Statistical Analysis of Structured Latent Attribute Models

by

Yuqi Gu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2020

Doctoral Committee:

Assistant Professor Gongjun Xu, Chair
Associate Professor Long Nguyen
Professor Naisyin Wang
Assistant Professor Zhenke Wu

Pink tulip blossom in the Diag on Central Campus in Ann Arbor, May, 2019. In memory of those bright and colorful days that Ann Arbor has generously offered me.

Yuqi Gu

yuqigu@umich.edu

ORCID iD: 0000-0002-4124-113X

To my beloved family:

my parents Xiaoguang Gu and Xiaoxu Chen, and my husband Xiaowei Wang.

ACKNOWLEDGEMENTS

It was during a flight from Detroit to Seattle in January 2020 that I started thinking about how to write this acknowledgments section. The airplane flew above the clouds, and the warm golden sunshine was all over there, although the weather beneath was rather gloomy. It reminded me of the days with ups and downs that I have experienced in the past five years in Michigan. Thinking of these suddenly filled my heart with a mix of excitement and sentimental thoughts. At that time, I really wished I could use words to describe even a tiny bit of how I am incredibly grateful to numerous people who made this magical journey possible, though I knew that words are deemed to be pale compared to what I feel deep inside.

First and foremost, I would like to express my deepest gratitude to my advisor Gongjun Xu, without whom I could not possibly have completed this dissertation nor have grown so much into the person that I am today. Gongjun is a remarkable researcher who is always sharp-minded and hardworking, an incredible mentor who is always supportive and encouraging, and an academic role model of mine that I know will continue to inspire me for many years to come. Gongjun is the best advisor one could ever hope for. From day one I started to work with him, I learned a special rule that his door would keep open whenever I have questions or have updates to share. His enormous availability to discuss research and offer insightful suggestions are really phenomenal. Working with Gongjun made it possible for me to touch upon and push forward my own limit. I especially thank Gongjun for introducing me and guiding me to work on the identifiability problems. These problems are both challenging and beautiful. Doing research on these and being able to make progress on them brought me great intellectual excitement and also self-confidence. Discovering the interesting structures of those necessary and sufficient identifiability conditions through rigorous proofs were undoubtedly among the happiest moments in my graduate school years. I am forever indebted to Gongjun for devoting so much time and energy to mentoring me and for bringing the best part of me. Gongjun sets an exemplary model and makes me also long for a career like his – doing great research and nurturing the next generation of researchers.

I am also very grateful to my doctoral dissertation committee members, Long Nguyen, Naisyin Wang, and Zhenke Wu, not only for their valuable time and effort in serving on my committee but also for their strong support to me throughout my graduate studies and in my job search. I took STATS 601, a Ph.D. level statistical machine learning course with Long in my first year in graduate school. Long's carefully designed lectures really showed how knowledgeable and passionate a scholar he is. This exactly triggered my interests in those elegant modern machine learning ideas and methods, including graphical models, latent variable models, and topic models; since then, I have always wanted to learn more about them. I got to know Naisyin when I served as a Graduate Student Instructor (GSI) for an undergraduate course that Naisyin was teaching. Our interactions in that semester let me understand how deeply Naisyin cares about the students and the GSI. Later through a reading group with her, I furthered learned how Naisyin is enthusiastically pursuing cutting edge research in privacy and has an incisive research taste. Naisyin's warm encouragement and support, especially those from a female researcher's perspective, meant a lot to me. I hope to continue learning from her and to give back to others like what she has done to me. Zhenke is my collaborator on the project about modeling structured data in biomedical research. Zhenke's knowledge in that field, willingness and ability in collaboration, and helpful advice on writing and presenting research was beneficial to me. I also would like to sincerely thank Xuming He for his helpful advice and valuable support.

During my job search process for a tenure-track position in the final year of my Ph.D. study, I also benefited from interacting with various other people. I am grateful to Liza Levina for her detailed and helpful comments on my job search documents and my practice job talk. I thank Jingshen Wang for encouragement and confidence in me. I thank Xu Shi and Kean Ming Tan for attending my practice talk and providing many helpful suggestions that greatly improved it. I sincerely thank all the faculty members in the Department of Statistics at the University of Michigan for their guidance and support during various stages in my graduate study. I heartily thank my wonderful friends in Ann Arbor for their friendship throughout these years.

I thank my parents for their unconditional love and support. My parents always strived to give me the best education. I lived with them until I went to Beijing to attend college. Even today, I clearly remember how I felt seeing my parents waving goodbye to me and then walking down the tree-lined road on the Tsinghua Campus in Summer 2011. During the nine years since then, I have not stayed with them for more than two weeks at one time. But what I know for sure is wherever I go, they will happily see me chase my dreams and be proud of me. Knowing this gives me constant courage in front of whatever challenges I will likely face in the future.

My special thanks go to my husband Xiaowei Wang, who always stays by my side and takes good care of me. I thank Xiaowei for sharing all my ups and downs, my laughter and tears, my success and failures. I thank Xiaowei for unwaveringly loving, hugging, and supporting me regardless of what I achieve but just because of who I am. Thinking of those precious moments between us, I am now so much looking forward to our future, where I know many more amazing life chapters will gradually unfold in front of us. We will go through them hand in hand.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES xi

LIST OF APPENDICES xii

ABSTRACT xiii

CHAPTER

I. Introduction 1

1.1 Setup of Structured Latent Attribute Models (SLAMs) 3
1.2 Model Examples: in Cognitive Diagnostic Modeling and in Machine Learning 8
1.3 Real Data Examples: in Designing Practice 13
1.4 Unique Challenges of SLAMs and Our Contributions 15

II. Necessary and Sufficient Condition for the Identifiability of the DINA Model 22

2.1 Main Theorem of Necessity and Sufficiency 23
2.2 Discussion 31

III. Partial Identifiability of Structured Latent Attribute Models 33

3.1 Issues with Existing Works and Open Problems 36
3.2 Identifiability Results for Two-Parameter Models 39
3.2.1 Strict and Partial Identifiability 43
3.2.2 Results for Q-restricted Latent Class Models 46
3.2.3 Applications 50

3.3 Identifiability Results for Multiparameter Models 52
3.3.1 Strict and Generic Identifiability 55
3.3.2 Results for Q-restricted Latent Class Models 58
3.3.3 Applications 60
3.4 Extensions to More Complex Models 61
3.4.1 Mixed-items Restricted Latent Class Models 61
3.4.2 Restricted Latent Class Models with Categorical Responses 63
3.4.3 Deep Restricted Boltzmann Machines 64
3.5 Discussion 66

IV. Necessary and Sufficient Conditions for the Identifiability of the Q-matrix 69

4.1 Definitions and Illustrations 72
4.1.1 Example of the Generic Identifiability Phenomenon with $Q_{4\times 2}$ 73
4.2 Identifiability of $(Q, \mathbf{\Theta}, \boldsymbol{p})$ under the DINA Model 78
4.3 Identifiability of $(Q, \mathbf{\Theta}, \boldsymbol{p})$ under General SLAMs 84
4.4 Discussion 88

V. Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models 92

5.1 Motivation for Latent Pattern Selection 94
5.2 Learnability Requirement and Conditions 97
5.3 Penalized Likelihood Approach to Pattern Selection 108
5.3.1 Shrinkage Estimation 108
5.3.2 Screening as a Preprocessing Step When $2^K \gg N$. 121
5.4 Simulation Studies 124
5.5 Data Analysis 134
5.6 Discussion 139

VI. Identification and Estimation of Hierarchical Latent Attribute Models 141

6.1 Hierarchical Latent Attribute Models 143
6.2 Joint Identifiability of Q-matrix and Attribute Hierarchy . . . 148
6.3 A Scalable Algorithm for Estimating HLAMs 151
6.4 Simulations and Real Data Analysis 154
6.5 Discussion 158

APPENDICES 161

BIBLIOGRAPHY 383

LIST OF FIGURES

Figure

1.1	Visualizing Q-matrix as a bipartite graph	5
1.2	Deep Boltzmann Machine	13
2.1	illustrating necessary and sufficient conditions on the Q-matrix in an example	26
3.1	(Deep) Restricted Boltzmann machine	65
3.2	Flowchart of the results in Sections 3.2 and 3.3	67
4.1	Illustration of nonidentifiability under $Q_{4\times 2}$ in scenario (a).	75
4.2	Illustration of generic identifiability under $Q_{4\times 2}$, which corresponds to simulation scenario (b).	75
4.3	The effect of the generic identifiability constraint (4.3). Red "$*$"s represent parameter sets with the 20% largest MSEs in Figure 4.2(a), with $N = 10^5$; blue "+"s represent the remaining parameter sets.	77
4.4	geometry of generic identifiability with $Q_{4\times 2} = (I_2; I_2)$.	77
5.5	Screening under noise rate 10%: plots (a), (c) are coverage probabilities of the true patterns, from the screening procedure under the two-parameter SLAM; plots (b), (d) are sizes of $\widehat{\mathcal{A}}_{\text{screen}}$. The "noise rate" refers to the value of $1 - \theta_j^+ = \theta_j^-$.	130
5.6	Screening under noise rate 20%: plots (a), (c) are coverage probabilities of the true patterns, from the screening procedure under the two-parameter SLAM; plots (b), (d) are sizes of $\widehat{\mathcal{A}}_{\text{screen}}$. The "noise rate" refers to the value of $1 - \theta_j^+ = \theta_j^-$.	131
5.7	Screening enhanced by Gibbs exploration: screening accuracy and size of $\widehat{\mathcal{A}}_{\text{screen}}$. Noise rate is $1 - \theta_j^+ = \theta_j^- = 20\%$.	132
5.9	Fraction Subtraction Data: different sets of estimated patterns (a)–(d) (black for "0" and white for "1") and the corresponding attribute structures (e)–(h) under various Υ's in Algorithm 2. Plot (h) here is equivalent to Figure 5.8(b).	135
6.1	Different attribute hierarchies among binary attributes, the first three for $K = 4$ (where $\lvert\{0,1\}^4\rvert = 16$) and the last for $K = 7$ (where $\lvert\{0,1\}^7\rvert = 128$). E.g., the set of allowed attribute patterns under hierarchy (a) is $\mathcal{A}_1 = \{(0000), (1000), (1100), (1001), (1101), (1111)\}$.	145

