern $\boldsymbol{\alpha}$ is defined as,*

$$\eta_{j,\boldsymbol{\alpha}}^{DINA} = \prod_{k=1}^{K} \alpha_k^{q_{jk}}. \tag{2}$$

The uncertainty is further incorporated by introducing the slipping and guessing parameters s_j and g_j for $j = 1, \ldots, J$. For each item j, the slipping parameter is the probability of a negative response for capable subjects, and the guessing parameter is the probability of a positive response for incapable subjects, as in, $s_j = \mathbb{P}(x_{ij} = 0 \mid \eta_{j,\boldsymbol{\alpha}_i} = 1)$ and $g_j = \mathbb{P}(x_{ij} = 1 \mid \eta_{j,\boldsymbol{\alpha}_i} = 0)$, where $\boldsymbol{\alpha}_i$ is the latent pattern for the ith subject. Therefore, we have

$$\theta_{j,\boldsymbol{\alpha}}^{DINA} = (1 - s_j)^{\eta_{j,\boldsymbol{\alpha}}^{DINA}} g_j^{1 - \eta_{j,\boldsymbol{\alpha}}^{DINA}}.$$

That is, the item response function is $1 - s_j$ if the ideal response is 1, and g_j otherwise.

Example 2 (DINO). *The DINO (Templin and Henson, 2006) model assumes a disjunctive relationship among attributes, where mastery of one of the required attributes of an item is necessary for a subject to be considered capable of providing a positive response. In the DINO model, the ideal response is defined as,*

$$\eta_{j,\boldsymbol{\alpha}}^{DINO} = 1 - \prod_{k=1}^{K} (1 - \alpha_k)^{q_{jk}}. \tag{3}$$

For the DINO model, the slipping parameters and guessing parameters are defined in the similar way as the DINA model, as in, $s_j = \mathbb{P}(x_{ij} = 0 \mid \eta_{j,\boldsymbol{\alpha}_i} = 1)$ *and* $g_j = \mathbb{P}(x_{ij} = 1 \mid \eta_{j,\boldsymbol{\alpha}_i} = 0)$. *Accordingly, we have*

$$\theta_{j,\boldsymbol{\alpha}}^{DINO} = (1 - s_j)^{\eta_{j,\boldsymbol{\alpha}}^{DINO}} g_j^{1 - \eta_{j,\boldsymbol{\alpha}}^{DINO}}.$$

Example 3 (GDINA). *The GDINA (de la Torre, 2011) model is a more general CDM, where all the interactions among the required latent attributes by each item are considered. The item response function for the GDINA model is defined as*

$$\theta_{j,\boldsymbol{\alpha}}^{GDINA} = f\Big(\sum_{S \subset \mathcal{K}_j} \beta_{j,S} \prod_{k \in S} \alpha_k \Big),$$

where $\mathcal{K}_j = \{k \in [K] : q_{j,k} = 1\}$ *is the set of required attributes by item* j, *and* $f(\cdot)$ *is a link function. The link function is usually taken to be the identity, log, or logistic link. In this work we use the identity link. The coefficients can be interpreted as following:* $\beta_{j,\emptyset}$ *is the probability of a positive response for the most incapable subjects with* $\boldsymbol{\alpha} = \mathbf{0}$*;* $\beta_{j,\{k\}}$ *is the increase in the probability of a positive response for the subjects with* $\alpha_k = 1$ *compared to those with* $\alpha_k = 0$*;* $\beta_{j,S}$ *is the increase in the response probability for subjects with* $\{\alpha_k = 1, k \in S\}$ *compared to those missing one of the attributes in* S. *By incorporating all the interactions among the required attributes, the GDINA model is one of the most general CDMs.*

From a broader perspective, the aforementioned three CDMs belong to a general family of finite mixture models called restricted latent class models (RLCMs; Haertel 1989; Xu 2017). One common restriction is that all the capable subjects with all the required attributes have the same and highest item response parameters, that is,

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \succeq \boldsymbol{q}_j} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \succeq \boldsymbol{q}_j} \theta_{j,\boldsymbol{\alpha}} \;\geq\; \theta_{j,\boldsymbol{\alpha}'} \;\geq\; \theta_{j,\mathbf{0}}, \text{ for any } \boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_j, \tag{4}$$

where we write $\boldsymbol{\alpha} \succeq \mathbf{q}_j$ if $\alpha_k \geq q_{j,k}$ for all $k = 1, ..., K$, and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$ otherwise.

To fit CDMs, popularly used parametric methods include marginal maximum likelihood estimation (MMLE) through EM algorithms (de la Torre, 2009, 2011) and MCMC techniques (DiBello et al., 2007; von Davier, 2005). Chiu, Köhn, Zheng, and Henson (2016) also proposed a joint maximum likelihood estimation (JMLE) method for fitting CDMs. The parametric estimation methods usually perform well when there are sufficiently large data. However, as found in recent studies (Chiu and Köhn, 2019a; Chiu, Sun, and Bian, 2018), they may either produce inaccurate estimates with small sample sizes or suffer from high computational costs. This has lead researchers to consider nonparametric methods, which are reviewed in Section 2.2.

2.2 Nonparametric Methods

As the name suggests, nonparametric methods no longer depend on parametric model assumptions. Instead of modeling item response functions, nonparametric methods directly classify the subjects to latent classes by minimizing the distance between subject's observed item responses and the centers of the latent classes. Two popular examples of nonparametric methods are the NPC and the GNPC methods, which compare the subject's observed item responses to the so-called ideal response vectors of each proficiency-class. Different CDMs define the ideal response vectors differently. For example, as specified in (2) or (3), the ideal response in the DINA or DINO model will be 1 only if the subject possesses all the required attributes or one of the required attributes, respectively. In the following, we give a brief introduction of the NPC and the GNPC methods. Please refer to Chiu and Köhn (2019b), Chiu and Douglas (2013) and Chiu et al. (2018) for more details.

For the NPC method, we use $M = 2^K$ to denote the total number of proficiency latent classes (i.e., attribute profiles), and for $m = 1, \ldots, M$, we write $\boldsymbol{\eta}_m = (\eta_{1,m}, \eta_{2,m}, \ldots, \eta_{J,m})$ as the ideal response vector for the mth proficiency-class, where $\eta_{j,m}$ can be the DINA or DINO ideal response. Given the ideal response vectors for each proficiency class, a subject is classified to the closest proficiency class that minimizes the distance between the subject's observed responses and the ideal responses:

$$\hat{\boldsymbol{\alpha}}_i = \underset{m \in \{1,2,\ldots,M\}}{\arg\min} \; d(\boldsymbol{x}_i, \boldsymbol{\eta}_m),$$

where $d(\cdot)$ is a distance function. For example, in Chiu and Douglas (2013), they used the Hamming

distance:

$$d_H(\boldsymbol{x}, \boldsymbol{\eta}) = \sum_{j=1}^{J} |x_j - \eta_j|.$$

In the NPC method, the ideal responses are either the DINA ideal responses or the DINO ideal responses, which are all binary; thus, the absolute difference will be 0 if the observed response is equal to the ideal response, and 1 otherwise. Moreover, because the observed and the ideal responses are all binary, the L_2 distance will lead to the same results as the Hamming distance in the NPC method.

Due to its dependence on the DINA or DINO model assumptions, which define two extreme relations between $\boldsymbol{q}$ and $\boldsymbol{\alpha}$, the NPC method may not be sufficiently flexible. The GNPC method addresses this issue by considering a more general ideal response that represents a weighted average of the ideal responses of the DINA and DINO models, as in:

$$\eta_{j,m}^{(w)} = w_{j,m}\eta_{j,m}^{\text{DINA}} + (1 - w_{j,m})\eta_{j,m}^{\text{DINO}},$$

where $w_{j,m}$ is the weight for the jth item and the mth proficiency class. We use $\boldsymbol{\eta}_m^{(w)} = (\eta_{1,m}^{(w)}, \ldots, \eta_{J,m}^{(w)})$ to denote the weighted ideal response vector for the mth proficiency class in the GNPC method. To get the estimates of the weights, Chiu et al. (2018) proposed to minimize the L_2 distance between the responses to item j and the weighted ideal responses $\eta_{j,m}^{(w)}$:

$$d_{jm} = \sum_{i \in C_m} \left(x_{ij} - \eta_{j,m}^{(w)}\right)^2, \tag{5}$$

where $\{C_m\}_{m=1}^{M}$ is the partition of the subjects into M proficiency classes. Minimizing (5) leads to

$$\hat{w}_{j,m} = 1 - \bar{x}_{j,C_m}, \quad \hat{\eta}_{j,m}^{(w)} = \bar{x}_{j,C_m},$$

where $\bar{x}_{j,C_m} = |C_m|^{-1} \sum_{i \in C_m} x_{ij}$, the mean of the jth item responses for subjects in the mth proficiency class, and $|C_m|$ is the number of subjects in C_m. Because the true memberships are unknown, they proposed to iteratively estimate the memberships and the ideal response vectors. Specifically, starting with an initial partition of the subjects, the ideal response vectors are chosen to minimize the L_2 distance $\sum_{m=1}^{M} \sum_{i \in C_m} \sum_{j=1}^{J} (x_{ij} - \eta_{j,m}^{(w)})^2$. The memberships of the subjects are then determined by minimizing the L_2 distance between the observed responses of a subject and the ideal response vectors estimated from the former step, as in, $\hat{\boldsymbol{\alpha}}_i = \arg\min_{m \in \{1,2,\ldots,M\}} d\big(\mathbf{x}_i, \hat{\boldsymbol{\eta}}_m^{(w)}\big)$.

To implement the GNPC method, start with some initial values at $t = 0$ step. At the $(t+1)$th step, update the estimates as follows:

$$\hat{\boldsymbol{\alpha}}_i^{(t+1)} = \underset{m\in\{1,2,\ldots,M\}}{\arg\min}\, d\big(\boldsymbol{x}_i, \hat{\boldsymbol{\eta}}_m^{(w)(t)}\big), \quad \hat{\eta}_{j,m}^{(w)(t+1)} = \bar{\boldsymbol{x}}_{j,\hat{C}_m^{(t+1)}},$$

where $\hat{\boldsymbol{\eta}}_m^{(w)(t)}$ is the estimated centroids obtained in step t, and $\hat{C}_m^{(t+1)}$ is the partition of the subjects based on $\big\{\hat{\boldsymbol{\alpha}}_i^{(t+1)}\big\}_{i=1}^N$. Chiu et al. (2018) demonstrated through simulation studies that, compared to parametric methods, the nonparametric methods generally performed better in small-scale test settings.

3 A General Estimation Framework

In this section, we propose a unified estimation framework that subsumes both the parametric and nonparametric models considered in Section 2. This approach would facilitate a better statistical understanding of the relationship between the two families of CDM estimations.

For the parametric methods, we shall focus on the joint estimation of the subjects' latent classes $(\boldsymbol{\alpha}_i)_{i=1}^n$ and the model parameters. Considering the joint maximum likelihood estimation for parametric CDMs and the nonparametric estimation approaches as introduced in Section 2, we can see that the $\boldsymbol{\theta}$ in the parametric models and the ideal response vectors $\boldsymbol{\eta}$ in the nonparametric methods are closely related, both denoting a certain "centroid" of the responses of the latent classes under different model assumptions. For instance, $\theta_{j,\boldsymbol{\alpha}} = P(x_j = 1 \mid \boldsymbol{\alpha})$ can be viewed as the statistical population average (center) of the responses to item j of those subjects with attribute profile $\boldsymbol{\alpha}$, whereas $\eta_{j,\boldsymbol{\alpha}}$ corresponds to the nonparametric clustering center of the responses to item j of those in cluster $\boldsymbol{\alpha}$. Therefore, similarly to the nonparametric clustering methods, the joint maximum likelihood estimation of parametric model can be viewed as minimization of some "distance" function, introduced by the negative log-likelihood, between the observed responses and the "centroid" responses $\boldsymbol{\theta}$.

Motivated by this observation, we propose a unified estimation framework for both the parametric and nonparametric methods. Specifically, we let $\boldsymbol{A} = \big(\boldsymbol{\alpha}_i\big)_{i=1}^N$ denote a class membership matrix for N subjects. Based on the membership matrix $\boldsymbol{A}$, we can obtain a partition of N subjects into 2^K proficiency classes, denoted by $\boldsymbol{C}(\boldsymbol{A}) = \big\{C_{\boldsymbol{\alpha}}(\boldsymbol{A}) : \boldsymbol{\alpha} \in \{0,1\}^K\big\}$, where $C_{\boldsymbol{\alpha}}(\boldsymbol{A})$ denotes the set of subjects whose latent patterns are specified as $\boldsymbol{\alpha}$ by $\boldsymbol{A}$. For each latent class $\boldsymbol{\alpha} \in \{0,1\}^K$, we use $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$ to denote the "centroid" parameters for both parametric and nonparametric methods. Our proposed

estimators for the latent attributes and centroid parameters are obtained by minimizing a loss function of $(\boldsymbol{A}, \boldsymbol{\mu})$ as follows:

$$L(\boldsymbol{A}, \boldsymbol{\mu}) := \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i \in C_{\boldsymbol{\alpha}}(\boldsymbol{A})} l(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}}), \tag{6}$$

and the corresponding estimators are $(\hat{\boldsymbol{A}}, \hat{\boldsymbol{\mu}}) = \underset{(\boldsymbol{A}, \boldsymbol{\mu})}{\arg\min} L(\boldsymbol{A}, \boldsymbol{\mu})$. In (6), $l(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}})$ is a loss function that measures the distance between the ith subject's response vector $\boldsymbol{x}_i$ and the centroid of latent class $\boldsymbol{\alpha}$. Specifically, the loss function takes the additive form $l(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}}) = \sum_{j=1}^{J} l(x_{ij}, \mu_{j,\boldsymbol{\alpha}})$, where we abuse the notation $l(\cdot, \cdot)$ a little, and when the loss function takes two vectors, we use it to denote the summation of the element-wise losses. In this work, we also assume that $l(x_{ij}, \mu_{j,\boldsymbol{\alpha}})$ is continuous in $\mu_{j,\boldsymbol{\alpha}}$. Note that (6) can also be expressed as

$$L(\boldsymbol{A}, \boldsymbol{\mu}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i \in C_{\boldsymbol{\alpha}}(\boldsymbol{A})} l(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}}) = \sum_{i=1}^{N} \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} I\{\boldsymbol{\alpha}_i = \boldsymbol{\alpha}\} \cdot l(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}}) = \sum_{i=1}^{N} l(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}_i}), \tag{7}$$

which corresponds to a joint estimation of $(\boldsymbol{A}, \boldsymbol{\mu})$ under the loss function $l(\cdot, \cdot)$. From the joint estimation perspective, we can show that, with appropriate loss functions (e.g., L_1, L_2, cross-entropy) and constraints on the centroids (e.g., centroids based on the ideal responses, weighted ideal responses, or specific CDM assumptions), the proposed framework can provide estimates for all the models discussed in Section 2. The examples below demonstrate how the NPC method, the GNPC method, and parametric estimation of the DINA and GDINA models can be derived from the proposed framework using various loss functions and centroid constraints.