6.2 A binary structural Q-matrix and the corresponding graphical model with directed edges from the latent to the observed variables representing dependencies. Below the observed variables in (b) are the row vectors of $Q_{6\times 3}$, i.e., the item loading vectors. There is $\mathcal{E} = \{1 \to 2,\ 1 \to 3\}$. 145
6.3 Among $K = 8$ attributes, on the upper-left is a diamond shape hierarchy, resulting in 15 patterns; and on the bottom-left is a tree shape hierarchy resulting in 10 patterns. The upper-right and bottom-right plots show how many times out of the 200 runs each true prerequisite relation is successfully recovered. The setting is $(N, J) = (1200, 1200)$ and $1 - \theta_j^+ = \theta_j^- = 20\%$. 157
6.4 Results of Austrian TIMSS 2011 data. 160
C.1 DINA: exhaustive search in the set of 5×2 Q-matrices with a true Q-matrix satisfying Conditions A, B and C in Theorem 1. 298
C.2 DINA: exhaustive search in 5×2 Q-matrices with a true Q-matrix Q^5 in (C.45) generically identifiable, corresponding to scenario (b.2) in Theorem 2. 300
C.3 DINA: exhaustive search in 5×2 Q-matrices with a true Q-matrix *not even locally* identifiable, corresponding to scenario (b.1) in Theorem 2. 301
C.4 DINA: true Q-matrix of size 20×3 is not complete and hence not identifiable. 306
C.5 DINA: true Q-matrix of size 20×5 is not complete and hence not identifiable. 307
C.6 GDINA: exhaustive search in 5×2 Q-matrices with a true Q satisfying Conditions D and E. 309
C.7 GDINA: exhaustive search in 5×2 Q-matrices with a true Q-matrix which leads to a not even locally generically identifiable model. . . . 312
C.8 GDINA: true Q is Q_3 with $(K, J) = (3, 20)$; each of the 70 colored line corresponds to one set of alternative parameters under $\bar{Q}_3$; all sets non-distinguishable. 316
C.9 GDINA: true Q is Q_4 with $(K, J) = (5, 20)$; each of the 70 colored line corresponds to one set of alternative parameters under $\bar{Q}_4$; all sets non-distinguishable. 318
D.1 Performance of Algorithm 1 across various values for threshold c. Setting is $K = 10$ and $1 - \theta_j^+ = \theta_j^- = 0.2$. In each scenario 200 runs are carried out, and the error bar is within one standard deviation of the mean accuracy. 344
D.2 Root Mean Square Errors (RMSEs) for estimating the true proportions of patterns decrease as sample size N increases. Results are based on 200 runs for each N. 346
D.3 Coverage probabilities of the true patterns, from the approximate screening procedure under the multi-parameter SLAM. Boxplots are from 200 runs in each scenario. 346

D.4 Sizes of the finally selected patterns $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$ under the two-parameter SLAM. The "noise" refers to the value of $1-\theta_j^+ = \theta_j^-$. The number of true patterns is $|\mathcal{A}_0| = 10$. 347
D.5 Different attribute structures corresponding to various Υ's in Algorithm 2. Plot (d) here is equivalent to Figure 5.10(b), the attribute structure selected by EBIC. 348
E.1 Boxplots of entry-differences between estimated Q_{iter} (or Q_{ini}) and the true structural matrix Q_{true}, with size 1000×7. Results are based on 200 independent runs. 379
E.2 Noisy binary matrix factorization and reconstruction with *randomly perturbed initialization*. Color yellow represents value "1" and color blue represents value "0". Only 3 stochastic EM iterations suffice for perfect estimation of the structural matrix Q. 380
E.3 Noisy binary matrix factorization and reconstruction with *entirely random initialization*. Color yellow represents value "1" and color blue represents value "0". Only 4 stochastic EM iterations of the proposed ADG-EM Algorithm 4 suffice for almost perfect decomposition and reconstruction. The stochastic Q after 4 iterations is identical to the true Q after column permutation. 381

LIST OF TABLES

Table

1.1 Q-matrix entry frequencies, TOEFL iBT field test, Reading Forms A & B . . . 14

1.2 Q-matrix, TIMSS 2003 8th Grade Data . . . 20

1.3 Q-matrix, Fraction Data . . . 21

5.1 Q-matrix in Su et al. (2013) for TIMSS 2003 8th Grade Data . . . 95

5.2 Pattern selection accuracies for two-parameter SLAM. Tuning parameter $\lambda \in \{-0.2, -0.4, \ldots, -3.8, -4.0\}$ in PEM (Algorithm 1) and $\Upsilon \in \{1.0, 0.9, \ldots, 0.4, 0.3\}$ in FP-VEM (Algorithm 2) are selected based on EBIC. . . . 126

5.3 Pattern selection accuracies for multi-parameter all-effect SLAM. Tuning parameter $\lambda \in \{-0.2, -0.4, \ldots, -4.0\}$ in PEM (Algorithm 1) and $\Upsilon \in \{1.0, 0.9, \ldots, 0.3\}$ in FP-VEM (Algorithm 2) are selected using EBIC. Signal strengths are $\theta_{j,\mathbf{0}_K} = 0.1$, $\theta_{j,\mathbf{1}_K} = 0.9$. . . . 133

6.1 Accuracy of learning the structural Q-matrix and the attribute hierarchy. The "noise" in the table refers to the value of $1 - \theta_j^+ = \theta_j^-$. Numbers in the column "size" record the *median values* of the cardinality of $|\mathcal{A}_{\text{final}}|$ (and $|\mathcal{A}_{\text{candi}}|$ in the parenthesis), based on 200 runs in each scenario. . . . 157

E.1 Statistical variations of results presented in Table 6.1 in the main text. The "IQR" stands for the inter quartile range of the three accuracy measures: $\text{acc}[Q]^{\mathcal{A}}$, TPR, and $1-$FDR and that of the size of $\mathcal{A}_{\text{final}}$. All results are based on 200 runs. In each scenario, the IQR is presented below the original number of the accuracy or the size of $\mathcal{A}_{\text{final}}$ in parenthesis. . . . 375

LIST OF APPENDICES

Appendix

A. Appendix of Chapter II 162

B. Appendix of Chapter III 181

C. Appendix of Chapter IV 254

D. Appendix of Chapter V 319

E. Appendix of Chapter VI 349

ABSTRACT

In modern psychological and biomedical research with diagnostic purposes, scientists often formulate the key task as inferring the fine-grained latent information under structural constraints. These structural constraints usually come from the domain experts' prior knowledge or insight. The emerging family of Structured Latent Attribute Models (SLAMs) accommodate these modeling needs and have received substantial attention in psychology, education, and epidemiology. SLAMs bring exciting opportunities and unique challenges. In particular, with high-dimensional discrete latent attributes and structural constraints encoded by a structural matrix, one needs to balance the gain in the model's explanatory power and interpretability, against the difficulty of understanding and handling the complex model structure.

This dissertation studies such a family of structured latent attribute models from theoretical, methodological, and computational perspectives. On the theoretical front, we present identifiability results that advance the theoretical knowledge of how the structural matrix influences the estimability of SLAMs. The new identifiability conditions guide real-world practices of designing diagnostic tests and also lay the foundation for drawing valid statistical conclusions. On the methodology side, we propose a statistically consistent penalized likelihood approach to selecting significant latent patterns in the population in high dimensions. Computationally, we develop scalable algorithms to simultaneously recover both the structural matrix and the dependence structure of the latent attributes in ultrahigh dimensional scenarios. These developments explore an exponentially large model space involving many dis-

crete latent variables, and they address the estimation and computation challenges of high-dimensional SLAMs arising from large-scale scientific measurements. The application of the proposed methodology to the data from international educational assessments reveals meaningful knowledge structures of the student population.

CHAPTER I

Introduction

In the era of data science, latent variable models have witnessed a tremendous surge of interest from a wide range of scientific applications and machine learning problems. On one hand, latent variable models have always played an important role in social and behavioral sciences to model constructs that are not directly measurable, such as extrovert personality or public opinion. On the other hand, latent variables are useful tools for dimension reduction in machine learning, and they hold huge representational and predictive power in deep neural networks.

The entire family of latent variable models can be categorized into four general types according to the nature of the observed and the latent variables. With the observed and latent variables both being continuous, the traditional factor analysis and probabilistic principal component analysis (Anderson and Rubin, 1956; Tipping and Bishop, 1999) can be used in modeling. To model continuous observed data using discrete latent variables, researchers have employed mixture models of continuous distributions, such as the Gaussian mixture model (Reynolds et al., 2000; Bishop, 2006), for explaining data heterogeneity and clustering subjects. When it comes to discrete observations, the item response theory models (Weiss and Yoes, 1991; Embretson and Reise, 2013) has been traditionally used in the field of psychometrics to draw continuous latent information from categorical data. Finally, when both the observed variables and the latent constructs of interest are discrete, the latent class model has been a popular modeling tool since decades ago (Lazarsfeld, 1959; Goodman, 1974).

In particular, in many areas of modern social and biomedical research, the key task can be formulated as inferring the fine-grained latent information from noisy measurements. Especially in many applications, formalizing the latent constructs as being discrete, instead of being continuous, allow for more interpretability and also naturally enables subsequent clustering of subjects. Further, in many real-world problems it is critical to incorporate structural information into the latent variable modeling process. Such discrete latent variable models with structural constraints have received a lot of attention in various fields, including psychology, epidemiology, and medicine. We term such models as *Structured Latent Attribute Models* (SLAMs), which generally fall into the last category of using discrete latent variables to explain discrete outcomes as described in the previous paragraph. However, SLAMs have the following key features distinct from the traditional latent class model: the first is that in a SLAM the latent variable per subject is characterized by a configuration of multiple fine-grained attributes; and the second is that the aforementioned structural constraints play an important role in describing the data generation process. Therefore, SLAMs can also be viewed as restricted latent class models (Xu, 2017; Gu and Xu, 2020a). These key features pose many interesting and challenging questions, requiring balancing the additional gain in the model's explanatory power and scientific interpretability, against the additional difficulty of understanding and handling the complex model structure.

This dissertation studies such a modern family of structured latent attribute models from theoretical, methodological, and computational perspectives. In the remaining part of this chapter, we first introduce the setup of SLAMs in Section 1.1. Then we review some popular model examples in Section 1.2 and some real-world designs in Section 1.3. Later in Section 1.4, we point out the unique challenges brought by SLAMs, summarize our contributions, and outline the structure of this dissertation.

1.1 Setup of Structured Latent Attribute Models

SLAMs offer a framework to achieve fine-grained inference on individuals' multiple latent attributes. This further provides the basis for clustering the population into subgroups based on the inferred attribute patterns. These models are central to a wide scope of applications, including the following examples.

(1) *Cognitive diagnosis in educational assessment.* Structured latent attribute models play a key role in cognitive diagnosis modeling in educational and psychological assessment. Cognitive diagnosis aims to make a classification-based decision on an individual's latent attributes, based on his or her observed responses to a set of designed diagnostic items (questions). The structural constraints usually come from the design matrix that specifies what latent attributes each item measures (e.g., Junker and Sijtsma, 2001; Henson et al., 2009; Rupp et al., 2010; de la Torre, 2011). See Section 1.3 for several data examples, including the Test of English as a Foreign Language (TOEFL) (e.g., von Davier, 2008) and Trends in International Mathematics and Science Study.

(2) *Psychiatric evaluation in clinical settings.* Structured latent attribute models have also been used in psychiatric evaluation. Here the responses are manifested symptoms and the latent patterns represent the profiles of presence/absence of a set of underlying psychological or psychiatric disorders. The structural constraints result from the fact that each symptom may be shared by multiple disorders, which are specified by psychiatric diagnosis guidelines. See examples in Templin and Henson (2006), Jaeger et al. (2006), and de la Torre et al. (2018).

(3) *Disease etiology detection in epidemiology.* Another application of structured

latent attribute models is the diagnosis of disease etiology in epidemiology (Wu et al., 2016, 2017; Deloria Knoll et al., 2017; O'Brien et al., 2019). Here the observed responses are imperfect laboratory measurements of subjects' biological samples, and the latent attribute patterns are the configurations of existence or absence of a set of pathogens underlying some disease. The structural constraints naturally arise from the fact that each measurement may only target certain pathogens.

In these applications, either the study design or the prior knowledge dictates that the observed variables depend on the latent ones in a highly structured fashion. For example, each test item in an educational assessment, by design, may only measure a particular subset of the skills, while in disease etiology research each laboratory measurement may target a specific set of pathogens. SLAMs incorporate these scientifically interpretable constraints through a key structure: a Q-matrix of binary entries. In a scenario with J observed measurements per subject that target K unobserved latent attributes, the Q-matrix has size $J \times K$. The concept of the Q-matrix was first proposed in Tatsuoka (1983) and later gained popularity in many *cognitive diagnostic models*, as will be reviewed in Section 1.2. Figure 1.1 illustrates the bipartite graph representation of a Q-matrix. The directed edges from the K latent attributes (in circles) to the J observed responses (in rectangles) represent the structured statistical dependence; these directed edges can be equivalently expressed as nonzero entries in a $J \times K$ binary matrix $Q = (q_{j,k})_{J\times K}$. On the latent side, arbitrary dependencies are allowed among the attributes, as indicated by the dotted edges in Figure 1.1. When $q_{j,k} = 1$, there exists statistical dependence of outcome j on latent attribute k, there is a directed edge from latent attribute k to observed item response j. We say attribute k is a parent attribute of item j if $q_{j,k} = 1$. Further, denote the the set of parent attributes of each item j by $\mathcal{K}_{\boldsymbol{q}_j} = \{k \in \{1,\dots,K\} : q_{j,k} = 1\}$.

$$Q = \begin{pmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \boldsymbol{q}_3 \\ \boldsymbol{q}_4 \\ \boldsymbol{q}_5 \\ \boldsymbol{q}_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

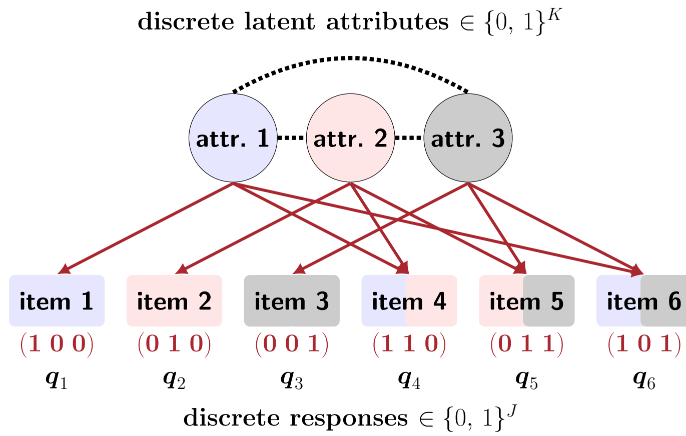


Figure 1.1: Visualizing Q-matrix as a bipartite graph

The most important statistical property of the family of SLAMs is characterized by the Q-matrix, or equivalently, by the bipartite graph between the latent attributes and the observed responses. Specifically, the following key property is shared by all the SLAMs considered in this dissertation.

Property 1. The distribution of the observed response R_j to the jth item depends *only on* the parent attributes of item j (that is, those in $\mathcal{K}_{\boldsymbol{q}_j}$), as specified by the entries of the Q-matrix.

In most real-world applications of SLAMs in psychological and educational measurement, the Q-matrix is pre-specified by practitioners and summarizes the information of the study design. This process is subjective and misspecification might exist, therefore in practice, sometimes researchers are interested in the identification and estimation of the Q-matrix itself. This dissertation will investigate both scenarios: both with a known Q-matrix and with an unknown Q-matrix.

SLAMs have close connections with many other statistical models. First, each possible configuration of K attributes forms a pattern defining a latent subpopulation. Therefore the model can be viewed as a structured *mixture model* (McLachlan and Peel, 2004) and also provides a framework for *model-based clustering* (Fraley and Raftery, 2002) of categorical data. Second, the probability distribution of a SLAM can be written as a mixture of higher-order tensors, relating the framework to *tensor decompositions* (Anandkumar et al., 2014). Third, SLAMs also connect with the mixed membership model for multivariate categorical data Erosheva et al. (2007) through a reformulation. Fourth, SLAMs share a similar spirit to the restricted/deep Boltzmann machines and deep belief networks in the *deep learning* literature (Goodfellow et al., 2016). This is because all of them assume both latent and observed variables are multivariate binary and that there are complex dependencies in between.

We now summarize the general model setup of SLAMs. A latent attribute pattern is denoted by a K-dimensional vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ of binary entries, where $\alpha_k \in \{0,1\}$ denotes the presence or absence of the kth attribute. Conditional on a subject's latent attribute pattern $\boldsymbol{\alpha} \in \{0,1\}^K$, his/her responses to the J items are assumed to be independent Bernoulli random variables with parameters $\theta_{1,\boldsymbol{\alpha}}, \dots, \theta_{J,\boldsymbol{\alpha}}$. Specifically, $\theta_{j,\boldsymbol{\alpha}} = \mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha})$ denotes the positive response probability, and is also called an item parameter of item j. We collect all the item parameters in the matrix $\boldsymbol{\Theta} = (\theta_{j,\boldsymbol{\alpha}})$, which has size $J \times 2^K$ with rows indexed by the J items and columns by the 2^K attribute patterns. For pattern $\boldsymbol{\alpha} \in \{0,1\}^K$, we denote its corresponding column vector in $\boldsymbol{\Theta}$ by $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}}$.

Corresponding to Property 1, the key assumption in a SLAM is that for a latent attribute pattern $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ and item j, the parameter $\theta_{j,\boldsymbol{\alpha}}$ is only determined by whether $\boldsymbol{\alpha}$ possesses the attributes in the set $\mathcal{K}_j = \{k \in \{1, \dots, K\} : q_{j,k} = 1\}$; that is, those attributes related to item j as specified in the Q-matrix. We will sometimes call the attributes in $\mathcal{K}_j$ the *required attributes of item* j. Under this assumption, all latent attribute patterns in the set

$$\mathcal{C}_j = \{\boldsymbol{\alpha} \in \{0,1\}^K : \boldsymbol{\alpha} \succeq \boldsymbol{q}_j\} \tag{1.1}$$

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

$$\max_{\boldsymbol{\alpha}\in\mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}\in\mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}} \text{ for any } j \in \{1,\dots,J\}. \tag{1.2}$$

We will call the set $\mathcal{C}_j$ a *constraint set*. Thus, the Q-matrix puts constraints on $\boldsymbol{\Theta}$ by forcing certain entries of it to be the same. Different SLAMs model the dependence of $\theta_{j,\boldsymbol{\alpha}}$ on the parent attributes in $\mathcal{K}_{\boldsymbol{q}_j}$ differently to encode different scientific assumptions; please see Examples I.1–I.3.

In addition to (1.2), another key assumption in SLAMs is the monotonicity assumption that

$$\theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'} \text{ for any } \boldsymbol{\alpha} \in \mathcal{C}_j,\ \boldsymbol{\alpha}' \notin \mathcal{C}_j. \tag{1.3}$$

Constraint (1.3) is commonly used in our motivating applications of cognitive diagnosis in educational assessments, where (1.3) indicates subjects mastering all required attributes of an item are more "capable" of giving a positive response to it (i.e., with a larger Bernoulli parameter $\theta_{j,\boldsymbol{\alpha}}$), than those who lack some required attributes. Nonetheless, our theoretical results of model identifiability in the following chapters also apply if (1.3) is relaxed to $\theta_{j,\boldsymbol{\alpha}} \neq \theta_{j,\boldsymbol{\alpha}'}$ for any $\boldsymbol{\alpha} \in \mathcal{C}_j$, $\boldsymbol{\alpha}' \notin \mathcal{C}_j$. This allows more flexibility in the model assumptions of SLAMs used in other applications.

Under the introduced notations, the probability mass function of a subject's response vector $\boldsymbol{R} = (R_1,\dots,R_J)^\top$ can be written as

$$\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{r_j} (1-\theta_{j,\boldsymbol{\alpha}})^{1-r_j}, \tag{1.4}$$

for $\boldsymbol{r} \in \{0,1\}^J$. Alternatively, the responses can be viewed as a J-th order tensor and the probability mass function of $\boldsymbol{R}$ can be written as a probability tensor as follows.

$$\mathbb{P}(\boldsymbol{R} \mid \boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{l=1}^{2^K} p_{\boldsymbol{\alpha}_l} \begin{pmatrix} \theta_{1,\boldsymbol{\alpha}_l} \\ 1-\theta_{1,\boldsymbol{\alpha}_l} \end{pmatrix} \circ \begin{pmatrix} \theta_{2,\boldsymbol{\alpha}_l} \\ 1-\theta_{2,\boldsymbol{\alpha}_l} \end{pmatrix} \circ \cdots \circ \begin{pmatrix} \theta_{J,\boldsymbol{\alpha}_l} \\ 1-\theta_{J,\boldsymbol{\alpha}_l} \end{pmatrix}, \tag{1.5}$$

where "$\circ$" denotes the tensor outer product and $\theta_{j,\boldsymbol{\alpha}}$'s are constrained by (1.2) and (1.3).