Example 4 (NPC)**.** *In the proposed framework, let the ideal responses under the NPC method be the centroids, that is,* $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = \boldsymbol{\eta}_{\boldsymbol{\alpha}}$*. If we use the* L_1 *loss function* $l(x_{ij}, \eta_{j,\boldsymbol{\alpha}}) = |x_{ij} - \eta_{j,\boldsymbol{\alpha}}|$*, then our proposed framework will become exactly the NPC method. Recall that in the NPC method, the ideal response vectors* $\boldsymbol{\eta}_{\boldsymbol{\alpha}}$ *are determined by pre-specified model assumptions (either the DINA or the DINO); thus, we only need to classify each subject to the closest proficiency class.*

Example 5 (GNPC)**.** *Recall that in the GNPC method, the ideal response is defined as* $\eta_{j,m}^{(w)} = w_{j,m}\eta_{j,m}^{DINA} + (1 - w_{j,m})\eta_{j,m}^{DINO}$*, a weighted average of the DINA ideal response and the DINO ideal response. Note that for proficiency classes and items such that* $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$*, we have* $\eta_{\boldsymbol{\alpha},j}^{DINA} = \eta_{\boldsymbol{\alpha},j}^{DINO} = 1$*, and for* $\boldsymbol{\alpha} \odot \boldsymbol{q}_j = \boldsymbol{0}$*, where* $\odot$ *denotes the elementwise multiplication of vectors, we have* $\eta_{\boldsymbol{\alpha},j}^{DINA} = \eta_{\boldsymbol{\alpha},j}^{DINO} = 0$*. In such cases, the weights in fact do not affect the weighted ideal responses since the*

DINA and the DINO models have the same ideal responses. Therefore, if we constrain $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (\mu_{j,\boldsymbol{\alpha}}, j = 1, \ldots, J)$ *in* (6)*, such that* $\mu_{j,\boldsymbol{\alpha}} = 1$ *if* $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$*,* $\mu_{j,\boldsymbol{\alpha}} = 0$ *if* $\boldsymbol{\alpha} \odot \boldsymbol{q}_j = \mathbf{0}$*, and* $\mu_{j,\boldsymbol{\alpha}} = \eta^{(w)}_{j,m}$ *as defined in the GNPC for the rest of the items, while at the same time use the* L_2 *loss function* $l(x_{ij}, \eta_{j,\boldsymbol{\alpha}}) = (x_{ij} - \eta_{j,\boldsymbol{\alpha}})^2$*, then the criterion in* (6) *is equivalent to the GNPC method.*

Example 6 (DINA)**.** *Let's consider the cross-entropy loss (i.e., the negative log-likelihood function),*

$$l(x_{ij}, \mu_{j,\boldsymbol{\alpha}}) = -\big(x_{ij} \log \mu_{j,\boldsymbol{\alpha}} + (1 - x_{ij}) \log(1 - \mu_{j,\boldsymbol{\alpha}})\big). \tag{8}$$

In addition, if we constrain the centroids to satisfy the following conditions:

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \succeq \boldsymbol{q}_j} \mu_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \succeq \boldsymbol{q}_j} \mu_{j,\boldsymbol{\alpha}} \geq \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j} \mu_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j} \mu_{j,\boldsymbol{\alpha}},$$

that is, all the capable subjects share the same higher item positive probabilities, whereas all the incapable subjects share the same lower item probabilities, then the proposed criterion (6) *becomes the JMLE criterion for the DINA model. Moreover, the centroids here correspond to item response parameters* $\boldsymbol{\theta}$ *for each latent class in the DINA model.*

Example 7 (GDINA)**.** *In Example 6, we can put the following constraints on the centroids:* $\mu_{j,\boldsymbol{\alpha}} = \mu_{j,\boldsymbol{\alpha}'}$*, if* $\boldsymbol{\alpha}_{\mathcal{K}_j} = \boldsymbol{\alpha}'_{\mathcal{K}_j}$*, where* $\boldsymbol{\alpha}_{\mathcal{K}_j} = (\alpha_k)_{k \in \mathcal{K}_j}$ *is the sub-vector of* $\boldsymbol{\alpha}$ *on the set* $\mathcal{K}_j$*, and* $\mathcal{K}_j = \{k \in [K] : q_{j,k} = 1\}$ *is the set of required attributes by item* j*. Equivalently, these constraints will result in the same centroid parameters for any two latent patterns sharing the same values on the required attributes of item* j*, which is a GDINA model assumption. Furthermore, if we take the same loss functions as in Example 6, it will result in the JMLE criterion for the GDINA model. Again, the centroids correspond to item response parameters* $\boldsymbol{\theta}$ *for each proficiency class.*

As demonstrated in the above examples, by taking different loss functions and different constraints on the centroid of each latent class, our proposal (6) provides a general estimation framework bridging both the parametric and nonparametric methods in the literature. The parametric estimation approaches mostly use the cross-entropy loss (negative log-likelihood) function, whereas the nonparametric approaches use the L_1 or L_2 distance measures. The analogous roles of negative log-likelihood for a parametric CDM and the distance function for a nonparametric CDM were also noted in Chiu et al. (2018).

It can be noted that the proposed estimation criterion (6) does not directly use the information

pertaining to the population distribution of the latent attribute profiles, which differentiates it from marginal likelihood estimation. As the population proportion of each latent class of attribute profiles may also provide useful information for the model estimation, we propose to further generalize (6) by including the proportion parameters in the loss function as follows:

$$L(\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\pi}) := \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i \in C_{\boldsymbol{\alpha}}(\boldsymbol{A})} \Big(l(\boldsymbol{x}_i, \boldsymbol{\mu}_{\boldsymbol{\alpha}}) + h(\pi_{\boldsymbol{\alpha}}) \Big), \tag{9}$$

where $l(\cdot, \cdot)$ is the loss function as in (6), and $h(\cdot)$ is a continuous nonincreasing regularization function of the proportion parameter $\pi_{\boldsymbol{\alpha}}$, which denotes the population proportion of latent class $\boldsymbol{\alpha}$. As can be seen from (9), the loss function L depends on both the centroids and the class proportions, with one part measuring the distance between a subject's response $\boldsymbol{x}_i$ and the centroid of a latent class $\boldsymbol{\mu}_{\boldsymbol{\alpha}}$, and the other part involving a regularization of class proportions.

Implicitly, Examples 4–7 take $h(\pi_{\boldsymbol{\alpha}}) = 0$. When we take the loss function $l(x_{ij}, \mu_{j,\boldsymbol{\alpha}})$ to be the cross-entropy loss function as in (8), and let $h(\pi_{\boldsymbol{\alpha}}) = -\log \pi_{\boldsymbol{\alpha}}$, then (9) becomes

$$L(\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i \in C_{\boldsymbol{\alpha}}(\boldsymbol{A})} \Big(l(x_{ij}, \mu_{j,\boldsymbol{\alpha}}) - \log \pi_{\boldsymbol{\alpha}} \Big) = -\sum_{i=1}^{N} \log \Big\{ \pi_{\boldsymbol{\alpha}_i} \times Lik(\boldsymbol{x}_i; \boldsymbol{\mu}_{\boldsymbol{\alpha}_i}) \Big\}, \tag{10}$$

where $Lik(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{\alpha}}) = P(\boldsymbol{x} \mid \boldsymbol{\mu}_{\boldsymbol{\alpha}})$ is the likelihood function for latent class $\boldsymbol{\alpha}$ and observation $\boldsymbol{x}$, and $\boldsymbol{\mu}_{\boldsymbol{\alpha}} = (\mu_{j,\boldsymbol{\alpha}}, j = 1, \ldots, J)$ is the corresponding model parameters with $\mu_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}} = P(x_{ij} = 1 \mid \boldsymbol{\alpha})$. Note that $\pi_{\boldsymbol{\alpha}_i} \times Lik(\boldsymbol{x}_i; \boldsymbol{\mu}_{\boldsymbol{\alpha}_i})$ in the RHS of (10) corresponds to the *complete-data* likelihood of $(\boldsymbol{\alpha}_i, \boldsymbol{x}_i)$; therefore, the loss function (10) is in fact the complete-data log-likelihood of $(\boldsymbol{A}, \boldsymbol{X})$.

The loss function (10) also corresponds to the extension of the classification maximum likelihood (CML) criterion (Celeux and Govaert, 1992) applied to the CDM setting. In Examples 6 and 7, using the loss function as in (10) corresponds to the CML criterion for the DINA or GDINA model respectively. It can be noted that the CML differs from the JMLE in that the former has an additional term $\log \pi_{\boldsymbol{\alpha}}$ in the loss function to make use of the information in the proportion parameters. The CML is also closely related to the EM estimation for the marginal MLE in that the CML directly maximizes the complete-data log-likelihood whereas the EM algorithm maximizes the expected complete-data log-likelihood with respect to the posterior distribution of the latent variables. Finally, it can also be underscored that, by incorporating a wide range of loss functions, the proposed criterion (9) is a generalization of the CML criterion (10).

To implement the unified estimation framework, we develop an algorithm to minimize (9). The algorithm is a general iterative algorithm to classify each subject to the closet proficiency class. Starting from initial values, the current loss for each subject's responses and the centroid of each latent class is first computed, after which the subject is assigned to the closest latent class that minimizes the loss. Based on the assigned memberships, the estimates for the centroids and class proportions are updated. The details of the steps are shown in Algorithm 1.

Algorithm 1: General Iterative Classification Algorithm

Input : Binary response matrix $\boldsymbol{X} \in \{0,1\}^{N\times J}$ and structural Q-matrix $\boldsymbol{Q} \in \{0,1\}^{J\times K}$

Initialize $\hat{\boldsymbol{A}}^{(0)}$, $\hat{\boldsymbol{\mu}}^{(0)}$ and $\hat{\boldsymbol{\pi}}^{(0)}$.

while *convergence not reached* **do**

At the $(t+1)^{th}$ iteration,

Step 1: Compute the current loss between $\boldsymbol{x_i}$ and the centroid of each proficiency class,

$$l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}^{(t)}) + h(\hat{\pi}_{\boldsymbol{\alpha}}^{(t)}),\ i = 1,\ldots,N,\ \boldsymbol{\alpha} \in \{0,1\}^K.$$

Step 2: Assign each $\boldsymbol{x}_i$ to the closest proficiency class, as in,

$$\hat{\boldsymbol{\alpha}}_i^{(t)} = \arg\min_{\boldsymbol{\alpha}}\ l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}^{(t)}) + h(\hat{\pi}_{\boldsymbol{\alpha}}^{(t)}),\ i = 1,\ldots,N.$$

and obtain the resulting partition $\hat{\boldsymbol{C}}^{(t)} := \boldsymbol{C}(\hat{\boldsymbol{A}}^{(t)})$.

Step 3: Compute the centroid and proportion of each proficiency class,

$$(\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}^{(t+1)}, \hat{\boldsymbol{\pi}}_{\boldsymbol{\alpha}}^{(t+1)}) = \arg\min_{(\boldsymbol{\mu},\boldsymbol{\pi})} \sum_{i\in\hat{C}_{\boldsymbol{\alpha}}^{(t)}} \Big(l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}^{(t)}) + h(\hat{\pi}_{\boldsymbol{\alpha}}^{(t)})\Big),\ \boldsymbol{\alpha} \in \{0,1\}^K.$$

Output: $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{\pi}}$.

In the CDM context, certain proficiency classes share the same item response parameters for each item given a particular Q-matrix. For example, for all CDMs, any $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$, has the same item response function; for the DINA model, there are only two levels of probabilities for each item's item response function, and the capable classes with $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$ share the same item response function $1 - s_j$, and the incapable classes with $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j$ share the same item response function g_j. Based on this observation, under certain model assumptions, the proficiency classes can be partitioned into some equivalence classes for each item according to the Q-matrix. Specifically, for item j, let $\tilde{A}_j = \{\tilde{A}_{j,\boldsymbol{\alpha}} = \{\boldsymbol{\alpha}' : \mu_{j,\boldsymbol{\alpha}} = \mu_{j,\boldsymbol{\alpha}'}\}\}$. Under this partitioning, the proficiency classes in the same equivalence class will have the same item response probability for this item. For example, in a DINA model with two latent attributes, if $\boldsymbol{q}_j = (1,0)$, then the proficiency classes can be partitioned into

$\Big\{\big\{(0,0),(0,1)\big\},\big\{(1,0),(1,1)\big\}\Big\}$, where $\boldsymbol{\alpha} \in \big\{(0,0),(0,1)\big\}$ will have the same item response function, g_j, and $\boldsymbol{\alpha} \in \big\{(1,0),(1,1)\big\}$ will also share the same item response function, $1-s_j$. Therefore, by incorporating information of the Q-matrix and certain model assumptions, we can develop an iterative classification algorithm tailored for CDMs that updates the centroids associated with equivalence classes together.

To illustrate, if we let the negative log-likelihood function be the loss function as specified in (8), then Step 3 in Algorithm 1 simplifies to

$$\hat{\pi}_{\boldsymbol{\alpha}}^{(t+1)} = \frac{\sum_{i=1}^{N} I\{\hat{\boldsymbol{\alpha}}_i^{(t)} = \boldsymbol{\alpha}\}}{N}, \quad \hat{\mu}_{j,\boldsymbol{\alpha}}^{(t+1)} = \frac{\sum_{\boldsymbol{\alpha}' \in \tilde{A}_{j,\boldsymbol{\alpha}}} \sum_{i \in \hat{C}_{\boldsymbol{\alpha}'}^{(t)}} x_{ij}}{\sum_{\boldsymbol{\alpha}' \in \tilde{A}_{j,\boldsymbol{\alpha}}} |\hat{C}_{\boldsymbol{\alpha}'}^{(t)}|},$$

where $|\cdot|$ is the cardinality of a set. Based on this simplification, the estimated proportion parameters are the sample proportions based on the estimated partition of the subjects, and the estimated centroids are the corresponding sample means of the equivalence classes also based on the estimated partition. Moreover, if fixed and equal proportions, together with L_2 loss $l(x_{ij}, \mu_{j,\alpha}) = (x_{ij} - \mu_{j,\alpha})^2$ are used, the algorithm becomes the iterative algorithm for the GNPC method outlined in Chiu et al. (2018).