1.2 Model Examples: in Cognitive Diagnostic Modeling and in Machine Learning

The structured latent attribute models have recently gained great interests in cognitive diagnosis with applications in educational assessment, psychiatric evaluation and many other disciplines (e.g., Rupp et al., 2010; de la Torre, 2011; Culpepper, 2015; Wang et al., 2018; Chen et al., 2018b). Cognitive diagnosis is the process of arriving at a classification-based decision about an individual's latent attributes, based on the observed surrogate responses to a set of items. Such diagnostic information plays an important role in constructing efficient, focused remedial strategies for improvement in individual performance.

The structured latent attribute models are important statistical tools in cognitive diagnosis to detect the presence or absence of multiple fine-grained attributes. Cognitive diagnosis models in the psychometrics literature mostly consist of binary attributes, while general diagnostic models with categorical attributes were also considered in von Davier (2008). This dissertation focuses on the case of binary attributes.

In the following, we review some popular cognitive diagnosis models and illustrate how they fall into the family of structured latent attribute models. We first introduce some notation. For two vectors $\boldsymbol{a} = (a_1, \ldots, a_K)$, $\boldsymbol{b} = (b_1, \ldots, b_K)$ of the same dimension K, we write $\boldsymbol{a} \succeq \boldsymbol{b}$ if $a_i \geq b_i$ for all $i = 1, \ldots, K$; and $\boldsymbol{a} \succneqq \boldsymbol{b}$ if $\boldsymbol{a} \succeq \boldsymbol{b}$ and $\boldsymbol{a} \neq \boldsymbol{b}$. Denote $\boldsymbol{a} - \boldsymbol{b} = (a_1 - b_1, \ldots, a_K - b_K)$ and $\boldsymbol{a} \vee \boldsymbol{b} = (\max\{a_1, b_1\}, \ldots, \max\{a_K, b_K\})$. We also denote the all-zero and all-one vectors by $\mathbf{0}$ and $\mathbf{1}$, respectively.

Example I.1 (Conjunctive DINA and Disjunctive DINO)**.** The Deterministic Input Noisy output "And" gate (DINA) model proposed in Junker and Sijtsma (2001) and the Deterministic Input Noisy output "Or" gate (DINO) model proposed in Templin and Henson (2006) are popular and basic diagnostic models, which adopt the conjunctive and disjunctive assumptions, respectively. Specifically, under DINA, a subject needs to master all the required attributes of an item to be "capable" of it, and mastering the attributes not required by the item will not compensate for the lack of the required ones. That is, the required attributes of an item act "conjunctively" to define two knowledge states, with the following positive response probability

$$\theta_{j,\boldsymbol{\alpha}}^{DINA} = \begin{cases} 1 - s_j, & \text{if } \boldsymbol{\alpha} \succeq \boldsymbol{q}_j, \\ g_j, & \text{otherwise.} \end{cases}$$

where s_j is the slipping parameter, which denotes the probability that a capable subject slips the positive response, and g_j is the guessing parameter, which denotes the probability that a non-capable subject coincidentally gives the positive response by guessing. Under DINO, a subject only needs to master one of the required attributes to be "capable" of an item. That is, the required attributes of an item act "disjunctively" and

$$\theta_{j,\boldsymbol{\alpha}}^{DINO} = \begin{cases} 1 - s_j, & \text{if } \exists k \text{ s.t. } \alpha_k = q_{j,k} = 1, \\ g_j, & \text{otherwise.} \end{cases}$$

where s_j and g_j are the slipping and guessing parameters. Both the DINA and DINO models assume $1 - s_j > g_j$ for all j.

Example I.2 (Main-Effect Cognitive Diagnosis Models)**.** An important family of cognitive diagnosis models assume that the $\theta_{j,\boldsymbol{\alpha}}$ depends on the main effects of those attributes required by item j, but not their interactions. The main-effect models assume the main effects of the required attributes in $\mathcal{K}_{\boldsymbol{q}_j}$ play a role in distinguishing the item parameters, which can be written as

$$\begin{aligned}\theta^{\text{main-eff}}_{j,\boldsymbol{\alpha}} &= f\Big(\beta_{j,0} + \sum\nolimits_{k\in\mathcal{K}_{\boldsymbol{q}_j}} \beta_{j,k}\alpha_k\Big) \\ &= f\Big(\beta_{j,0} + \sum_{k=1}^{K} q_{j,k}\beta_{j,k}\alpha_k\Big),\end{aligned} \tag{1.6}$$

where $f(\cdot)$ is a link function. Note that not all β-coefficients in the second equivalent definition in the above equation are included in the model. For an attribute $k \in \{1, \dots, K\}$, $\beta_{j,k} \neq 0$ only if $q_{j,k} = 1$. We interpret this as $f(\beta_{j,0})$ denoting the probability of a positive response when none of the required attributes are present in $\boldsymbol{\alpha}$; when $q_{j,k} = 1$, $\beta_{j,\{k\}}$ is included in the model, representing the change in the positive response probability resulting from the mastery of a single attribute k. Different link functions $f(\cdot)$ lead to different models. Specifically, the popular reduced Reparameterized Unified Model (reduced-RUM; DiBello et al., 1995) has $f(\cdot)$ being the exponential function $\theta^{RUM}_{j,\boldsymbol{\alpha}} = \theta^{+}_{j}\prod_{k=1}^{K} r_{j,k}^{q_{j,k}(1-\alpha_k)}$, where $\theta^{+}_{j} = P(R_j = 1 \mid \boldsymbol{\alpha} \succeq \boldsymbol{q}_j)$ represents the positive response probability of a capable subject of j, and $r_{j,k} \in (0,1)$ is the parameter penalizing not possessing attribute k required by item j. Equivalently, the item parameter in reduced-RUM can be written as $\log\theta^{RUM}_{j,\boldsymbol{\alpha}} = \beta_{j,0} + \sum_{k=1}^{K}\beta_{j,k}(q_{j,k}\alpha_k)$, where $\beta_{j,k} \geq 0$ for $q_{j,k} = 1$. The Linear Logistic Model (LLM; Maris, 1999) has $f(\cdot)$ being the sigmoid function with $\text{text}(\theta^{LLM}_{j,\boldsymbol{\alpha}}) = \beta_{j,0} + \sum_{k=1}^{K}\beta_{j,k}(q_{j,k}\alpha_k)$. And the Additive Cognitive Diagnosis Model (ACDM; de la Torre, 2011) with $f(\cdot)$ the identity function.

Example I.3 (All-Effect Cognitive Diagnosis Models)**.** Another type of multi-parameter SLAMs are the all-effect models. These models assume that the positive response probability depends on the main effects and the interaction effects of the parent attributes of the item. The item parameter of an all-effect model can be written as

$$\begin{aligned}\theta_{j,\boldsymbol{\alpha}}^{\text{all-eff}} &= f\Big(\sum\nolimits_{S\subseteq\mathcal{K}_{\boldsymbol{q}_j}}\beta_{j,S}\prod\nolimits_{k\in S}\alpha_k\Big) \qquad (1.7)\\ &= f\Big(\beta_{j,\varnothing}+\sum_{k=1}^{K}q_{j,k}\beta_{j,\{k\}}\alpha_k+\sum_{1\leq k\neq k'\leq K}q_{j,k}q_{j,k'}\beta_{j,\{k,k'\}}\alpha_k\alpha_{k'}+\cdots\\ &\qquad \beta_{j,\{1,2,\cdots,K\}}\prod_{k=1}^{K}(q_{j,k}\alpha_k)\Big).\end{aligned}$$

Still note that not all β-coefficients in the second equivalent definition in the above equation are modeled. For a subset $\mathcal{S}$ of the K attributes $\{1,\ldots,K\}$, $\beta_{j,\mathcal{S}}\neq 0$ only if $\prod_{k\in\mathcal{S}}q_{j,k}=1$. When $q_{j,k}=1$, $\beta_{j,\{k\}}$ is included in the model, representing the change in the positive response probability resulting from the mastery of a single attribute k; when $q_{j,k}=q_{j,k'}=1$, $\beta_{j,\{k,k'\}}$ is included in the model, representing the change in the positive response probability resulting from the interaction effect of mastering both k and k', etc. When the link function $f(\cdot)$ is the identity, (1.7) gives the Generalized DINA (GDINA) model proposed by de la Torre (2011). Note that the DINA model is a submodel of the GDINA model by setting all the $\beta_{j,\,S}$ coefficients in (1.7), other than $\beta_{j,\,\varnothing}$ and $\beta_{j,\,\mathcal{K}_{\boldsymbol{q}_j}}$, to zero. Similar to the GDINA model, the LCDM adopts the logistic link function and assumes that $\text{logit}(\theta_{j,\,\boldsymbol{\alpha}}^{LCDM})=\sum_{S\subseteq\mathcal{K}_{\boldsymbol{q}_j}}\beta_{j,\,S}\prod_{k\in S}\alpha_k$. When the link function $f(\cdot)$ is the sigmoid function, (1.7) gives the Log-linear Cognitive Diagnosis Models (LCDMs) proposed by Henson et al. (2009); see also the General Diagnostic Models (GDMs) proposed in von Davier (2008).

All the cognitive diagnosis models reviewed in Examples I.1–I.3 are structured latent attribute models. Other than these examples in the psychometrics literature, the following is another example of SLAM in the deep learning literature.

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

$$\mathbb{P}(\boldsymbol{R}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}) \propto \exp\Big(- \boldsymbol{R}^\top \boldsymbol{W}^Q \boldsymbol{\alpha}^{(1)} - (\boldsymbol{\alpha}^{(1)})^\top \boldsymbol{U} \boldsymbol{\alpha}^{(2)} - \boldsymbol{f}^\top \boldsymbol{R} - \boldsymbol{b}_1^\top \boldsymbol{\alpha}^{(1)} - \boldsymbol{b}_2^\top \boldsymbol{\alpha}^{(2)} \Big), \tag{1.8}$$

where $\boldsymbol{f} \in \mathbb{R}^J$, $\boldsymbol{b}_i \in \mathbb{R}^{K_i}$ for $i = 1, 2$, and $\boldsymbol{W}^Q = (w_{j,k}) \in \mathbb{R}^{J \times K_1}$, $\boldsymbol{U} \in \mathbb{R}^{K_1 \times K_2}$ are model parameters; Figure 3.1 gives an example of a DBM with a 5×4 Q-matrix. For $\boldsymbol{f} = (f_1, \ldots, f_J)^\top$ and $\boldsymbol{\alpha}^{(1)} = (\alpha_1^{(1)}, \ldots, \alpha_{K_1}^{(1)})$, the conditional distribution of an observed variable R_j given the latent variables is

$$\mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \cdots) = \mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha}^{(1)}) = \frac{\exp\Big(\sum_{k=1}^{K_1} w_{j,k} \alpha_k^{(1)} + f_j \Big)}{1 + \exp\Big(\sum_{k=1}^{K_1} w_{j,k} \alpha_k^{(1)} + f_j \Big)}, \tag{1.9}$$

where " $\cdots$ " represents deeper latent layers that potentially exist in a DBM. Moreover, from (1.8) we have $\mathbb{P}(\boldsymbol{R} \mid \boldsymbol{\alpha}^{(1)}) = \prod_{j=1}^J \mathbb{P}(R_j \mid \boldsymbol{\alpha}^{(1)})$, so a DBM satisfies the local independence assumption that the R_j's are conditionally independent given the $\boldsymbol{\alpha}^{(1)}$. Therefore, a DBM can be viewed as a multi-parameter main-effect SLAM in (1.6) with a sigmoid link function. Viewing a DBM in this way, (B.73) gives the item parameter $\theta_{j,\boldsymbol{\alpha}^{(1)}}$, and the constraint set of each item j also takes the form $\mathcal{C}_j = \{\boldsymbol{\alpha}^{(1)} \in \{0,1\}^{K_1} : \boldsymbol{\alpha}^{(1)} \succeq \boldsymbol{q}_j\}$.