4 Analysis of the Proposed Framework

In this section, we provide a theoretical analysis of the proposed framework. We show that, under certain conditions, the proposed estimation framework will give consistent estimates. The consistency results can be regarded as extensions of those for the NPC and the GNPC methods developed in Wang and Douglas (2015) and Chiu and Köhn (2019a). In addition to the asymptotic results, we also provide an analysis of the proposed algorithm in the finite sample situations.

As we introduced in Section 2.1, all the parametric CDMs belong to the family of latent class models. Hence, in our following analysis, we assume a general latent class model as the underlying model. Our results below are also easily adapted to the Q-matrix restricted latent class models. We use $\theta_{j,\boldsymbol{\alpha}}^0$ to denote the true probability of providing a positive response for the jth item and latent pattern $\boldsymbol{\alpha}$, as in, $\theta_{j,\boldsymbol{\alpha}}^0 = \mathbb{P}(x_j = 1 \mid \boldsymbol{\alpha})$, and we use $\boldsymbol{\theta}_{\boldsymbol{\alpha}}^0 = (\theta_{1,\boldsymbol{\alpha}}^0, \ldots, \theta_{J,\boldsymbol{\alpha}}^0)$ to denote item probability vector for latent pattern $\boldsymbol{\alpha}$. We let $\boldsymbol{A}^0 = (\boldsymbol{\alpha}_i^0)_{i=1}^N$ denote the true latent pattern matrix of the N subjects to be classified. Before we establish the consistency results, we first make some mild assumptions.

Assumption 1. *The loss function $l(x, \mu)$ is Hölder continuous in μ on $[\tau, 1-\tau]$ for any $\tau \in (0, 0.5)$,*

and the total loss (9) is minimized at class means given the subjects' membership, as in, $\hat{\mu}_{j,\boldsymbol{\alpha}} = \sum_{i \in C_{\boldsymbol{\alpha}}} x_{ij}/|C_{\boldsymbol{\alpha}}|$.

Assumption 2. *$h(\cdot)$ in (9) is a continuous nonincreasing function of the proportion parameters.*

Assumption 3. *There exist constants $\delta_1, \delta_2 > 0$ such that $\lim_{J\to\infty}\{\min_{\boldsymbol{\alpha}\neq\boldsymbol{\alpha}'} J^{-1}\|\boldsymbol{\theta}^0_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}'}\|_1\} \geq \delta_1$, and $\delta_2 \leq \min_{j,\boldsymbol{\alpha}} \theta^0_{j,\boldsymbol{\alpha}} < \max_{j,\boldsymbol{\alpha}} \theta^0_{j,\boldsymbol{\alpha}} \leq 1 - \delta_2$, where $\|\cdot\|_1$ denotes the L_1 norm.*

Assumption 4. *There exists $\delta \geq 1$ such that*

$$\left|E[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}})] - E[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}^0_i})]\right| \geq \left|\theta^0_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}^0_i}\right|^{\delta}, \ \forall\ \boldsymbol{\alpha} \neq \boldsymbol{\alpha}^0_i. \tag{11}$$

One can easily check that the L_2 and cross-entropy (negative log-likelihood) loss functions given in Section 3 satisfy Assumption 1. Note that the second part of Assumption 1 is a natural requirement for the consistent estimation of $\theta^0_{j,\boldsymbol{\alpha}}$, as $\theta^0_{j,\boldsymbol{\alpha}}$ represents the population average of the responses of subjects in latent class $\boldsymbol{\alpha}$, that is, $\theta^0_{j,\boldsymbol{\alpha}} = \mathbb{P}(x_j = 1 \mid \boldsymbol{\alpha})$. Given the true memberships of the subjects, for an estimator $\hat{\mu}_{j,\boldsymbol{\alpha}}$ that is consistent for $\theta^0_{j,\boldsymbol{\alpha}}$, it must satisfy $\left|\hat{\mu}_{j,\boldsymbol{\alpha}} - \sum_{i\in C_{\boldsymbol{\alpha}}} x_{ij}/|C_{\boldsymbol{\alpha}}|\right| \to 0$ in probability by the law of large number. An interesting counterexample is the L_1 loss function, which does not satisfy this assumption because given the memberships, $\hat{\mu}_{j,\boldsymbol{\alpha}}$ that minimizes the L_1 loss function is the sample median instead of the sample mean. Since in the CDM setting the responses are binary, the sample median would be either 0 or 1, which makes $\hat{\mu}_{j,\boldsymbol{\alpha}}$ under the L_1 loss not a consistent estimator of $\theta^0_{j,\boldsymbol{\alpha}}$ even when the true memberships are known. In other words, the L_1 loss cannot provide consistent estimation of the centroid parameters while the L_2 and cross-entropy losses can, as to be shown in the following theorems. More generally, following the M-estimation theory (van der Vaart, 2000), the second part of Assumption 1 can be further relaxed to requiring $E_{\theta^0_{j,\boldsymbol{\alpha}}}[l(x_{ij}, \mu_{j,\boldsymbol{\alpha}})]$ has a unique minima at $\theta^0_{j,\boldsymbol{\alpha}}$ and some additional technical conditions. For the presentation brevity, here we shall use the current assumption, which is already broad enough for practical use.

Assumptions 2 and 3 ensure the identifiability of the model, and also keep the true parameters away from the boundaries of the parameter space. Particularly, the assumption $\lim_{J\to\infty}\{\min_{\boldsymbol{\alpha}\neq\boldsymbol{\alpha}'} J^{-1}\|\boldsymbol{\theta}^0_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}'}\|_1\} \geq \delta_1$ implies that there is sufficient information to distinguish any two different classes $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$, thus ensuring the completeness (Chiu et al., 2009) and identifiability conditions (Gu and Xu, 2020). It is also similar to Condition (b) in Wang and Douglas (2015):

Condition(b). *Define the set $\mathcal{A}_{m,m'} = \{j \mid \eta_{mj} \neq \eta_{m'j}\}$, where m and m' index the attribute profiles of different proficiency classes among all the $M = 2^K$ realizable proficiency classes; Card$(\mathcal{A}_{m,m'}) \to \infty$*

as $J \to \infty$.

Condition (b) in Wang and Douglas (2015) and Assumption 3 in our work are essentially stating that for any two different proficiency classes, there are infinitely many items such that the item response functions for these two proficiency classes are different.

The condition (11) in Assumption 4 also holds for the aforementioned loss functions in Section 3. For example, it is easy to check the condition (11) for the L_2 loss and the cross-entropy loss. For the L_2 loss, we have $E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}})\big] - E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}^0_i})\big] = (\theta^0_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}^0_i})^2$. For the cross-entropy loss, we have

$$
\begin{aligned}
& E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}})\big] - E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}^0_i})\big] \\
=& -\theta^0_{j,\boldsymbol{\alpha}_0} \log(\theta^0_{j,\boldsymbol{\alpha}}) - (1-\theta^0_{j,\boldsymbol{\alpha}^0_i}) \log(1-\theta^0_{j,\boldsymbol{\alpha}}) + \theta^0_{j,\boldsymbol{\alpha}^0_i} \log(\theta^0_{j,\boldsymbol{\alpha}^0_i}) + (1-\theta^0_{j,\boldsymbol{\alpha}^0_i}) \log(1-\theta^0_{j,\boldsymbol{\alpha}^0_i}) \\
=& D_{KL}\Big(p(\theta^0_{j,\boldsymbol{\alpha}}) \;||\; p(\theta^0_{j,\boldsymbol{\alpha}^0_i})\Big) \geq \frac{1}{2}\Big(\big|\theta^0_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}^0_i}\big| + \big|(1-\theta^0_{j,\boldsymbol{\alpha}}) - (1-\theta^0_{j,\boldsymbol{\alpha}^0_i})\big|\Big)^2 \\
=& 2(\theta^0_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}^0_i})^2,
\end{aligned}
$$

where $D_{KL}(\cdot \,||\, \cdot)$ is the Kullback-Leibler divergence, $p(\cdot)$ is the mass function of a Bernoulli distribution, and the last inequality follows from Theorem 1.3 in Popescu et al. (2016).

Similar to the analysis of the joint maximum likelihood estimation in Chiu et al. (2016), we assume that there is a calibration dataset that would give a statistically consistent estimator of the calibration subjects' latent class membership $\hat{\boldsymbol{A}}_c$, in the sense that $\mathbb{P}(\hat{\boldsymbol{A}}_c \neq \boldsymbol{A}^0_c) \to 0$ as $J \to \infty$. We use N_c and $\boldsymbol{A}^0_c$ to denote the sample size and the true membership matrix of the calibration dataset, respectively. Here the subscript c denotes the calibration dataset. Similar assumption is also made in the consistency theories of the GNPC method in Chiu and Köhn (2019a). In the next theorem, we show that the consistent membership estimator will give consistent estimators for the centroids of the latent classes.

Theorem 1. *Suppose the data conforms to CDMs that can be expressed in terms of general latent class models, and Assumptions 1-3 hold. Further assume that* $J \exp\big(-N_c\epsilon\big) \to 0$ *as* $J, N_c \to \infty$ *for any* $\epsilon > 0$. *If* $\hat{\boldsymbol{A}}_c$ *is a consistent estimator of* $\boldsymbol{A}^0_c$, *then* $\hat{\boldsymbol{\mu}}$ *is also consistent for* $\boldsymbol{\theta}^0$ *as* $J, N_c \to \infty$, *that is,* $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\theta}^0\|_\infty \xrightarrow{P} 0$ *as* $J, N_c \to \infty$, *where* $\|\cdot\|_\infty$ *is the supremum norm.*

Theorem 1 states that if we could get a consistent estimate of the calibration subjects' membership $\hat{\boldsymbol{A}}_c$, then the estimated centroids $\hat{\boldsymbol{\mu}}$ are also consistent for the true item response probabilities $\boldsymbol{\theta}^0$ in a uniform sense that all item parameters can be uniformly consistently estimated. The detailed proof is in online Appendix A.1. This result is similar to Lemmas 1 and 2 in Chiu and Köhn (2019a) under the GNPC framework, and Theorem 2 in Chiu et al. (2016) under the JMLE framework. Note that for the

GNPC method, the centroids are weighted averages of the ideal responses from the DINA and DINO models. As discussed in Example 5, if the DINA and DINO models have the same ideal responses (i.e., $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$ or $\boldsymbol{\alpha} \odot \boldsymbol{q}_j = \mathbf{0}$), then the corresponding centroid will be fixed to be 0 or 1, which thus does not lead to a consistent estimation of the corresponding item response probability $\theta^0_{j,\boldsymbol{\alpha}}$; however, note that for the nonparametric GNPC method, such a fixed centroid does not necessarily lead to inconsistency of $\hat{\boldsymbol{\alpha}}$. Here we allow all the centroid parameters to be free, and the consistency estimation is ensured as in Theorem 1.

The next theorem shows that if we start with a consistent membership $\hat{\boldsymbol{A}}_c$ obtained from the calibration dataset, and use the estimated centroids to classify the subjects, then the resulting classifications are also consistent for each subject.

Theorem 2. *Suppose Assumptions 1–4 and the assumptions of Theorem 1 hold. If we start with a consistent $\hat{\boldsymbol{A}}_c$ obtained from a calibration dataset to estimate the centroid $\hat{\boldsymbol{\mu}}$, then $\hat{\boldsymbol{\alpha}}_i$ obtained by Algorithm 1 is also a consistent estimator of $\boldsymbol{\alpha}_i$ for each $i = 1, \ldots, N$.*

To establish the consistency in Theorem 2, the following two lemmas are needed.

Lemma 1. *Suppose Assumptions in Theorem 2 hold. For each subject i, the true attribute pattern minimizes $E[l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}) + h(\hat{\pi}_{\boldsymbol{\alpha}})]$ with probability approaching 1 as $J \to \infty$, as in,*

$$P\Big(\boldsymbol{\alpha}_i^0 = \arg\min_{\boldsymbol{\alpha}} \; E\big[l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}) + h(\hat{\pi}_{\boldsymbol{\alpha}})\big]\Big) \to 1 \quad \text{as } J \to \infty.$$

Lemma 2. *Suppose Assumptions in Theorem 2 hold, then we have*

$$P\Big(\max_{\boldsymbol{\alpha}} \Big| \frac{1}{J} \sum_{j=1}^{J} \big(l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) - E[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}})]\big) \Big| \geq \epsilon \Big) \to 0, \;\; \text{as } J \to \infty.$$

Lemma 1 extends Proposition 1 in Wang and Douglas (2015) and Lemma 3 in Chiu and Köhn (2019a) to more general loss functions. Lemma 2 generalizes Proposition 3 in Wang and Douglas (2015) and Lemma 4 in Chiu and Köhn (2019a). The detailed proofs of Lemma 1, Lemma 2, and Theorem 2 are given in online Appendices A.2 – A.4. Note that Theorem 2 only gives the consistency for each $\boldsymbol{\alpha}_i$; however, we can further establish uniform consistency for all $\boldsymbol{\alpha}_i$, $i = 1, \ldots, N$, as shown in Theorem 3.

Theorem 3. *Suppose all the assumptions of Theorem 2 hold. Further assume that $N > J$, $N_c > J$ and for any $\epsilon > 0$, $N \exp(-J\epsilon) \to 0$. If we start with a consistent $\hat{\boldsymbol{A}}_c$ obtained from a calibration dataset, then $\hat{\boldsymbol{\alpha}}_i$ obtained from Algorithm 1 is uniformly consistent for $\boldsymbol{\alpha}_i$, for all $i = 1, \ldots, N$.*

Uniform consistency has also been established for specific nonparametric methods, such as Theorem 2 in Wang and Douglas (2015) and Theorem 2 in Chiu and Köhn (2019a). Our uniformly consistent result in Theorem 3 can be regarded as their generalization. Specifically, in Wang and Douglas (2015), they showed the uniform consistency for the NPC method, where the loss function is taken to be L_1 loss and the centroids are fixed, to be the ideal responses of the DINA or DINO model. In Chiu and Köhn (2019a), the authors generalize the uniform consistency for the NPC method to the GNPC method, where the loss function is L_2 loss and the centroids are weighted averages of ideal responses from the DINA and the DINO models.

The above analysis pertains the asymptotic properties of our framework. For finite-sample situations, we have the following theoretical properties for the proposed iterative algorithms in Section 3, which are established following the theory in Celeux and Govaert (1992).