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

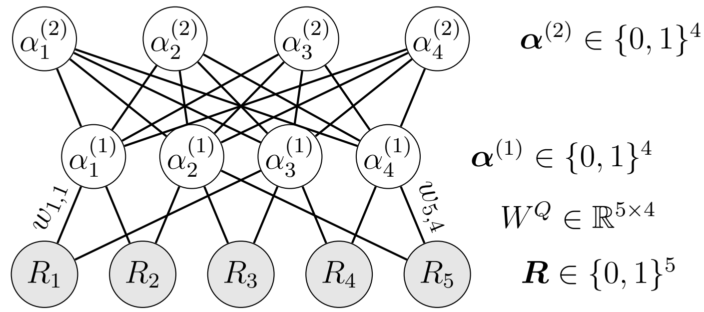


Figure 1.2: Deep Boltzmann Machine

1.3 Real Data Examples: in Designing Practice

To further illustrate the structural constraints induced by the Q-matrix, we next present several real-world applications that utilize SLAMs as cognitive diagnosis modeling tools.

Example I.5 (TOEFL Internet-based Testing Data)**.** TOEFL, short for Test of English as a Foreign Language, is a standardized test to measure English language ability of non-native speakers. Restricted latent class models have been used to analyze the TOEFL data by researchers at Educational Testing Service (ETS; e.g., von Davier, 2005, 2008). For instance, von Davier (2008) proposed a general diagnostic model (GDM), which was used to analyze the TOEFL reading section of two parallel forms, A and B, with their Q-matrices analyzed and specified by content experts. In particular, the forms A and B contain 39 and 40 items with four latent attributes: α_1*: word meaning,* α_2*: specific information,* α_3*: connect information, and* α_4*: synthesize and organize.* Table 1.1 gives the summary of the two Q-matrices by presenting each

$\boldsymbol{q}$-vector's frequencies in them. For instance, the first line in Table 1.1 reads (1, 0, 0, 0) for the row $\boldsymbol{q}$-vector and (9, 9) for the frequencies. This means that there are nine items with $\boldsymbol{q}$-vector (1, 0, 0, 0) in form A and nine in form B, respectively. Under the restrictions induced by the Q-matrices, the diagnostic models used to analyze the TOEFL data fall in the family of restricted latent class models.

Table 1.1: Q-matrix entry frequencies, TOEFL iBT field test, Reading Forms A & B

Q-matrix row $\boldsymbol{q}$-vectors				$\boldsymbol{q}$-vector frequency	
α_1 word meaning	α_2 specific information	α_3 connect information	α_4 synthesize & organize	Form A	Form B
1	0	0	0	9	9
0	1	0	0	8	11
1	1	0	0	1	1
0	0	1	0	10	10
1	0	1	0	0	1
0	1	1	0	2	0
0	1	0	1	1	0
0	0	1	1	7	8
1	0	1	1	1	0

Example I.6 (Trends in International Mathematics and Science Study)**.** Trends in International Mathematics and Science Study (TIMSS) is a large scale cross-country assessment, administered by the International Association for the Evaluation of Educational Achievement. TIMSS evaluates the mathematics and science abilities of fourth and eighth graders every four years since 1995 and covers more than 40 countries. The TIMSS data allows one to analyze trends in student progress that can provide feedback for future improvement in areas needing further instruction (Lee et al., 2011). Researchers have used the cognitive diagnosis models to analyze the TIMSS data (e.g., Lee et al., 2011; Choi et al., 2015; Yamaguchi and Okada, 2018). For instance, a 43×12 Q-matrix constructed by mathematics educators and researchers was specified for the TIMSS 2003 eighth grade mathematics assessment (Choi et al., 2015). A total number of 12 fine-grained attributes are identified, which fall in five big categories of skill domains measured by the eighth grade exam, Number, Algebra, Geometry, Measurement, and Data. The Q-matrix is presented in Table 1 in the Supplementary Material. Choi et al. (2015) used DINA model to fit the dataset containing responses sampled from 8912 U.S. and 5309 Korean students. Main-Effect and All-Effect diagnostic models have also been applied to analyze the TIMSS data (e.g., Yamaguchi and Okada, 2018).

Example I.7 (Fraction Subtraction Data)**.** The fraction subtraction dataset is widely analyzed in the psychometrics literature (de la Torre and Douglas, 2004a; DeCarlo, 2011; Henson et al., 2009; de la Torre, 2011). The dataset contains 536 middle school students' binary responses to 20 fraction subtraction items that were designed for diagnostic assessment. Table 1.3 presents the Q-matrix specified in de la Torre and Douglas (2004a), which corresponds to the $K = 8$ skill attributes regarding doing fraction and subtraction. The eight attributes are (α_1) Convert a whole number to a fraction; (α_2) Separate a whole number from a fraction; (α_3) Simplify before subtracting; (α_4) Find a common denominator; (α_5) Borrow from whole number part; (α_6) Column borrow to subtract the second numerator from the first; (α_7) Subtract numerators; (α_8) Reduce answers to simplest form. Many researchers have used various structured latent attribute models models to fit this dataset (e.g., de la Torre and Douglas, 2004b; DeCarlo, 2011; Henson et al., 2009; de la Torre, 2011).

1.4 Unique Challenges of SLAMs and Our Contributions

The family of SLAMs bring advantages both in representational power and in scientific interpretability. As mentioned earlier, multiple latent attributes can represent various meaningful real-world constructs, and also the structural matrix Q can encode the information of study design or scientific prior knowledge. However, despite the popularity and advantages of SLAMs, this family of models also bring several unique challenges and yield important open problems.

The first challenge is the fundamental identifiability issue associated with SLAMs. Indeed, this has long been recognized as a problem, as pointed out by practitioners and researchers in the literature. The following are quotes from researchers in the educational and psychological measurements, just to name a few:

(a) Maris and Bechger (2009): *"Identifiability of the parameters from the observations remains problematic for most diagnostic classification models [SLAMs]. [For these models] the problem is much harder and much less trivial."*

(b) Huebner (2010): *"Identification of parameters is increasingly difficult with increasing numbers of skills in the model"*

(c) von Davier (2014): *"The literature on assessing identifiability of diagnostic models [SLAMs] is sparse at best... There is little [study] to be found."*

Model identifiability is the first and foremost prerequisite for drawing any valid statistical inference. In statistical terms, a model is identifiable if all the parameters can be uniquely determined by the distribution of the observed data. For SLAMs, identifiability issues are challenging to address, due to (1) the discreteness nature of all the random variables, (2) the existence of many latent attributes, and (3) the complex constraints imposed by the Q-matrix.

As previously mentioned, SLAMs can be viewed as *restricted* latent class models. The study of identifiability of latent class models dates back to decades ago (McHugh, 1956; Teicher, 1967; Goodman, 1974). For unrestricted latent class models, Gyllenberg et al. (1994) showed the model is not identifiable in the sense that, there always exists some set of parameters, such that one can construct a different set of parameters which lead to the same distribution of the responses. Such nonidentifiablity has likely impeded statisticians from looking further into this problem (Allman et al., 2009). Due to the difficulty of establishing strict identifiability in such scenarios, Elmore et al. (2005) and Allman et al. (2009) studied the *generic* identifiability of these models. The idea of generic identifiability is closely related to concepts in algebraic geometry and implies that the model parameters are identifiable almost everywhere in the parameter space, excluding only a Lebesgue measure zero set. Allman et al. (2009) established generic identifiability results for various latent variable models, including the unrestricted latent class models. The complex constraints of SLAMs pose additional challenge to the study of model identifiability. The existing results of generic identifiability in Allman et al. (2009) do not apply to SLAMs, because the restrictions imposed by the structural matrix Q already constrain the model parameters of a SLAM into a measure-zero (and hence potentially unidentifiable) subset of the parameter space of an unrestricted latent class model.

Another type of challenges accompanying the application of SLAMs is the estimation and computation difficulty in high dimensions. Since the latent attributes are modeled as multivariate categorical, given a moderate to large number of discrete attributes K, the size of the latent pattern space grows exponentially with K. This poses big challenges to both estimation and computation methodology. In real-world applications of SLAMs, the number of potential latent classes can be much larger than the sample size. For instance, the Trends in International Mathematics and Science Study (TIMSS) is an international educational assessment that provides reliable and timely data on the mathematics and science achievement of middle school students. In a TIMSS dataset with eighth-graders, the number of attributes of interest is $K = 15$, leading to $2^{15} = 32768$ configurations of binary latent patterns; while the available sample size is only hundreds. For interpretability, it is often assumed that only a small subset of attribute patterns exist. In these high-dimensional settings with such "sparsity" structure, existing estimation methods tend to over-select too many latent classes, and also incur excessive computational cost. Therefore, valid statistical methods and scalable computational algorithms to attack combinatorial estimation problems in high-dimensional settings are severely called for.

This dissertation addresses these research questions and has several contributions outlined as follows. On the theory of identifiability, Chapter II[1] fully answers the question that under what conditions the popular and basic DINA model (Junker and Sijtsma, 2001) is identifiable, by providing the necessary and sufficient conditions for strict identifiability. Chapter III[2] develops practical partial and generic identifiability theory for a general family of SLAMs, motivated by real-world needs of designing cognitive diagnostic tests with minimal restrictions. The new theory is applied to give the affirmative answer of identifiability to models with several aforementioned real world designs, for the first time in the literature. Chapter IV[3] addresses a further question, which goes beyond merely identifying the model parameters. Rather, here the main goal is to identify the key latent structure, that is, the Q-matrix itself. This chapter includes various results of identifying the Q-matrix, which is a technically much more challenging than establishing identifiability given a known Q-matrix.