Proposition 1. *Any sequence* $(\boldsymbol{A}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)})$ *obtained by Algorithm 1 decreases the criterion (9) and the sequence* $L(\boldsymbol{A}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)})$ *converges to a stationary value. Moreover, if for any fixed* $\boldsymbol{A}$*, the minima of the loss function* $L(\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\pi})$ *is well-defined, then the sequence* $(\boldsymbol{A}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)})$ *also converges to a stationary point.*

Proposition 1 indicates that the update sequence $(\boldsymbol{A}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)})$ from the proposed algorithm converges to a stationary point of the proposed criterion (9) with finite samples. Additionally, all the loss functions in the examples in Section 3 satisfy the condition that the minima is well-defined. Now, consider a smoothed version of $L(\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\pi})$,

$$L(\boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i=1}^{n} u_{i\boldsymbol{\alpha}} \Big(l(\boldsymbol{x_i}, \boldsymbol{\mu_\alpha}) + h(\pi_{\boldsymbol{\alpha}}) \Big),$$

where $\boldsymbol{U} = \{u_{im}\} \in [0,1]^{n \times 2^K}$ is a matrix with nonnegative entries and each column sums to one, which is called a standard classification matrix in Celeux and Govaert (1992). Recall that $L(\boldsymbol{A}, \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{\boldsymbol{\alpha}} \sum_{i \in C_{\boldsymbol{\alpha}}(\boldsymbol{A})} \Big(l(\boldsymbol{x}_i, \boldsymbol{\mu_\alpha}) + h(\pi_{\boldsymbol{\alpha}}) \Big) = \sum_{\boldsymbol{\alpha}} \sum_{i=1}^{n} I(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}) \Big(l(\boldsymbol{x}_i, \boldsymbol{\mu_\alpha}) + h(\pi_{\boldsymbol{\alpha}}) \Big)$. Therefore, $L(\boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\pi})$ can be regarded as a smoothed version, where the hard membership matrix $\boldsymbol{A}$ is replaced by $\boldsymbol{U}$. Note that the minimum of $L(\boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\pi})$ is attained when $\boldsymbol{U}$ is equal to some hard membership matrix $\boldsymbol{A}$.

Proposition 2. *Assume that* $L(\boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\pi})$ *has a local minimum at* $(\boldsymbol{U}^*, \boldsymbol{\mu}^*, \boldsymbol{\pi}^*)$ *and that the Hessian of* $L(\boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\pi})$ *exists and is positive definite at* $(\boldsymbol{U}^*, \boldsymbol{\mu}^*, \boldsymbol{\pi}^*)$*. Then there is a neighborhood of* $(\boldsymbol{U}^*, \boldsymbol{\mu}^*, \boldsymbol{\pi}^*)$ *such that starting with any* $(\boldsymbol{U}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\pi}^{(0)})$ *in that neighborhood, the resulting sequence* $(\boldsymbol{A}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)})$ *of the Algorithm 1 converges to* $(\boldsymbol{U}^*, \boldsymbol{\mu}^*, \boldsymbol{\pi}^*)$ *at a linear rate.*

Proposition 2 states that if we start with a good initial value which is close enough to the optimal point, then the update sequence will also converge to the optimal point. These two propositions give good finite-sample properties of our proposed estimation framework. The detailed proofs of Proposition 1 and Proposition 2 are given in online Appendix A.6 and A.7, respectively.

5 Simulation Studies

We conducted comprehensive simulations under a variety of settings to compare the performance of different methods. The methods compared were:

- NPC: the baseline method, where the centers are the ideal responses from the DINA model, and the loss function is the L_1 loss;
- GNPC: the centers are weighted averages of the ideal responses from the DINA and DINO models, and the loss function is the L_2 loss;
- GNPC + log penalty: add log penalties on the proportion parameters to the GNPC method, where the loss function is L_2 loss for the centroids plus the summation of the log functions of the proportion parameters;
- JMLE: the Joint Maximum Likelihood Estimate, where the centroid parameters are to be estimated, and the loss function is the negative log-likelihood;
- CMLE: the Classification Maximum Likelihood Estimate, where the centers and the loss function are the same as JMLE but with an additional term of class proportions as specified in (10);
- MMLE: the Marginal Maximum Likelihood Estimate obtained from the traditional EM algorithm under the DINA or GDINA model assumption.

MMLE, as one of, if not the the most commonly used estimation algorithm in the CDM literature, was included in the comparison to provide a more comprehensive understanding of how different CDM estimation methods perform.

For the underlying true models, we considered two different settings: all items conformed to the DINA, or all items conformed to the GDINA model. Following the simulation design in Chiu et al. (2018), the subjects' true latent attribute patterns were either drawn from a uniform distribution or derived from the multivariate normal threshold model. More specifically, for the uniform setting, each

latent pattern $\boldsymbol{\alpha}$ had the same probability $1/2^K$ of being drawn. For the multivariate normal setting, each subject's attribute profile was linked to a latent continuous ability vector $\boldsymbol{z} = (z_1, \ldots, z_K)' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with values along the main diagonal of $\boldsymbol{\Sigma}$ setting to 1 and the off-diagonal entries setting to either $r = 0.4$ or 0.8 for different levels of correlation. The latent continuous ability vector $\boldsymbol{z}$ was randomly sampled, and the kth entry of the attribute pattern was determined by

$$\alpha_{ik} = \begin{cases} 1, & z_{ik} \geq \Phi^{-1}\left(\frac{k}{K+1}\right) \\ 0, & \text{otherwise.} \end{cases}$$

where Φ is the inverse cumulative distribution function of standard normal distribution.

We considered different numbers of latent attributes ($K = 3$ or 5), different sample sizes ($N = 30$, 50, 200 or 500) and different number of items ($J = 30$ or 50). To ensure identifiability, we set the first two $K \times K$ submatrices of the Q-matrix to be identity matrices. The remaining items were randomly generated from all the possible latent patterns. When $K = 5$, the Q-matrix contained items that measured up to three attributes and was constructed the same way as that for $K = 3$. When the underlying model was the DINA or DINO model, different signal strengths were considered. Specifically we set $s = g = 0.1$ or 0.3. When the true model was the GDINA model, we also considered two different signal strength levels. One had the same item response functions as those specified in Chiu et al. (2018), which are listed in Table 1; the other setting contained larger noise as listed in Table 2.

$P(\boldsymbol{\alpha}_1)$	$P(\boldsymbol{\alpha}_2)$	$P(\boldsymbol{\alpha}_3)$	$P(\boldsymbol{\alpha}_4)$	$P(\boldsymbol{\alpha}_5)$	$P(\boldsymbol{\alpha}_6)$	$P(\boldsymbol{\alpha}_7)$	$P(\boldsymbol{\alpha}_8)$
0.2	0.9						
0.1	0.8						
0.1	0.9						
0.2	0.5	0.4	0.9				
0.1	0.3	0.5	0.9				
0.1	0.2	0.6	0.8				
0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.9

Table 1: Item response parameters for GDINA with small noise

$P(\boldsymbol{\alpha}_1)$	$P(\boldsymbol{\alpha}_2)$	$P(\boldsymbol{\alpha}_3)$	$P(\boldsymbol{\alpha}_4)$	$P(\boldsymbol{\alpha}_5)$	$P(\boldsymbol{\alpha}_6)$	$P(\boldsymbol{\alpha}_7)$	$P(\boldsymbol{\alpha}_8)$
0.3	0.7						
0.3	0.8						
0.3	0.4	0.7	0.8				
0.3	0.4	0.6	0.7				
0.2	0.3	0.6	0.7				
0.2	0.3	0.3	0.4	0.4	0.5	0.6	0.7

Table 2: Item response parameters for GDINA with large noise

To evaluate the performance of different methods, two metrics were used: the pattern-wise agreement rate (PAR) and the attribute-wise agreement rate (AAR), as defined below,

$$\text{PAR} = \frac{\sum_{i=1}^{N} I\{\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\}}{N}, \quad \text{AAR} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} I\{\hat{\alpha}_{ik} = \alpha_{ik}\}}{NK}.$$

For parametric methods including JMLE, CMLE, MMLE of the DINA and the GDINA models, we also calculate the Mean Squared Errors (MSEs) for item response probabilities θ's of each latent class. For each setting, we repeated 100 times and reported the obtained means of PAR, AAR and MSE. Note that the aforementioned methods are iterative in nature, hence, would be affected by how they are initialized. For comparability purposes, we treated the NPC method as the baseline in this work, and used its results to initialize all the other methods. Using the NPC to perform the initialization is a reasonable choice given its non-iterative nature. In the following result plots, we use DINA or GDINA to stand for the results of MMLE obtained by the EM algorithm under the corresponding model assumptions.

Result I: DINA

Figures 1 and 2 present the PARs and AARs when the underlying process followed the DINA model. Under the independent attribute (i.e., uniform) setting, the NPC performed the best, as expected, in almost all the cases. The JMLE performed similarly to the CMLE in most cases, and slightly better than CMLE when there were more latent attributes ($K = 5$) – this was so because the JMLE method correctly assumed that the true latent patterns were uniformly distributed. The GNPC produced similar results to the JMLE and the CMLE in most cases, but much worse results with large noises ($s = g = 0.3$) and more items ($J = 50$). Adding the log penalty to the GNPC method degraded the results under the uniform setting especially when the sample size was large, which is also expected since the GNPC method implicitly assumes a uniform penalty on the latent classes. In comparison, the

MMLE of the DINA and GDINA models did not perform as well as the others. This was particularly true when the noise was large and sample size was small.

Under the dependent attribute (i.e., multivariate normal) settings, although the NPC still performed the best in almost all the cases with moderate correlation ($r = 0.4$), it performed poorly with larger correlation ($r = 0.8$) and sample size ($N = 200$ or 500) as a consequence of more unequal latent patterns proportions. The MMLE of the DINA provided the best results when the sample size was larger ($N = 200/500$) and the correlation was large ($r = 0.8$), but did not perform well with smaller sample sizes. The GNPC and JMLE performed similarly when the noise was small, but the GNPC was much worse than the JMLE when the noise was large. Adding the log penalty on the proportions improved the performance of the GNPC method under the correlated settings, though still not as good as the CMLE method. In contrast, the CMLE performed uniformly well in almost all cases, and its advantages became more apparent when there were more latent attributes, and the correlation and the noise were large. Specifically, the CMLE performed similarly to the NPC when the sample sizes were small, and the MMLE of the DINA when the sample sizes were large. In almost all of the conditions, the MMLE of the GDINA did not perform as well as the other methods, which was not unexpected as the DINA was the true model.

Mean Squared Errors (MSEs) for the item response probabilities θ's using parametric methods including JMLE, CMLE and MMLE for the DINA and GDINA models are plotted in Figure 3. From the results, we can see that across different settings the MMLE for the DINA model gave the best item response probability estimates, while the MMLE for the GDINA model performed the worst. The JMLE and CMLE methods provided similar results. It is actually not surprising that the MMLE for the DINA performed the best for item response probability estimation since it correctly assumed a two-level DINA model and directly estimated the corresponding guessing and slipping parameters, while other methods did not have such prior knowledge about the underlying model.

Result II: GDINA

Figures 4 and 5 show the PARs and the AARs when the data conformed to the GDINA model under different settings. Based on the results, when the latent attributes were independent, the GNPC performed generally the best across the settings, whereas the JMLE ,the CMLE and the MMLE of the GDINA model improved with increasing sample size. As in the DINA cases, the log penalty on the proportions degraded the performance of the GNPC method under the independent setting. The

JMLE provided comparable or slightly better results than the CMLE, particularly when K was larger. As mentioned earlier, this is because the JMLE correctly assumed a uniform prior distribution for the latent attributes, whereas the CMLE, although made no assumptions, needed to estimate additional parameters.

Under the correlated latent attributes settings, adding the log penalties on the proportions to the GNPC method greatly improved the performance especially when the noise was large or correlation was high. GNPC+log penalty tended to provide the best results with small sample sizes, and the CMLE and the MMLE of the GDINA gave the best results with larger samples. Moreover, with larger noises, the CMLE method provided better results than the MMLE of the GDINA model particularly when there were more latent attributes. As the correlation became larger, with large sample sizes, the performance of the CMLE method became more similar to that of the MMLE of the GDINA, and better than the JMLE method, due to the proportions of latent attribute patterns no longer being equal. Based on the above analysis, one can note that the CMLE method was more robust to large noise.

The MSEs for the parametric methods including JMLE, CMLE and MMLE for the GDINA model under the GDINA settings are given in Figure 6. These three methods gave similar results in most cases, while JMLE and CMLE performed better than the MMLE of the GDINA settings especially when the number of attributes was large or noises were large.

Summary and Recommendations

Based on the above simulation results, we can see that there is no dominating method that performed uniformly better than other methods across all the simulation settings. Hence, the choice of the method should be based on the specific setting and other information we have at hand. In the following, we provide recommendations in practice under different circumstances.

If we can safely assumed that the true underlying model is the DINA model, then the NPC method would give good results if the latent attributes are independent. When the latent attributes are moderately correlated, either the NPC or the CMLE method is recommended. When the correlations are high among the latent attributes, the NPC and the CMLE would perform well with small sample sizes, whereas the CMLE and the MMLE of the DINA model would give better results with sufficiently large data sizes.

In situations where the true model is the GDINA model, the GNPC method will perform generally

well if the latent attributes are independent. When the latent attributes are correlated and the sample size is small, the GNPC augmented by the log penalties on the proportion parameters is preferred. However, when the sample is sufficiently large, the CMLE method is more robust. The CMLE method also performs well with small sample sizes when the noise is large.

Finally, if the true data-generating models are unknown, the CMLE method is recommended when the latent attributes are correlated. If the latent attributes are independent, the GNPC method is preferred. Moreover, when the sample size is large enough, the MMLE method for the GDINA model is also recommended. If the noise is small, the GNPC method will also perform well when the sample size is small, and augmenting the GNPC method with the log penalties on the proportion parameters will improve its performance under the correlated setting.

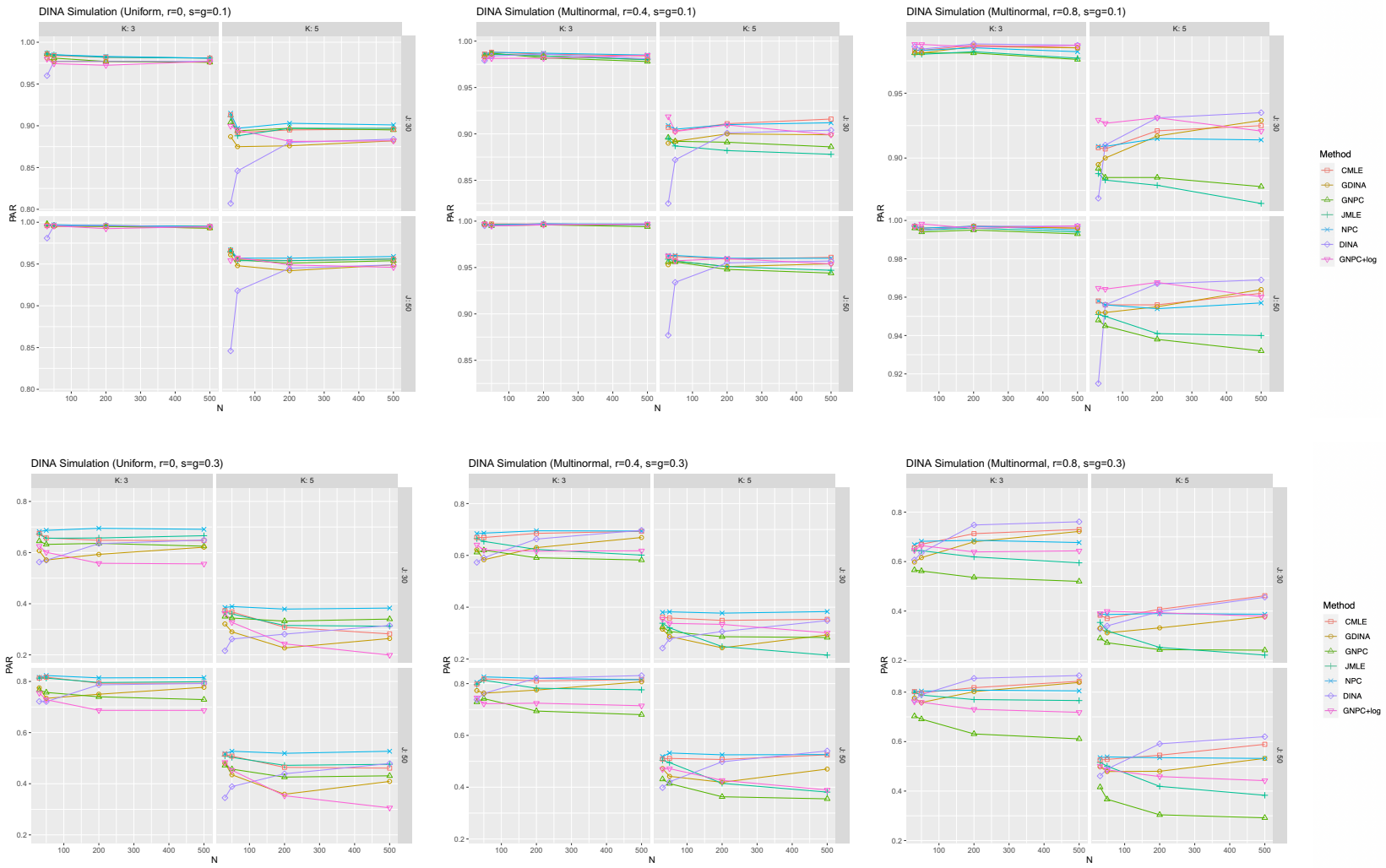

Figure 1: PARs when the data conformed to the DINA model

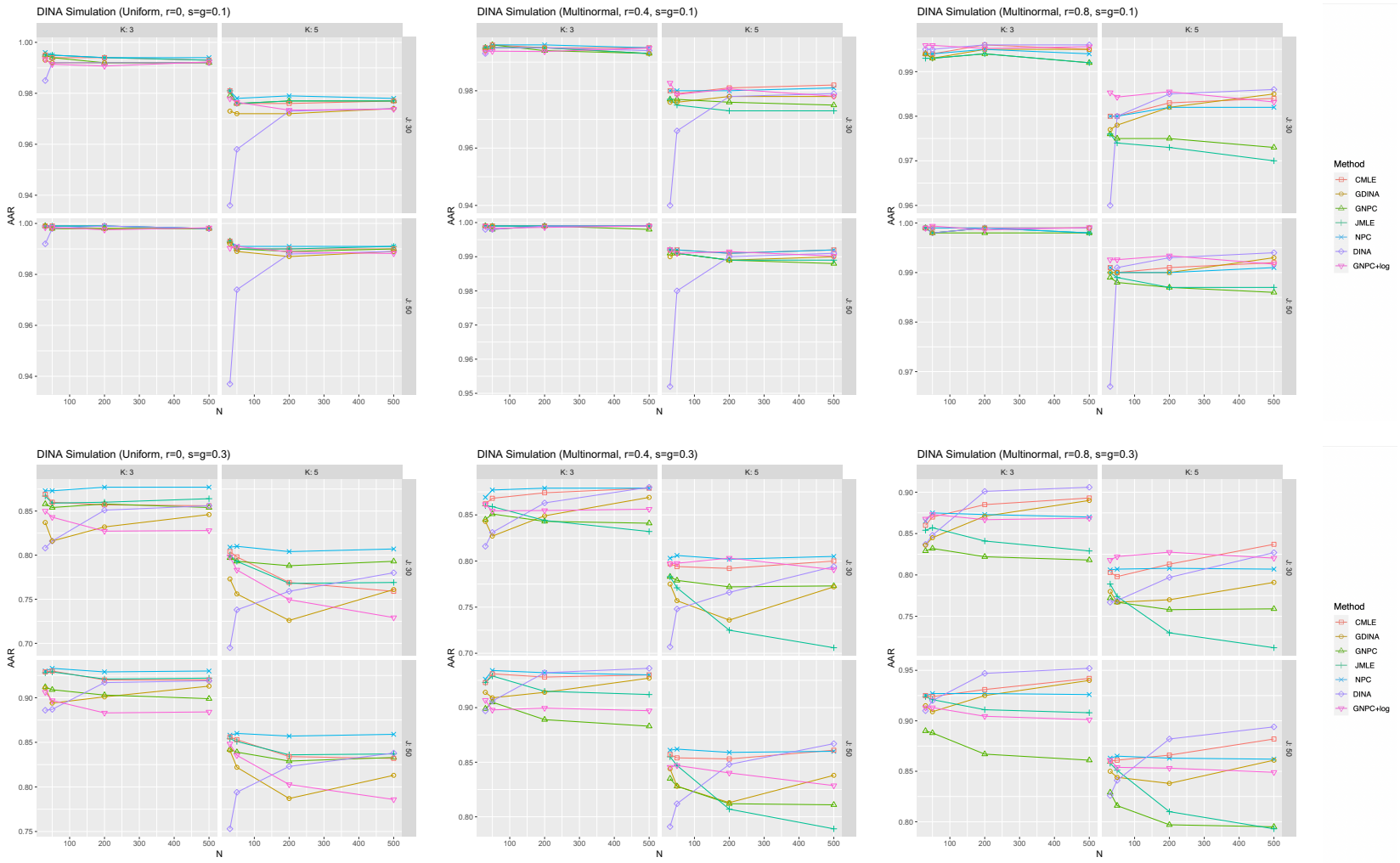

Figure 2: AARs when the data conformed to the DINA model

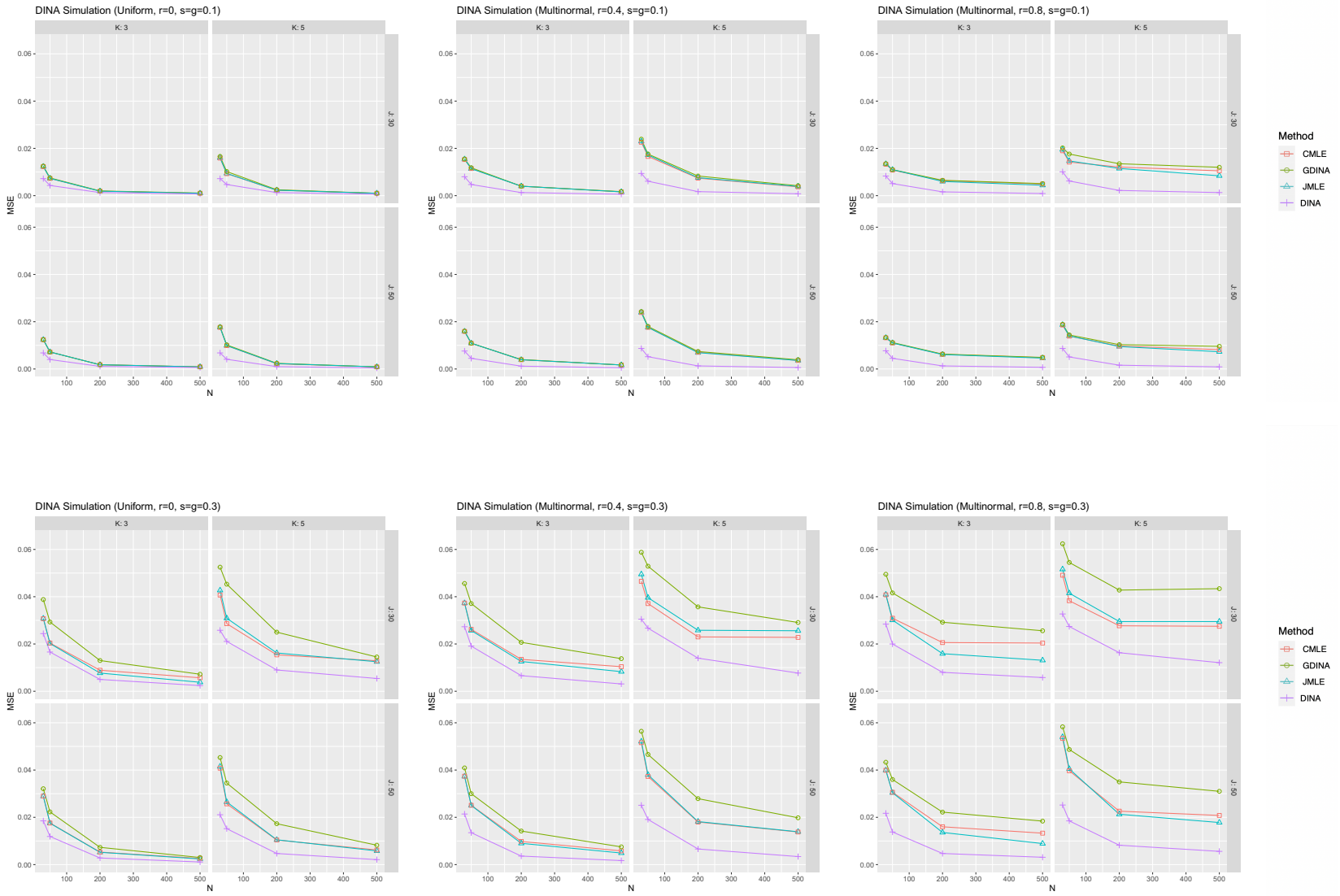

Figure 3: MSE when the data conformed to the DINA model

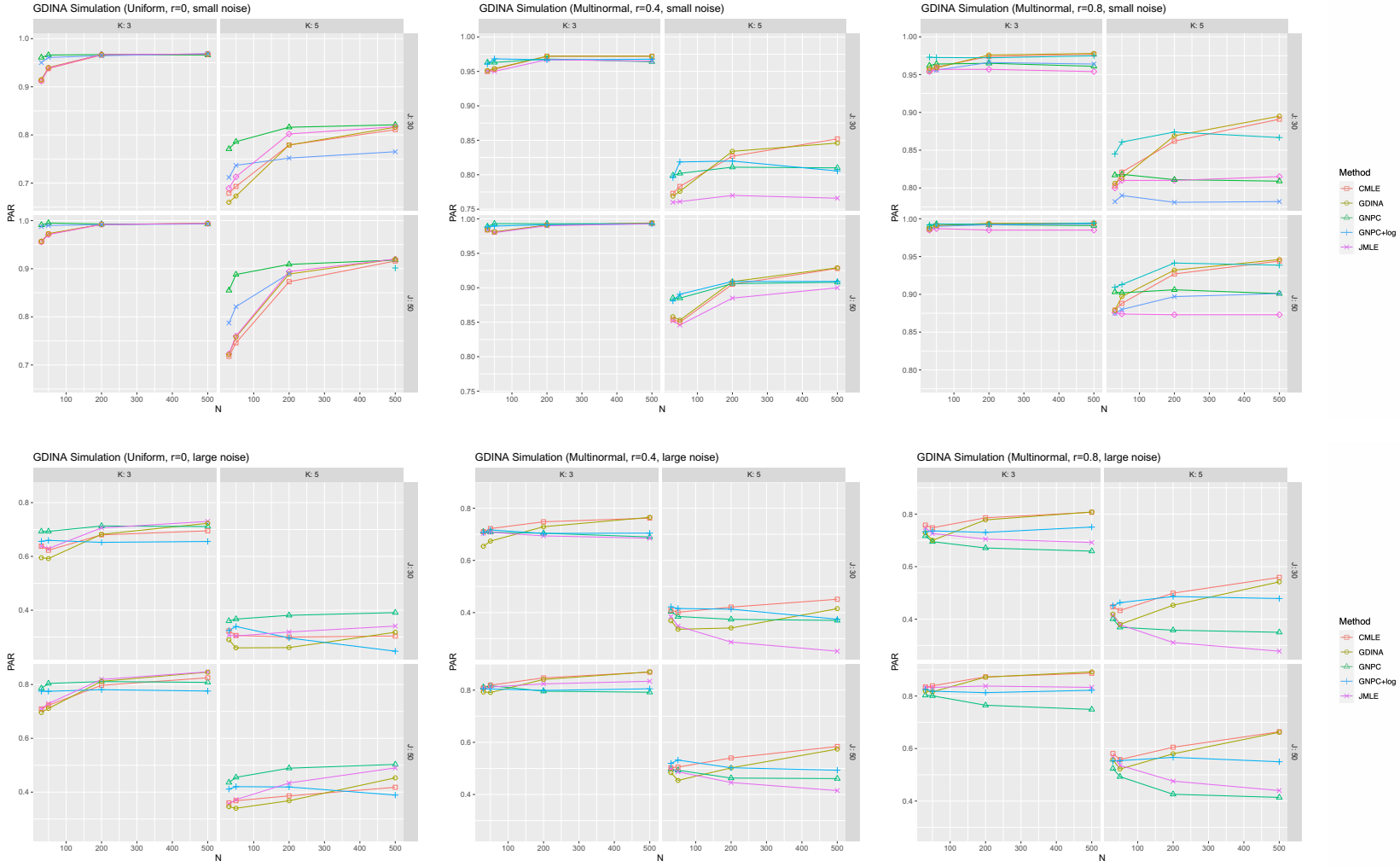

Figure 4: PARs when the data conformed to the GDINA model

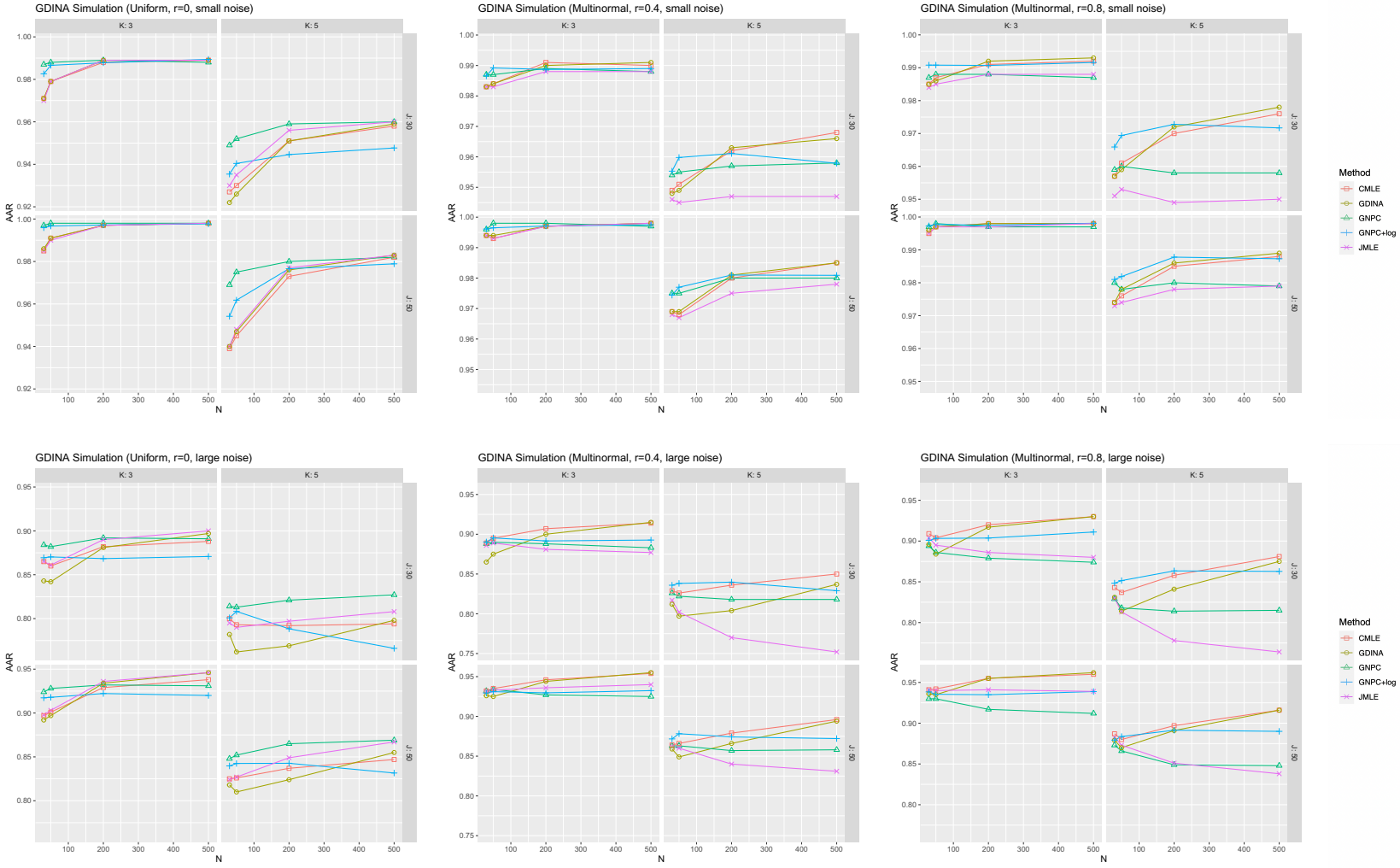

Figure 5: AARs when the data conformed to the GDINA model

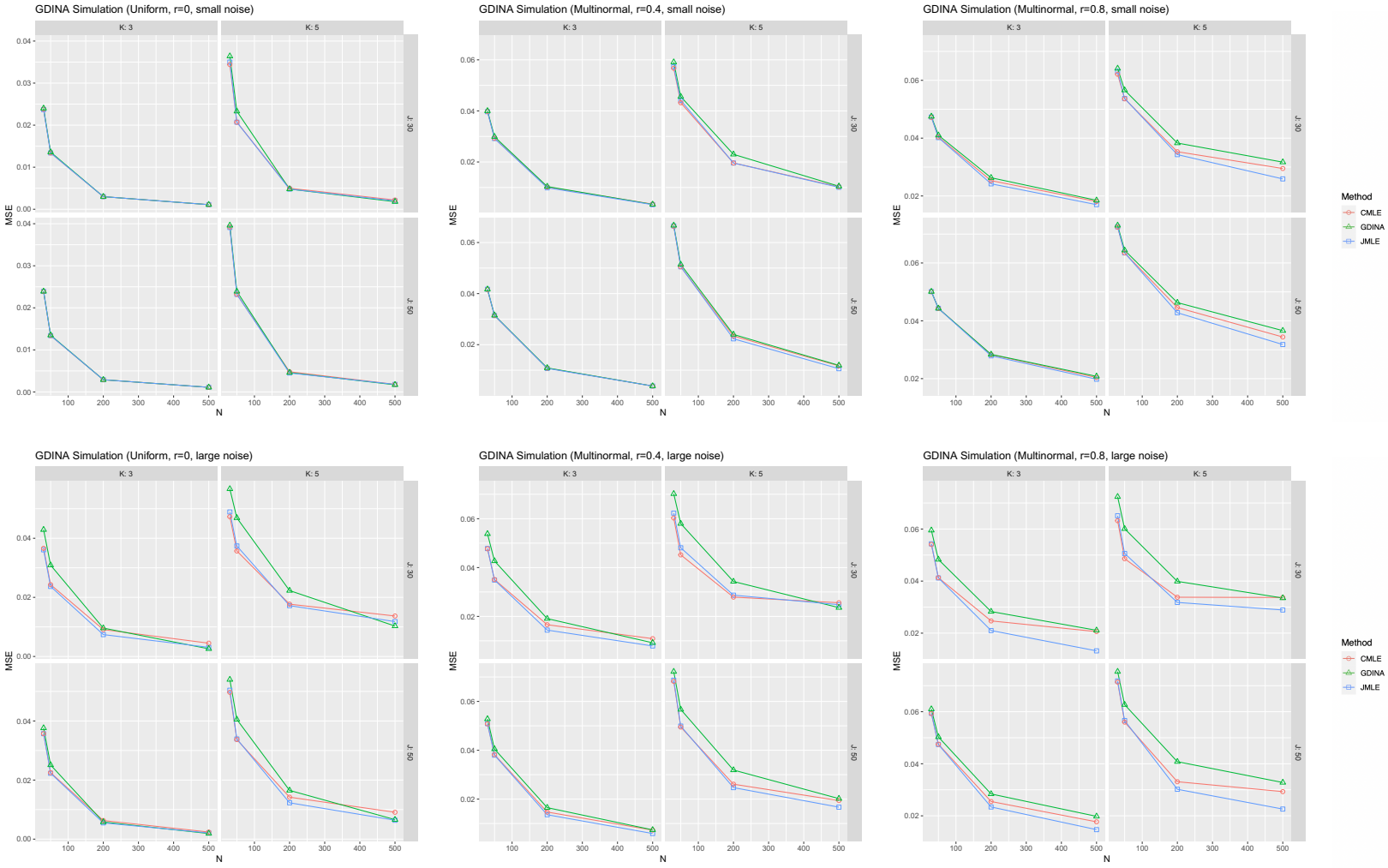

Figure 6: MSEs when the data conformed to the GDINA model

6 Discussion

In this paper, a unified estimation framework is proposed to bridge the parametric and nonparametric methods of cognitive diagnosis, and corresponding computational algorithms are developed. Specifically, by choosing different loss functions and potentially imposing additional constraints on the centroids of the proficiency classes, the proposed framework essentially provides estimations for both parametric cognitive diagnosis models and nonparametric methods for classifying subjects to proficiency classes. Moreover, we also provide theoretical analysis and establish consistency theories of the proposed framework. The simulation studies under various settings demonstrate the advantages and disadvantages of different methods.

In our proposed framework (9), we decompose the loss function into two additive parts. In addition to the losses between the responses and class centroids, we also put a regularization term on the class proportions. The regularization term can also play a role in selecting significant latent classes in the population. For instance, similar to the CML in Examples 6 and 7, a log-type penalty $h(\pi_{\boldsymbol{\alpha}}) = -\lambda \log(\pi_{\boldsymbol{\alpha}})$, where $\lambda > 0$ is a tuning parameter and $\pi_{\boldsymbol{\alpha}}$ is the proportion parameter for the latent pattern $\boldsymbol{\alpha}$, can be used. Such a log-type penalty penalizes smaller proportions more heavily, and as recently shown in Gu and Xu (2019), can effectively select significant latent classes in the population. Alternatively, to perform such latent class selection, the use of Lasso or elastic-net type penalty can be explored in the future.

Another interesting problem is the uncertainty quantification of the latent pattern classification. Since in the proposed framework we directly assign the latent patterns by minimizing a loss function, the subjects' latent patterns are treated as fixed effects instead of random variables. Based on the clustering literature, it is theoretically challenging to quantify the uncertainty of clustering accuracy. One practical approach is to use bootstrap, where we resample the data multiple times and use the bootstrapped samples to quantify the estimation and classification uncertainty. It is also possible to further model latent pattern probabilities and use large deviation theory to approximate the misclassification errors. For instance, Liu et al. (2015) studied the asymptotic misclassification error rate for CDMs under the assumption that the item parameters are pre-calibrated. However, in the proposed framework, the item parameters and the latent patterns are unknown and jointly estimated, and we focus on a more complicated double asymptotic regime, where the sample size N and the number of items J both go to infinity, making uncertainty quantification even more challenging. This interesting problem will be explored further in the future.