On the methodological and computational side, Chapter V[4] deals with the challenge in modern applications of SLAMs is the high-dimensional latent attribute patterns. The methodological contribution in this chapter is a penalized likelihood method to select significant latent patterns in the high-dimensional scenario. The computational contribution includes a scalable screening algorithm as a preprocessing step that drastically reduces the computational cost of the method. Going a step further from learning general sparse latent patterns, Chapter VI addresses the identification and estimation problem of hierarchical latent attribute models. These models incorporate an additional ingredient on top of SLAMS: hierarchical constraints on which configurations of the attributes are allowed. This chapter addresses the ques-

[1]mainly corresponding to Gu and Xu (2019b), *Psychometrika*.

[2]mainly corresponding to Gu and Xu (2020a), accepted by the *Annals of Statistics*.

[3]mainly corresponding to Gu and Xu (2020b), accepted by *Statistica Sinica*.

[4]mainly corresponding to Gu and Xu (2019a), *Journal of Machine Learning Research*.

tion of identifiability under arbitrary attribute hierarchies, and further proposes a scalable algorithm for estimating both the latent structural matrix and the attribute hierarchy from the noisy data. Each chapter from Chapter II to Chapter VI has a corresponding appendix containing all the technical proofs and additional numerical results. All the appendices come after the main chapters.

Table 1.2: Q-matrix, TIMSS 2003 8th Grade Data

Item ID	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}
1	1	0	1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0
3	1	0	1	0	0	0	0	0	0	1	0	0
4	0	1	1	0	0	0	0	0	0	1	0	0
5	0	0	0	0	0	0	1	0	0	0	1	0
6	0	0	0	0	0	0	1	0	0	1	1	0
7	0	0	0	0	0	0	1	0	0	0	1	0
8	0	1	0	1	0	0	0	0	0	0	0	0
9	0	0	0	1	0	0	1	0	0	0	0	0
10	0	0	0	1	1	1	0	0	0	0	0	0
11	0	0	0	1	1	1	0	0	0	0	0	0
12	1	0	0	0	1	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	1	1	0	0	0
14	0	0	0	0	0	0	0	1	1	0	0	0
15	0	0	0	0	0	0	0	1	0	0	0	0
16	0	0	0	1	1	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	1	0	0
18	1	1	1	0	0	0	0	0	0	1	0	0
19	0	0	0	1	0	0	0	0	0	0	0	1
20	1	0	0	0	0	0	1	0	0	0	0	0
21	1	1	0	0	0	0	0	0	0	0	0	0
22	0	0	1	1	1	0	0	0	0	0	0	0
23	1	0	0	0	0	0	0	0	0	1	0	0
24	0	0	0	0	0	0	0	0	1	1	0	0
25	0	0	0	0	0	0	1	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0	0	1
27	1	0	0	0	0	0	0	0	0	0	0	0
28	1	1	0	0	0	0	0	0	0	0	0	0
29	1	1	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	1	1	0
31	0	0	1	0	0	0	1	0	0	0	1	0
32	1	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	1	1	0	0	0	0
34	0	1	0	0	1	0	0	0	0	0	0	0
35	0	0	0	0	0	0	1	0	0	1	0	0
36	0	0	0	0	1	0	0	1	1	0	0	0
37	0	0	0	0	0	0	1	0	0	0	0	0
38	0	0	0	0	1	1	0	0	0	0	0	0
39	1	0	0	0	0	0	0	0	0	0	0	0
40	1	1	1	0	0	0	0	0	0	0	0	0
41	1	0	0	0	0	0	0	0	0	0	0	0
42	1	1	0	1	0	0	0	0	0	0	0	0
43	1	0	0	0	0	0	0	0	0	0	0	1

Table 1.3: Q-matrix, Fraction Data

Item ID	Content	$K = 8$ attributes							
		α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
1	$\frac{5}{3} - \frac{3}{4}$	0	0	0	1	0	1	1	0
2	$\frac{3}{4} - \frac{3}{8}$	0	0	0	1	0	0	1	0
3	$\frac{5}{6} - \frac{1}{9}$	0	0	0	1	0	0	1	0
4	$3\frac{1}{2} - 2\frac{3}{2}$	0	1	1	0	1	0	1	0
5	$4\frac{3}{5} - 3\frac{4}{10}$	0	1	0	1	0	0	1	1
6	$\frac{6}{7} - \frac{4}{7}$	0	0	0	0	0	0	1	0
7	$3 - 2\frac{1}{5}$	1	1	0	0	0	0	1	0
8	$\frac{2}{3} - \frac{2}{3}$	0	0	0	0	0	0	1	0
9	$3\frac{7}{8} - 2$	0	1	0	0	0	0	0	0
10	$4\frac{4}{12} - 2\frac{7}{12}$	0	1	0	0	1	0	1	1
11	$4\frac{1}{3} - 2\frac{4}{3}$	0	1	0	0	1	0	1	0
12	$\frac{11}{8} - \frac{1}{8}$	0	0	0	0	0	0	1	1
13	$3\frac{3}{8} - 2\frac{5}{6}$	0	1	0	1	1	0	1	0
14	$3\frac{4}{5} - 3\frac{2}{5}$	0	1	0	0	0	0	1	0
15	$2 - \frac{1}{3}$	1	0	0	0	0	0	1	0
16	$4\frac{5}{7} - 1\frac{4}{7}$	0	1	0	0	0	0	1	0
17	$7\frac{3}{5} - \frac{4}{5}$	0	1	0	0	1	0	1	0
18	$4\frac{1}{10} - 2\frac{8}{10}$	0	1	0	0	1	1	1	0
19	$4 - 1\frac{4}{3}$	1	1	1	0	1	0	1	0
20	$4\frac{1}{3} - 1\frac{5}{3}$	0	1	1	0	1	0	1	0

CHAPTER II

Necessary and Sufficient Condition for the Identifiability of the DINA Model

The DINA model introduced in Example I.1 is a very popular and basic Cognitive Diagnostic Model (CDM). It also serves as a submodel for some general structured latent attribute models, such as the GDINA model introduced in Example I.3. Recently there have been several studies on the identifiability of the CDMs, including the DINA model (e.g., Xu and Zhang, 2016). However, the existing works mostly focus on developing sufficient conditions for identifiability, which might impose stronger than needed or sometimes even impractical constraints on designing identifiable cognitive diagnostic tests. It remains an open problem in the literature what would be the minimal requirement, i.e., the necessary and sufficient conditions, for the models to be identifiable. In particular, for the DINA model, Xu and Zhang (2016) proposed a set of sufficient conditions and a set of necessary conditions for the identifiability of the slipping, guessing and population proportion parameters. However, as pointed out by the authors, there is a gap between the two sets of conditions; see Xu and Zhang (2016) for examples and discussions.

This chapter addresses this open problem by developing the necessary and suf-

This chapter contains the main part of Gu and Xu (2019b), *Psychometrika*.

ficient condition for the identifiability of the DINA model. Furthermore, we show that the identifiability condition ensures the statistical consistency of the maximum likelihood estimators of the model parameters. The proposed condition not only guarantees identifiability, but also gives the minimal requirement that the DINA model needs to meet in order to be identifiable. The identifiability result can be directly applied to the DINO model (Templin and Henson, 2006) through the duality of the DINA and DINO models (Chen et al., 2015). For general CDMs such as the LCDM and GDINA models, since the DINA model can be considered as a submodel of them, the proposed condition also serves as a necessary requirement. From a practical perspective, the necessary and sufficient condition only depends on the Q-matrix structure and hence is easily checkable. Such condition would provide a practical guideline for designing statistically valid and estimable cognitive tests.

The rest of this chapter is organized as follows. Section 2.1 states the main result and includes several illustrating examples. Section 2.2 gives a brief discussion. The proofs of the main results are included in Appendix A.

2.1 Main Theorem of Necessity and Sufficiency

We first introduce the important concept of the "completeness" of a Q-matrix, which was first introduced in Chiu et al. (2009). A Q-matrix is said to be complete if it can differentiate all latent attribute profiles, in the sense that under the Q-matrix, different attribute profiles have different response distributions. In this study of the DINA model, completeness of the Q-matrix means that $\{\boldsymbol{e}_k^\top : k = 1, \ldots, K\} \subseteq \{\boldsymbol{q}_j : j = 1, \ldots, J\}$, equivalently, for each attribute there is some item which requires that and solely requires that attribute. Up to some row permutation, a complete Q-matrix under the DINA model contains a $K \times K$ identity matrix. Under the DINA model, completeness of the Q-matrix is necessary for identifiability of the population proportion parameters $\boldsymbol{p}$ (Xu and Zhang, 2016).

Besides the completeness, an additional necessary condition for identifiability was also specified in Xu and Zhang (2016) that each attribute needs to be related with at least three items. For ease of discussion, the set of necessary conditions in Xu and Zhang (2016) are summarized as follows.

Condition II.1. The Q-matrix is complete under the DINA model and without loss of generality, we assume the Q-matrix takes the following form:

$$Q = \begin{pmatrix} I_K \\ \hdashline Q^\star \end{pmatrix}_{J\times K}, \tag{2.1}$$

where I_K denotes the $K \times K$ identity matrix and $Q^\star$ is a $(J-K)\times K$ submatrix of Q.

Condition II.2. Each of the K attributes is required by at least 3 items.

Though necessary, Xu and Zhang (2016) recognized that Condition 1 is not sufficient. To establish identifiability, the authors also proposed a set of sufficient conditions, which however is not necessary. For instance, the Q-matrix in (2.2), which is given on page 633 in Xu and Zhang (2016), does not satisfy their sufficient condition but still gives an identifiable model.

$$Q = \begin{pmatrix} & & I_4 & \\ \hdashline 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2.2}$$

In particular, their sufficient condition C4 requires that for each $k \in \{1, \ldots, K\}$, there exist two subsets S_k^+ and S_k^- of the items (not necessarily nonempty or disjoint) in $Q^\star$ such that S_k^+ and S_k^- have attribute requirements that are identical except in the kth attribute, which is required by an item in S_k^+ but not by any item in S_k^-. However, the first attribute in (2.2) does not satisfy this condition. Examples of this kind of

Q-matrices not satisfying their C4 but still identifiable are not rare and can be easily constructed as shown below in (2.3).

$$Q = \begin{pmatrix} & I_3 & \\ \hdashline 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} & I_3 & \\ \hdashline 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} & I_3 & \\ \hdashline 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} & & I_4 & \\ \hdashline 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}. \tag{2.3}$$

It has been an open problem in the literature what would be the minimal requirement of the Q-matrix for the model to be identifiable. This chapter solves this problem and shows shat Condition 1 together with the following Condition 2 are sufficient and necessary for the identifiability of the DINA model parameters.