One constraint of all the methods discussed in this paper pertain to the assumption that the Q-matrix is known and accurately specified. In practice, the Q-matrix may not be given or subjectively specified by domain experts, with possible misspecifications. There are some existing methods for estimating the Q-matrix in the literature (Chen, Culpepper, Chen, and Douglas, 2018; Chen, Liu, Xu, and Ying, 2015; Chung and Johnson, 2018; Culpepper, 2019; Liu, Xu, and Ying, 2012; Xu and Shang, 2018). Developing computational methods and theories for estimating CDMs with unknown Q-matrix under our proposed general framework is a natural next step that is left for future work. Another possible extension is to consider hierarchical structures among the latent attributes (Leighton, Gierl, and Hunka, 2004; Templin and Bradshaw, 2014; Ma and Xu, 2021), which may exclude some latent patterns in the subjects' population. Our proposed framework and computational algorithms should be easily adapted if the latent hierarchical structure is given. Our theoretical analysis will also be readily carried over to the hierarchical setting.

References

Celeux, G. and G. Govaert (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis 14*(3), 315–332.

Chen, Y., S. A. Culpepper, Y. Chen, and J. Douglas (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika 83*(1), 89–108.

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

Chiu, C.-Y. and J. Douglas (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification 30*(2), 225–250.

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

Chiu, C.-Y. and H.-F. Köhn (2019a). Consistency theory for the general nonparametric classification method. *Psychometrika 84*(3), 830–845.

Chiu, C.-Y. and H.-F. Köhn (2019b). Nonparametric methods in cognitively diagnostic assessment. *Handbook of Diagnostic Classification Models*, 107–132.

Chiu, C.-Y., H.-F. Köhn, Y. Zheng, and R. Henson (2016). Joint maximum likelihood estimation for diagnostic classification models. *Psychometrika 81*(4), 1069–1092.

Chiu, C.-Y., Y. Sun, and Y. Bian (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika 83*(2), 355–375.

Chung, M. and M. S. Johnson (2018). An MCMC algorithm for estimating the Q-matrix in a Bayesian framework. *arXiv preprint arXiv:1802.02286*.

Culpepper, S. (2019, 6). Estimating the cognitive diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika 84*(2), 333–357.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics 34*(1), 115–130.

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

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

DiBello, L., L. Roussos, and W. Stout (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. CR Rao, & S. Sinharay (eds.), handbook of statistics, vol. 26: Psychometrics (pp. 970–1030).

George, A. C. and A. Robitzsch (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology 11*(3), 189–205.

Gu, Y. and G. Xu (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research 20*(2019).

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement 26*(4), 301–321.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Ph. D. thesis, ProQuest Information & Learning.

Henson, R. A., J. L. Templin, and J. T. Willse (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika 74*(2), 191.

Junker, B. W. and K. Sijtsma (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement 25*(3), 258–272.

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

Liu, J., G. Xu, and Z. Ying (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement 36*(7), 548–564.

Liu, J., Z. Ying, and S. Zhang (2015). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika 80*(2), 468–490.

Ma, C. and G. Xu (2021). Hypothesis testing for hierarchical structures in cognitive diagnosis models. *Journal of Data Science*, 1–24, DOI 10.6339/21–JDS1024.

Popescu, P. G., S. S. Dragomir, E. I. Sluşanschi, and O. N. Stănăşilă (2016). Bounds for kullback-leibler divergence. *Electronic Journal of Differential Equations 2016*.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement 20*(4), 345–354.

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

Templin, J. L. and R. A. Henson (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods 11*(3), 287.

van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series 2005*(2), i–35.

Wang, S. and J. Douglas (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika 80*(1), 85–100.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics 45*(2), 675–707.

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

SUPPLEMENTAL MATERIAL

A Appendix

In the appendix, we provide detailed proofs of the Lemmas and Theorems in Section 4.

A.1 Proof of Theorem 1

Proof. Our proof uses similar arguments as in Chiu et al. (2016). First consider the case when the true membership $\boldsymbol{A}_c^0$ is known. Since $\hat{\mu}_{j,\boldsymbol{\alpha}} = \sum_{i \in C_{\boldsymbol{\alpha}}} x_{ij}/|C_{\boldsymbol{\alpha}}| := \bar{x}_{j,\boldsymbol{\alpha}}$, by Hoeffding's inequality (**?**), for any $\epsilon > 0$,

$$\begin{aligned} P\big(\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}}\|_\infty \geq \epsilon \;\big|\; \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\big) &= P\Big(\max_j |\bar{x}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}}| \geq \epsilon \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\Big) \\ &\leq \sum_{j=1}^{J} P\Big(|\bar{x}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}}| \geq \epsilon \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\Big) \\ &\leq \; 2J \exp\big(-2|C_{\boldsymbol{\alpha}}| \cdot \epsilon^2\big). \end{aligned}$$

Since $\lim_{n\to\infty} |C_{\boldsymbol{\alpha}}|/N_c \to \pi_{\boldsymbol{\alpha}}$ almost surely and $J \exp\big(-N_c\epsilon\big) \to 0$ for any $\epsilon > 0$, we have $J \exp\big(-2|C_{\boldsymbol{\alpha}}| \cdot \epsilon^2\big) = J \exp\Big(-2\big(1+o(1)\big)N_c \cdot \pi_{\boldsymbol{\alpha}} \cdot \epsilon^2\Big) \to 0$ almost surely.

Now consider the case when $\hat{\boldsymbol{A}}_c$ is consistent for $\boldsymbol{A}_c^0$, that is, $P(\hat{\boldsymbol{A}}_c \neq \boldsymbol{A}_c^0) \to 0$.

Then for any $\epsilon > 0$, we have

$$\begin{aligned} &P\big(\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}}\|_\infty \geq \epsilon\big) \\ &\leq P\big(\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}}\|_\infty \geq \epsilon \;\big|\; \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\big) \cdot P\big(\hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\big) + P\big(\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}}\|_\infty \geq \epsilon \;\big|\; \hat{\boldsymbol{A}}_c \neq \boldsymbol{A}_c^0\big) \cdot P\big(\hat{\boldsymbol{A}}_c \neq \boldsymbol{A}_c^0\big) \\ &\leq P\big(\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}}\|_\infty \geq \epsilon \;\big|\; \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\big) + P\big(\hat{\boldsymbol{A}}_c \neq \boldsymbol{A}_c^0\big) \\ &\xrightarrow{P} 0, \qquad \text{as } J \to \infty. \end{aligned}$$

Therefore we have $\|\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}}\|_\infty \xrightarrow{P} 0$. Since there are finitely many $\boldsymbol{\alpha}$'s, we have $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\theta}^0\|_\infty \xrightarrow{P} 0$ □

A.2 Proof of Lemma 1

Proof. Let $\tilde{\boldsymbol{\alpha}}_i$ denote the latent attribute pattern that minimizes $E[l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}) + h(\hat{\pi}_{\boldsymbol{\alpha}})]$, that is,

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}_i :=& \arg\min_{\boldsymbol{\alpha}} \big\{ E\big[l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}) + h(\hat{\pi}_{\boldsymbol{\alpha}}) \big] \big\} \\
=& \arg\min_{\boldsymbol{\alpha}} E\Big[\sum_{j=1}^{J} l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) + h(\hat{\pi}_{\boldsymbol{\alpha}}) \Big] \\
=& \arg\min_{\boldsymbol{\alpha}} \Big\{ \frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) \big] + \frac{1}{J} h(\hat{\pi}_{\boldsymbol{\alpha}}) \Big\}.
\end{aligned}$$

For the second term, under the Assumption 2, since $\hat{\pi}_{\boldsymbol{\alpha}}$ is asymptotically bounded and $h(\cdot)$ is continuous, hence $h(\hat{\pi}_{\boldsymbol{\alpha}})$ is also bounded, and we have $h(\hat{\pi}_{\boldsymbol{\alpha}})/J \to 0$ as $J \to \infty$, which is asymptotically negligible. For the first term, we need to compare $\frac{1}{J}\sum_{j=1}^{J} E\big[l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) \big]$ and $\frac{1}{J}\sum_{j=1}^{J} E\big[l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}_i^0}) \big]$ for any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}_i^0$.

$$\begin{aligned}
&\frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) \big] - \frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}_i^0}) \big] \\
=& \Big(\frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) \big] - \frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}}) \big] \Big) + \Big(\frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}}) \big] - \frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}_i^0}) \big] \Big) \\
&+ \Big(\frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}_i^0}) \big] - \frac{1}{J} \sum_{j=1}^{J} E\big[l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}_i^0}) \big] \Big) \\
:=& E_1 + E_2 + E_3. \qquad \text{(S1)}
\end{aligned}$$

Since $\hat{\boldsymbol{\mu}}$ is consistent for $\boldsymbol{\theta}^0$, by Assumption 1, we have $E_1 \xrightarrow{P} 0$ and $E_3 \xrightarrow{P} 0$. Specifically, first consider the case when $\hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0$. By Assumption 3, we know that the true item response probabilities are bounded. There exists $\delta_2 \in (0, 0.5)$ such that $\delta_2 \leq \min_{j,\boldsymbol{\alpha}} \theta^0_{j,\boldsymbol{\alpha}} < \max_{j,\boldsymbol{\alpha}} \theta^0_{j,\boldsymbol{\alpha}} \leq 1 - \delta_2, \forall 1 \leq j \leq J, \boldsymbol{\alpha} \in \{0,1\}^K$. Let's now look at the probability that $\hat{\mu}_{j,\boldsymbol{\alpha}}$ is also bounded. Specifically, we consider $P(\hat{\mu}_{j,\boldsymbol{\alpha}} \geq 1 - \delta_2/2 \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0)$ and $P(\hat{\mu}_{j,\boldsymbol{\alpha}} \leq \delta_2/2 \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0)$ respectively. Since $\hat{\mu}_{j,\boldsymbol{\alpha}} = \sum_{i \in C_{\boldsymbol{\alpha}}} x_{ij}/|C_{\boldsymbol{\alpha}}| := \bar{x}_{j,\boldsymbol{\alpha}}$, we have

$$\begin{aligned}
P(\hat{\mu}_{j,\boldsymbol{\alpha}} \geq 1 - \delta_2/2 \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0) =& \; P(\bar{x}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}} \geq 1 - \delta_2/2 - \theta^0_{j,\boldsymbol{\alpha}} \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0) \\
\leq& \; \exp\big(-2|C_{\boldsymbol{\alpha}}|(1 - \delta_2/2 - \theta^0_{j,\boldsymbol{\alpha}})^2 \big) \\
\leq& \; \exp\big(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2 \big).
\end{aligned}$$

Similarly, we also have $P(\hat{\mu}_{j,\boldsymbol{\alpha}} \le \delta_2/2 \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0) \le \exp\big(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2\big)$. Therefore,

$$P\big(\min_j \hat{\mu}_{j,\boldsymbol{\alpha}} \le \delta_2/2 \text{ or } \max_j \hat{\mu}_{j,\boldsymbol{\alpha}} \ge 1-\delta_2/2 \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\big) \le 2J\exp\big(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2\big).$$

Moreover, since under the Assumption 1, the loss function is assumed to be Hölder continuous, that is, there exist $c>0$ and $\beta>0$, such that for any $\mu_1,\mu_2 \in (\delta_2/2, 1-\delta_2/2)$, we have $|l(x,\mu_1)-l(x,\mu_2)| \le c|\mu_1-\mu_2|^\beta$ for $x=0$ or 1. Then

$$\begin{aligned}
|E_1| &= \Big|\frac{1}{J}\sum_{j=1}^J E\big[l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}})\big] - \frac{1}{J}\sum_{j=1}^J E\big[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big]\Big| \\
&\le \frac{1}{J}\sum_{j=1}^J E\Big[\big|l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}}) - l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big|\Big] \\
&\le \frac{1}{J}\sum_{j=1}^J E\big[c|\hat{\mu}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}}|^\beta\big] \\
&\le c\max_j\{E\big[|\hat{\mu}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}}|^\beta\big]\}
\end{aligned}$$

Therefore for any $\epsilon>0$,

$$\begin{aligned}
&P(|E_1|>\epsilon) \\
&\le P(|E_1|>\epsilon \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0) + P(\hat{\boldsymbol{A}}_c \ne \boldsymbol{A}_c^0) \\
&\le P(E_1|>\epsilon \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0, \delta_2/2 < \hat{\mu}_{j,\boldsymbol{\alpha}} < 1-\delta_2/2, j=1,\dots,J) \\
&\quad + P\big(\min_j \hat{\mu}_{j,\boldsymbol{\alpha}} \le \delta_2/2 \text{ or } \max_j \hat{\mu}_{j,\boldsymbol{\alpha}} \ge 1-\delta_2/2 \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\big) + P(\hat{\boldsymbol{A}}_c \ne \boldsymbol{A}_c^0) \\
&\le P(||\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}}||_\infty > (\epsilon/c)^{1/\beta}) + 2J\exp(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2) + P(\hat{\boldsymbol{A}}_c \ne \boldsymbol{A}_c^0) \\
&\le 2J\exp(-2|C_{\boldsymbol{\alpha}}|(\epsilon/c)^{2/\beta}) + 2J\exp(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2) + P(\hat{\boldsymbol{A}}_c \ne \boldsymbol{A}_c^0) \qquad \text{(S2)} \\
&= 2J\exp\Big(-2\big(1+o(1)\big)N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot(\epsilon/c)^{2/\beta}\Big) + 2J\exp\Big(-\big(1+o(1)\big)N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot\delta_2^2/2\Big) + P(\hat{\boldsymbol{A}}_c \ne \boldsymbol{A}_c^0) \\
&\quad \to 0,
\end{aligned}$$

where (S2) follows from Theorem 1. Similarly we can show that $E_3 \xrightarrow{P} 0$ as well.

For the second term, by Assumption 4, we have

$$E_2 = \frac{1}{J}\sum_{j=1}^J E\big[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big] - \frac{1}{J}\sum_{j=1}^J E\big[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}^0_i})\big] \ge \frac{1}{J}\sum_{j=1}^J |\theta^0_{j,\boldsymbol{\alpha}^0_i} - \theta^0_{j,\boldsymbol{\alpha}}|^\delta, \qquad \text{(S3)}$$

for any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}_i^0$. Since in Assumption 3, there exists $\delta_1 > 0$ such that $\lim_{J\to\infty} \min_{\boldsymbol{\alpha}\neq\boldsymbol{\alpha}'} ||\boldsymbol{\theta}^0_{\boldsymbol{\alpha}} - \boldsymbol{\theta}^0_{\boldsymbol{\alpha}'}||_1/J > \delta_1$, then for a small enough $c_0 > 0$, there exists $c_1 > 0$ such that $\big|\{j : |\theta^0_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}'}| \geq c_0\}\big| \geq c_1 J$ for any $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ and large enough J. That is, there should be as many items as of order J that can differentiate two different classes. Otherwise, $\big|\{j : |\theta^0_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}'}| \geq c_0\}\big|/J \to 0$, which contradicts with the Assumption 3 for a small enough c_0. Then in (S3), we have $E_2 \geq c_1 c_0^{\delta}$ as $J \to \infty$. Therefore, the true attribute pattern minimizes $E[l(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}; \hat{\pi}_{\boldsymbol{\alpha}})]$ with probability approaching 1. □

A.3 Proof of Lemma 2

Proof. We first decompose the probability in Lemma 2 into two parts:

$$P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}}) - E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})]\big)\Big| \geq \epsilon\Big) \tag{S4}$$

$$\leq P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}}) - l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big)\Big| \geq \epsilon/2\Big) + P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}}) - E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})]\big)\Big| \geq \epsilon/2\Big). \tag{S5}$$

The first term in (S5) goes to zero since $\hat{\boldsymbol{\theta}}$ is uniform consistent for $\boldsymbol{\theta}^0$. Specifically, from Lemma 1, we have $P(\hat{\mu}_{j,\boldsymbol{\alpha}} \leq \delta_2/2 \text{ or } \hat{\mu}_{j,\boldsymbol{\alpha}} \geq 1-\delta_2/2 \mid \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0) \leq 2\exp\big(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2\big)$. Moreover, due to the Hölder continuity of the loss function, we have $|l(x,\mu_1) - l(x,\mu_2)| \leq c|\mu_1-\mu_2|^{\beta}$ for $x = 0$ or 1. Then

$$\begin{aligned}
&P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}}) - l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big)\Big| \geq \epsilon/2 \;\Big|\; \delta_2/2 < \hat{\mu}_{j,\boldsymbol{\alpha}} \leq 1-\delta_2/2, \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\Big)\\
\leq\; &\sum_{\boldsymbol{\alpha}} P\Big(\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}}) - l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big)\Big| \geq \epsilon/2 \;\Big|\; \delta_2/2 < \hat{\mu}_{j,\boldsymbol{\alpha}} \leq 1-\delta_2/2,\ \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\Big)\\
\leq\; &2^K \sum_{j=1}^{J} P\Big(\Big|l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}}) - l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\Big| \geq \epsilon/2 \;\Big|\; \delta_2/2 < \hat{\mu}_{j,\boldsymbol{\alpha}} \leq 1-\delta_2/2, \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\Big)\\
\leq\; &2^K \sum_{j=1}^{J} P\Big(\Big|\hat{\mu}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}}\Big|^{\beta} \geq \epsilon/2c \;\Big|\; \delta_2/2 < \hat{\mu}_{j,\boldsymbol{\alpha}} \leq 1-\delta_2/2, \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\Big)\\
=\; &2^K \sum_{j=1}^{J} P\Big(\Big|\bar{x}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}}\Big| \geq (\epsilon/2c)^{1/\beta} \;\Big|\; \delta_2/2 < \hat{\mu}_{j,\boldsymbol{\alpha}} \leq 1-\delta_2/2, \hat{\boldsymbol{A}}_c = \boldsymbol{A}_c^0\Big)\\
\leq\; &2^{K+1} J \exp\big(-2|C_{\boldsymbol{\alpha}}|(\epsilon/2c)^{2/\beta}\big).
\end{aligned}$$