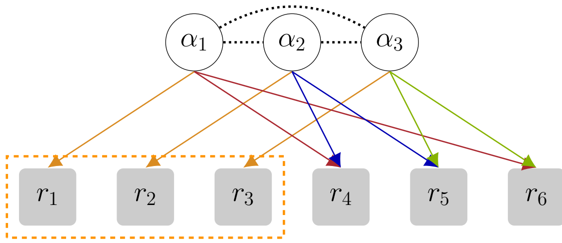
Condition II.3. Any two different columns of the sub-matrix $Q^\star$ in (2.1) are distinct.

We have the following identifiability result.

Theorem II.1 (Sufficient and Necessary Condition)**.** *Conditions II.1, II.2, II.3 are sufficient and necessary for the identifiability of all the DINA model parameters.*

Remark II.1. From the model construction, when there are some items that require none of the attributes, all the DINA model parameters are $(\boldsymbol{s}, \boldsymbol{p})$ and $\boldsymbol{g}^- = (g_j : \forall j \text{ such that } \boldsymbol{q}_j \neq \mathbf{0})^\top$. Theorem II.1 also applies to this special case that the proposed conditions still remain sufficient and necessary for the identifiability of $(\boldsymbol{s}, \boldsymbol{g}^-, \boldsymbol{p})$, under a Q-matrix containing some all-zero $\boldsymbol{q}$-vectors. See Proposition A.2 in the Appendix for more details.

Conditions II.1, II.2, and II.3 are easy to verify. Equivalently, these conditions can be written as three topological properties A, B and C of the bipartite graph corresponding to the Q-matrix, as shown in the example in Figure 2.1. Based on Theorem II.1, it is recommended in practice to design the Q-matrix such that it is



A. a perfect matching: orange edges

B. each attribute has ≥ 3 children

C. removing the perfect matching, the K attributes has distinct children sets:

- red, blue, and green edges point to different sets of r_j's

Figure 2.1: illustrating necessary and sufficient conditions on the Q-matrix in an example

complete, has each attribute required by at least 3 items, and has K distinct columns in the sub-matrix $Q^\star$. Otherwise, the model parameters would suffer from the non-identifiability issue. We use the following examples to illustrate the theoretical result.

Example II.1. From Theorem II.1, the Q-matrices in (2.2) and (2.3) satisfy both Conditions II.1, II.2, II.3 and therefore give identifiable models, while the results in Xu and Zhang (2016) cannot be applied since their condition C4 does not hold. On the other hand, the Q-matrices below in (2.4) satisfy the necessary conditions in Xu and Zhang (2016), but they do not satisfy our Condition 2, so the corresponding models are not identifiable.

$$Q = \begin{pmatrix} \multicolumn{3}{c}{I_3} \\ \hdashline 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} \multicolumn{3}{c}{I_3} \\ \hdashline 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} \multicolumn{3}{c}{I_3} \\ \hdashline 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} \multicolumn{4}{c}{I_4} \\ \hdashline 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}. \tag{2.4}$$

Example II.2. To illustrate the necessity of Condition II.3, we consider a simple case when $K = 2$. If Conditions II.1 and II.2 are satisfied but Condition II.3 does not hold, the Q-matrix can only have the following form up to some row permutations,

$$
Q = \begin{pmatrix} \multicolumn{2}{c}{I_2} \\ \hdashline 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}_{J\times 2}, \tag{2.5}
$$

where the first two items give an identity matrix while the next J_0 items require none of the attributes and the last $J-2-J_0$ items require both attributes. Under the Q-matrix in (2.5), we next show the model parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are not identifiable by constructing a set of parameters $(\bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}}) \neq (\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ which satisfy (5.9). Recall from the model setup in Section 2 that for any item $j \in \{3, \ldots, J_0+2\}$ that has $\boldsymbol{q}_j = \boldsymbol{0}$, the guessing parameter is not needed by the DINA model and for notational convenience, we set $g_j \equiv \bar{g}_j \equiv 0$. We take $\bar{\boldsymbol{s}} = \boldsymbol{s}$, $\bar{g}_j = g_j$ for $j = J_0+3, \ldots, J$, and $\bar{p}_{(11)} = p_{(11)}$. Next we show the remaining parameters $\big(g_1, g_2, p_{(00)}, p_{(10)}, p_{(01)}\big)$ are not identifiable. From Definition 1, the non-identifiability occurs if the following equations hold (see the Supplementary Material for the computational details): $P\big((R_1, R_2) = (r_1, r_2) \mid Q, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}}\big) = P\big((R_1, R_2) = (r_1, r_2) \mid Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}\big)$ for all $(r_1, r_2) \in \{0,1\}^2$, where (R_1, R_2) are the first two entries of the random response vector $\boldsymbol{R}$. These equations can be further expressed as the following equations in (2.6):

$$
(r_1, r_2) = \begin{cases}
(0,0): \ \bar{p}_{(00)} + \bar{p}_{(10)} + \bar{p}_{(01)} + p_{(11)} = p_{(00)} + p_{(10)} + p_{(01)} + p_{(11)}; \\
(1,0): \ \bar{g}_1[\bar{p}_{(00)} + \bar{p}_{(01)}] + (1-s_1)[\bar{p}_{(10)} + p_{(11)}] \\
\qquad\qquad = g_1[p_{(00)} + p_{(01)}] + (1-s_1)[p_{(10)} + p_{(11)}]; \\
(0,1): \ \bar{g}_2[\bar{p}_{(00)} + \bar{p}_{(10)}] + (1-s_2)[\bar{p}_{(01)} + p_{(11)}] \\
\qquad\qquad = g_2[p_{(00)} + p_{(10)}] + (1-s_2)[p_{(01)} + p_{(11)}]; \\
(1,1): \ \bar{g}_1\bar{g}_2\bar{p}_{(00)} + \bar{g}_1(1-s_2)\bar{p}_{(01)} + (1-s_2)\bar{g}_2\bar{p}_{(10)} + (1-s_1)(1-s_2)p_{(11)} \\
\qquad = g_1 g_2 \bar{p}_{(00)} + g_1(1-s_2)p_{(01)} + (1-s_2)g_2 p_{(10)} + (1-s_1)(1-s_2)p_{(11)}.
\end{cases} \tag{2.6}
$$

For any $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$, there are 4 constraints in (2.6) but 5 parameters $\big(\bar{g}_1, \bar{g}_2, \bar{p}_{(00)}, \bar{p}_{(10)}, \bar{p}_{(01)}\big)$

to solve. Therefore there are infinitely many solutions and $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are non-identifiable.

Example II.3. We provide a numerical illustration of Example 2. Without loss of generality, we take $J_0 = 0$, since whether there exist zero $\boldsymbol{q}$-vector items makes no impact on the nonidentifiability phenomenon as illustrated in (2.6). We take $J = 10$ and set the true parameters to be $(p_{(00)},\ p_{(10)},\ p_{(01)},\ p_{(11)}) = (0.1,\ 0.3,\ 0.4,\ 0.2)$ and $s_j = g_j = 0.2$ for $j \in \{1, \ldots, 10\}$. We first generate a random sample of size $N = 200$. From the data, we obtain one set of maximum likelihood estimators as follows:

$$(\widehat{p}_{(00)},\ \widehat{p}_{(10)},\ \widehat{p}_{(01)},\ \widehat{p}_{(11)}) = (\mathbf{0.22346},\ \mathbf{0.26298},\ \mathbf{0.32847},\ 0.18509);$$
$$\widehat{\boldsymbol{s}} = (0.1269,\ 0.1541,\ 0.0000,\ 0.2015,\ 0.1549,\ 0.2638,\ 0.3551,\ 0.1903,\ 0.1843,\ 0.1468);$$
$$\widehat{\boldsymbol{g}} = (\mathbf{0.1678},\ \mathbf{0.2011},\ 0.2330,\ 0.1990,\ 0.2007,\ 0.2316,\ 0.2155,\ 0.1720,\ 0.2197,\ 0.1805).$$

Based on (2.6), we can construct infinitely many sets of $(\bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}})$ that are also maximum likelihood estimators. For instance, we take $\bar{\boldsymbol{s}} = \widehat{\boldsymbol{s}}$, $\bar{g}_j = \widehat{g}_j$ for $j = 3, \ldots, 10$, $\bar{p}_{(11)} = \widehat{p}_{(11)}$, and $\bar{p}_{(00)} = 0.998 \cdot \widehat{p}_{(00)}$. Then solve (2.6) for the remaining parameters $\bar{p}_{(10)}$, $\bar{p}_{(01)}$, $\bar{g}_1$ and $\bar{g}_2$ to get

$$\bar{p}_{(00)} = \mathbf{0.22301}, \quad \bar{p}_{(01)} = \mathbf{0.33306}, \quad \bar{p}_{(10)} = \mathbf{0.25884}, \quad \bar{g}_1 = \mathbf{0.2561}, \quad \bar{g}_2 = \mathbf{0.1073}.$$

The two different sets of values $(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{g}}, \widehat{\boldsymbol{p}})$ and $(\bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}})$ both give the identical log-likelihood value -1132.1264, which confirms the non-identifiablility.

To further illustrate the above argument does not depend on the sample size, we generate a random sample of size $N = 10^5$ and obtain the following estimators:

$$(\widehat{p}_{(00)},\ \widehat{p}_{(10)},\ \widehat{p}_{(01)},\ \widehat{p}_{(11)}) = (\mathbf{0.10436},\ \mathbf{0.29933},\ \mathbf{0.39845},\ 0.19786);$$
$$\widehat{\boldsymbol{s}} = (0.1968,\ 0.1932,\ 0.2007,\ 0.2065,\ 0.2015,\ 0.2000,\ 0.2001,\ 0.1949,\ 0.1985,\ 0.2036);$$
$$\widehat{\boldsymbol{g}} = (\mathbf{0.1993},\ \mathbf{0.2006},\ 0.1995,\ 0.2010,\ 0.1971,\ 0.1983,\ 0.1995,\ 0.2022,\ 0.1989,\ 0.1988).$$

Similarly, we set $\bar{\boldsymbol{s}} = \widehat{\boldsymbol{s}}$, $\bar{g}_j = \widehat{g}_j$ for $j = 3, \ldots, 10$, $\bar{p}_{(11)} = \widehat{p}_{(11)}$, and $\bar{p}_{(00)} = 0.998 \cdot \widehat{p}_{(00)}$. Solving (2.6) gives

$$\bar{p}_{(00)} = \mathbf{0.10415}, \quad \bar{p}_{(01)} = \mathbf{0.40161}, \quad \bar{p}_{(10)} = \mathbf{0.29638}, \quad \bar{g}_1 = \mathbf{0.3212}, \quad \bar{g}_2 = \mathbf{0.0458}.$$

where the two different sets of values $(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{g}}, \widehat{\boldsymbol{p}})$ and $(\bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}})$ both lead to the identical log-likelihood value -571659.1708. This illustrates that the non-identifiability issue depends on the model setting instead of the sample size. In practice, as long as Conditions 1 and 2 do not hold, we may suffer from similar non-identifiability issues no matter how large the sample size is.