Then we have

$$
\begin{aligned}
&P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}})-l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big)\Big|\geq\epsilon/2\;\Big|\;\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0\Big)\\
&\leq\sum_{\boldsymbol{\alpha}}\sum_{j=1}^{J}\Big[P(\hat{\mu}_{j,\boldsymbol{\alpha}}<\delta_2/2\text{ or }\hat{\mu}_{j,\boldsymbol{\alpha}}>1-\delta_2/2\mid\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0)\\
&\qquad\qquad+P\big(\big|\hat{\mu}_{j,\boldsymbol{\alpha}}-\theta^0_{j,\boldsymbol{\alpha}}\big|\geq(\epsilon/2c)^{1/\beta}\;\big|\;\delta_2/2<\hat{\mu}_{j,\boldsymbol{\alpha}}\leq1-\delta_2/2,\;\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0\big)\Big]\\
&\leq 2^{K+1}J\exp(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2)+2^{K+1}J\exp(-2|C_{\boldsymbol{\alpha}}|(\epsilon/2c)^{2/\beta})\\
&=2^{K+1}J\exp\Big(-\big(1+o(1)\big)N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot\delta_2^2/2\Big)+2^{K+1}J\exp\Big(-2\big(1+o(1)\big)N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot(\epsilon/2c)^{2/\beta}\Big)\\
&\to 0,\text{ as }J\to\infty.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}})-l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big)\Big|\geq\epsilon/2\Big)\\
&\leq P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}})-l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big)\Big|\geq\epsilon/2\;\Big|\;\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0\Big)\cdot P(\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0)+P(\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0)\\
&\leq P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}})-l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big)\Big|\geq\epsilon/2\;\Big|\;\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0\Big)+P(\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0)\\
&\to 0,\text{ as }J\to\infty.
\end{aligned}
$$

Next we need to bound the second term. By Assumption 3, $\theta^0_{j,\boldsymbol{\alpha}}$'s are uniformly bounded and thus $l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})$'s are also uniformly bounded. There exists $M>0$ such that $\big|l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})\big|\leq M$ for any j and $\boldsymbol{\alpha}$. Then by Hoeffding's inequality (**?**), we have

$$
P\Big(\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})-E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})]\big)\Big|\geq\epsilon/2\Big)\leq 2\exp\big(-J\epsilon^2/2M^2\big),
$$

and therefore

$$
\begin{aligned}
&P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})-E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})]\big)\Big|\geq\epsilon/2\Big)\\
\leq\;&\sum_{\boldsymbol{\alpha}}P\Big(\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})-E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})]\big)\Big|\geq\epsilon/2\Big)
\end{aligned}
$$

$$\leq\ 2^{K+1}\exp\big(-J\epsilon^2/2M^2\big)\longrightarrow 0,\ \text{as}\ J\to\infty.$$

□

A.4 Proof of Theorem 2

Proof. Since $\hat{\boldsymbol{A}}_c$ is consistent for $\boldsymbol{A}_c^0$, by Theorem 1, $\hat{\boldsymbol{\mu}}$ is consistent for $\boldsymbol{\theta}^0$. Note that $\hat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i^0$ is equivalent to that

$$\frac{1}{J}\sum_{j=1}^{J} l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}_i^0}) + \frac{1}{J}h(\hat{\pi}_{\boldsymbol{\alpha}_i^0}) > \frac{1}{J}\sum_{j=1}^{J} l(x_{ij},\hat{\mu}_{j,\hat{\boldsymbol{\alpha}}_i}) + \frac{1}{J}h(\hat{\pi}_{\hat{\boldsymbol{\alpha}}_i}). \tag{S6}$$

From Assumptions 1 and 4 and the proof of Lemma 1, we know

$$\frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}_i^0})] < \frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij},\theta^0_{j,\hat{\boldsymbol{\alpha}}_i})] - c_1 c_0^{\delta} \tag{S7}$$

Let $c_2 = c_1 c_0^{\delta}$ and take $\epsilon = c_2/4$ in Lemma 2, and consider the event

$$B_\epsilon(J) := \big\{\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}}) - E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}})]\big)\Big| < \epsilon\big\}.$$

Since $h(\hat{\pi}_\alpha)$ is bounded, there exists some J_0 such that for any $J \geq J_0$, we have $\big|\frac{1}{J}h(\hat{\pi}_{\boldsymbol{\alpha}_i^0}) - \frac{1}{J}h(\hat{\pi}_{\hat{\boldsymbol{\alpha}}_i})\big| < c_2/4$. When $B_{c_2/4}(J)$ occurs, it implies that

$$\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\boldsymbol{\alpha}_i^0}) - E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}_i^0})]\big)\Big| < c_2/4,$$

and

$$\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij},\hat{\mu}_{j,\hat{\boldsymbol{\alpha}}_i}) - E[l(x_{ij},\theta^0_{j,\hat{\boldsymbol{\alpha}}_i})]\big)\Big| < c_2/4.$$

Then in equation (S6),

$$\text{LHS} < \frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij},\theta^0_{j,\boldsymbol{\alpha}_i^0})] + c_2/4 + \frac{1}{J}h(\hat{\pi}_{\boldsymbol{\alpha}_i^0}),$$

and

$$\text{RHS} > \frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij}, \theta^0_{j,\hat{\boldsymbol{\alpha}}_i})] - c_2/4 + \frac{1}{J}h(\hat{\pi}_{\hat{\boldsymbol{\alpha}}_i}),$$

which implies that

$$\begin{aligned}
\frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij}, \theta^0_{j,\hat{\boldsymbol{\alpha}}_i})] &< \frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}^0_i})] + c_2/2 + \frac{1}{J}h(\pi_{\boldsymbol{\alpha}^0_i}) - \frac{1}{J}h(\pi_{\hat{\boldsymbol{\alpha}}_i}) \\
&< \frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}^0_i})] + 3c_2/4 \\
&< \frac{1}{J}\sum_{j=1}^{J} E[l(x_{ij}, \theta^0_{j,\hat{\boldsymbol{\alpha}}_i})],
\end{aligned}$$

where the last inequality is from equation (S6) and results in a contradiction. It indicates that $\{\hat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}^0_i\} \subset B_{c_2/4}(J)^c$ for J large enough. And therefore we have

$$\begin{aligned}
P\Big(\hat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}^0_i\Big) \leq & P\big(B_{c_2/4}(J)^c\big) \\
\leq & P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\big(l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) - E[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}})]\big)\Big| \geq c_2/4\Big) \\
& \longrightarrow 0, \text{ as } J \to \infty. \quad \text{(by Lamma 2)}
\end{aligned}$$

□

A.5 Proof of Theorem 3

Proof. Following the proof of Theorem 2, we have

$$\begin{aligned}
& P\Big(\bigcup_i \{\hat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}^0_i\} \;\Big|\; \hat{\boldsymbol{A}}_c = \boldsymbol{A}^0_c\Big) \\
\leq & \sum_i P\Big(\{\hat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}^0_i\} \;\Big|\; \hat{\boldsymbol{A}}_c = \boldsymbol{A}^0_c\Big) \\
\leq & N \cdot P\Big(B_{c_2/4}(J)^c \;\Big|\; \hat{\boldsymbol{A}}_c = \boldsymbol{A}^0_c\Big) \\
\leq & N \cdot P\Big(\max_{\boldsymbol{\alpha}}\Big|\frac{1}{J}\sum_{j=1}^{J}\Big(l(x_{ij}, \hat{\mu}_{j,\boldsymbol{\alpha}}) - E\big[l(x_{ij}, \theta^0_{j,\boldsymbol{\alpha}})\big]\Big)\Big| \geq c_2/4 \;\Big|\; \hat{\boldsymbol{A}}_c = \boldsymbol{A}^0_c\Big) \\
\leq & 2^{K+1}NJ\exp(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2) + 2^{K+1}NJ\exp(-2|C_{\boldsymbol{\alpha}}|(c_2/8c)^{2/\beta}) + 2^{K+1}N\exp\big(-Jc_2^2/32M^2\big)
\end{aligned}$$

$$\leq 2^{K+1}N^2\exp(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2) + 2^{K+1}N^2\exp(-2|C_{\boldsymbol{\alpha}}|(c_2/8c)^{2/\beta}) + 2^{K+1}N\exp\big(-Jc_2^2/32M^2\big).$$

Under the Assumption 2, we have $\lim_{n\to\infty}|C_{\boldsymbol{\alpha}}|/N_c \to \pi_{\boldsymbol{\alpha}}$ almost surely; therefore $N^2\exp(-|C_{\boldsymbol{\alpha}}|\delta_2^2/2) = N^2\exp\Big(-\big(1+o(1)\big)N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot\delta_2^2/2\Big)$ and $N^2\exp(-2|C_{\boldsymbol{\alpha}}|(c_2/8c)^{2/\beta}) = N^2\exp\Big(-2\big(1+o(1)\big)N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot(c_2/8c)^{2/\beta}\Big)$. Then we have

$$\begin{aligned}
&P\Big(\bigcup_i\{\hat{\boldsymbol{\alpha}}_i\neq\boldsymbol{\alpha}_i^0\}\Big)\\
&\leq P\Big(\bigcup_i\{\hat{\boldsymbol{\alpha}}_i\neq\boldsymbol{\alpha}_i^0\}\;\Big|\;\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0\Big)P\Big(\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0\Big) + P\Big(\bigcup_i\{\hat{\boldsymbol{\alpha}}_i\neq\boldsymbol{\alpha}_i^0\}\;\Big|\;\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0\Big)P\Big(\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0\Big)\\
&\leq P\Big(\bigcup_i\{\hat{\boldsymbol{\alpha}}_i\neq\boldsymbol{\alpha}_i^0\}\;\Big|\;\hat{\boldsymbol{A}}_c=\boldsymbol{A}_c^0\Big) + P\Big(\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0\Big)\\
&\leq 2^{K+1}N^2\exp\Big(-\big(1+o(1)\big)N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot\delta_2^2/2\Big) + 2^{K+1}N^2\exp\Big(-\big(1+o(1)\big)2N_c\cdot\pi_{\boldsymbol{\alpha}}\cdot(c_2/8c)^{2/\beta}\Big)\\
&\quad + 2^{K+1}N\exp\big(-Jc_2^2/32M^2\big) + P\Big(\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0\Big)\\
&\leq 2^{K+1}N^2\exp\Big(-\big(1+o(1)\big)J\cdot\pi_{\boldsymbol{\alpha}}\cdot\delta_2^2/2\Big) + 2^{K+1}N^2\exp\Big(-2\big(1+o(1)\big)J\cdot\pi_{\boldsymbol{\alpha}}\cdot(c_2/8c)^{2/\beta}\Big)\\
&\quad + 2^{K+1}N\exp\big(-Jc_2^2/32M^2\big) + P\Big(\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0\Big)\\
&= 2^{K+1}\Big[N\exp\Big(-\big(1+o(1)\big)J\cdot\pi_{\boldsymbol{\alpha}}\cdot\delta_2^2/4\Big)\Big]^2 + 2^{K+1}\Big[N\exp\Big(-\big(1+o(1)\big)J\cdot\pi_{\boldsymbol{\alpha}}\cdot(c_2/8c)^{2/\beta}\Big)\Big]^2\\
&\quad + 2^{K+1}N\exp\big(-Jc_2^2/32M^2\big) + P\Big(\hat{\boldsymbol{A}}_c\neq\boldsymbol{A}_c^0\Big)\\
&\to 0,\ \text{as } J\to\infty.
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\alpha}}_i$'s are uniformly consistent for $\boldsymbol{\alpha}_i$'s for all $i=1,\ldots,N$. □

A.6 Proof of Proposition 1

Proof. Our proof uses similar arguments as in Celeux and Govaert (1992). In Step 3 of Algorithm 1, we have

$$L(\boldsymbol{A}^{(t)},\boldsymbol{\mu}^{(t+1)},\boldsymbol{\pi}^{(t+1)}) \leq L(\boldsymbol{A}^{(t)},\boldsymbol{\mu}^{(t)},\boldsymbol{\pi}^{(t)}).$$

Moreover, since $\hat{\boldsymbol{\alpha}}_i^{(t+1)} = \arg\min_{\boldsymbol{\alpha}} l(\boldsymbol{x}_i,\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}^{(t+1)}) + h(\hat{\pi}_{\boldsymbol{\alpha}}^{(t+1)})$, which is equivalent to that $l\big(\boldsymbol{x}_i,\hat{\boldsymbol{\mu}}_{\hat{\boldsymbol{\alpha}}_i^{(t+1)}}^{(t+1)}\big) + h\big(\hat{\pi}_{\hat{\boldsymbol{\alpha}}_i^{(t+1)}}^{(t+1)}\big) \leq l(\boldsymbol{x}_i,\hat{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}^{(t+1)}) + h(\hat{\pi}_{\boldsymbol{\alpha}}^{(t+1)})$ for any $\boldsymbol{\alpha}\neq\hat{\boldsymbol{\alpha}}_i^{(t+1)}$, we have

$$L(\boldsymbol{A}^{(t+1)},\boldsymbol{\mu}^{(t+1)},\boldsymbol{\pi}^{(t+1)}) \leq L(\boldsymbol{A}^{(t)},\boldsymbol{\mu}^{(t)},\boldsymbol{\pi}^{(t)}). \tag{S8}$$

Therefore the criterion (9) is decreasing.

In the finite sample setting, since there is finite number of partitions into 2^K classes, the decreasing sequence $L(\boldsymbol{A}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}^{(t)})$ also takes a finite number of values, which makes it converge to a stationary value. Moreover, since the minima of the loss function is well-defined, the sequence $(\boldsymbol{A}^{(t)}, \boldsymbol{\mu}^{(t)} f, \boldsymbol{\pi}^{(t)})$ also converges. □

A.7 Proof of Proposition 2

Proof. Our proof directly follows that in Celeux and Govaert (1992). Since

$$\begin{aligned} L(\boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\pi}) &= \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i=1}^{n} u_{i\boldsymbol{\alpha}} \Big(l(\boldsymbol{x_i}, \boldsymbol{\mu_\alpha}) + h(\pi_{\boldsymbol{\alpha}}) \Big) \\ &\geq \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i=1}^{n} u_{i\boldsymbol{\alpha}} \min_{\boldsymbol{\alpha'}} \Big(l(\boldsymbol{x_i}, \boldsymbol{\mu_{\alpha'}}) + h(\pi_{\boldsymbol{\alpha'}}) \Big) \\ &\geq \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \sum_{i=1}^{n} \min_{\boldsymbol{\alpha'}} \Big(l(\boldsymbol{x_i}, \boldsymbol{\mu_{\alpha'}}) + h(\pi_{\boldsymbol{\alpha'}}) \Big), \end{aligned}$$

where the RHS is attained when $\boldsymbol{U}$ is equivalent to some partition, the Algorithm 1 can be regarded as an alternating optimization algorithm to minimize $L(\boldsymbol{U}, \boldsymbol{\mu}, \boldsymbol{\pi})$. Specifically, the Algorithm 1 is in fact a grouped coordinate descent method. Following the Theorem 2.2 of **?**, the Proposition 2 is proved. □