Identifiability is the prerequisite and a necessary condition for consistent estimation. Here we say a parameter is consistently estimable if we can construct a consistent estimator for the parameter. That is, for parameter β, there exists $\widehat{\beta}_N$ such that $\widehat{\beta}_N - \beta \to 0$ in probability as the sample size $N \to \infty$. When the identifiability conditions are satisfied, we show that the maximum likelihood estimators (MLEs) of the DINA model parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are statistically consistent as $N \to \infty$. For the observed responses $\{\boldsymbol{R}_i : i = 1, \ldots, N\}$, we can write their likelihood function as

$$L_N(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p};\, \boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = \prod_{i=1}^{N} P(\boldsymbol{R} = \boldsymbol{R}_i \mid Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}), \tag{2.7}$$

where $P(\boldsymbol{R} = \boldsymbol{R}_i \mid Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is as defined in (1.4), with $\boldsymbol{\Theta}$ there replaced by the slipping and guessing parameters $\boldsymbol{s}$ and $\boldsymbol{g}$ in the DINA model. Let $(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{g}}, \widehat{\boldsymbol{p}})$ be the corresponding MLEs based on (2.7). We have the following corollary.

Corollary II.1. *When Conditions II.1, II.2, and II.3 are satisfied, the MLEs* $(\widehat{\boldsymbol{s}}, \widehat{\boldsymbol{g}}, \widehat{\boldsymbol{p}})$ *are consistent as* $N \to \infty$.

The results in Theorem II.1 and Corollary V.1 can be directly applied to the DINO model through the duality of the DINA and DINO models (see Proposition 1 in Chen et al., 2015). Specifically, when Conditions II.1, II.2, and II.3 are satisfied, the guessing, slipping, and population proportion parameters in the DINO model are identifiable and can also be consistently estimated as $N \to \infty$.

Moreover, the proof of Corollary V.1 can be directly generalized to the other CDMs that the MLEs of the model parameters, including the item parameters and population proportion parameters, are consistent as $N \to \infty$ if they are identifiable. Therefore under the sufficient conditions for identifiability of general CDMs developed in the literature such as Xu (2017), the model parameters are also consistently estimable. Although the minimal requirement for identifiability and estimability of general CDMs are still unknown, the proposed Conditions II.1, II.2, and II.3 are necessary since the DINA model is a submodel of them. For instance, Xu (2017) requires two identity matrices in the Q-matrix to obtain identifiability, which automatically satisfies Conditions II.1, II.2, and II.3 in this chapter.

We next present an example to illustrate that when the proposed conditions are satisfied, the MLEs of the DINA model parameters are consistent.

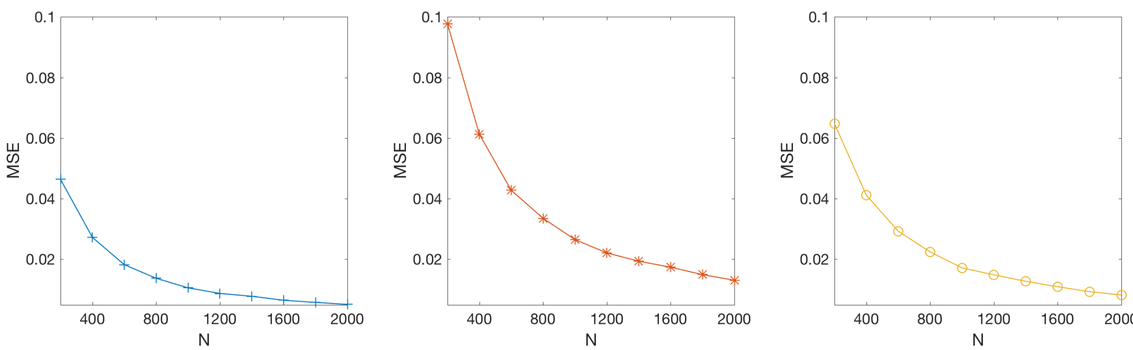
Example II.4. We perform a simulation study with the following Q-matrix that satisfies the proposed sufficient and necessary conditions. The true parameters are set to be $p_{\boldsymbol{\alpha}} = 0.125$ for all $\boldsymbol{\alpha} \in \{0,1\}^3$, and $s_j = g_j = 0.2$ for $j = 1, \ldots, 6$.

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

For each sample size $N = 200 \cdot i$ where $i = 1, \ldots, 10$, we generate 1000 independent datasets, and use the EM algorithm with random initializations to obtain the MLEs of model parameters for each dataset. The mean squared errors (MSEs) of the parameters $\boldsymbol{s}$, $\boldsymbol{g}$, $\boldsymbol{p}$ computed from the 1000 runs are shown in Table 2.1 and Figure 2.2. One can see that the MSEs keep decreasing as the sample size N increases, matching the theoretical result in Corollary V.1.

N	400	800	1200	1600	2000
$\boldsymbol{p}$	0.0272	0.0137	0.0087	0.0065	0.0051
$\boldsymbol{s}$	0.0613	0.0335	0.0221	0.0174	0.0131
$\boldsymbol{g}$	0.0411	0.0224	0.0149	0.0109	0.0082

Table 2.1: MSEs of DINA Model Parameters



(a) MSEs of $\boldsymbol{p}$ (b) MSEs of $\boldsymbol{s}$ (c) MSEs of $\boldsymbol{g}$

Figure 2.2: MSE of DINA Model Parameters versus Sample Size N

2.2 Discussion

This chapter presents the sufficient and necessary condition for identifiability of the DINA and DINO model parameters and establishes the consistency of the maximum likelihood estimators. As discussed in the previous section, the results would also shed light on the study of the sufficient and necessary conditions for general CDMs.

This chapter treats the attribute profiles as random effects from a population distribution. Under this setting, the identifiability conditions ensure the consistent estimation of the model parameters. However, generally in statistics and psychometrics, identifiability conditions are not always sufficient for consistent estimation. An example of identifiable but not consistently estimable is the fixed effects CDMs, where the subjects' attribute profiles are taken as model parameters. Consider a simple example of the DINA model with nonzero but known slipping and guessing parameters. Under the fixed effects setting, the model parameters include $\{\boldsymbol{\alpha}_i, i = 1, \ldots, N\}$, which are identifiable if the Q-matrix is complete (e.g., Chiu et al., 2009). But with fixed number of items, even when the sample size N goes to infinity, the parameters $\{\boldsymbol{\alpha}_i, i = 1, \ldots, N\}$ cannot be consistently estimated. In this case, to have the consistent estimation of each $\boldsymbol{\alpha}$, the number of items needs to go to infinity and the number of identity sub-Q-matrices also needs to go to infinity (Wang and Douglas, 2015), equivalently, there are infinitely many sub-Q-matrices satisfying Conditions II.1, II.2, and II.3.

When the identifiability conditions are not satisfied, we may expect to obtain partial identification results that certain parameters are identifiable while others are only identifiable up to some transformations. For instance, when Condition II.1 is satisfied, the slipping parameters are all identifiable and guessing parameters of items $(K + 1, \ldots, J)$ are also identifiable. It is also possible in practice that there exist certain hierarchical structures among the latent attributes. For instance, an attribute may be a prerequisite for some other attributes. In this case, some entries of $\boldsymbol{p}$ are restricted to be 0. It would also be interesting to consider the identifiability conditions under these restricted models. For these cases, weaker conditions are expected for identifiability of the model parameters. In particular, completeness of the Q-matrix may not be needed. Indeed, these problems are pursued in the following chapters

CHAPTER III

Partial Identifiability of Structured Latent Attribute Models

The necessary and sufficient conditions in the previous Chapter II sometimes can be hard to satisfy in practice, especially the Condition II.1 about the existence of an identity submatrix I_K in the Q-matrix. For many popular designed Q-matrices including the two from the TOEFL iBT tests in Example I.5, the Q-matrix from the Trends in International Mathematics and Science Study in Example I.6, and the Q-matrix from the fraction subtraction data in Example I.7, there does not exist an identity submatrix in the Q and whether the models are identifiable remain open problems. To address these questions, this chapter develops practical identifiability theory for a general family of SLAMs including both DINA and other more complicated models, motivated by real-world needs of designing cognitive diagnostic tests with minimal restrictions.

As introduced in Chapter I, a SLAM is also a restricted latent class model, where the Q-matrix imposes restrictions on the parameter space of a latent class model. So from now on, we call the DINA and the DINO models the *two-parameter Q-restricted latent class models*, since each item has exactly two item parameters, and we call the

This chapter consists of the main part of Gu and Xu (2020a), accpted by *Annals of Statistics*.

main-effect and all-effect models as *multiparameter Q-restricted latent class models*. In this chapter, we will use the term structured latent attribute model and the term restricted latent class model interchangeably.

We now restate the definition of the constraint set $\mathcal{C}_j$ for each item j, as mentioned earlier in Chapter I. For any item j, there exists an item-specific set of latent classes $\mathcal{C}_j$; and the classes in $\mathcal{C}_j$ share the same value of positive response probability (i.e., item parameter $\theta_{j,\boldsymbol{\alpha}}$), which is higher than those of the other latent classes. In other words, the set $\mathcal{C}_j$ also has the following equivalent definition,

$$\mathcal{C}_j = \Big\{\boldsymbol{\alpha} \in \mathcal{A} :\ \theta_{j,\boldsymbol{\alpha}} = \max_{\boldsymbol{\alpha}^\star \in \mathcal{A}} \theta_{j,\boldsymbol{\alpha}^\star}\Big\}. \tag{3.1}$$

The latent classes in $\mathcal{C}_j$ then correspond to those subjects who are "most capable" of giving a positive response to item j, and for each $j \in \mathcal{S}$,

$$\max_{\boldsymbol{\alpha}\in\mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}\in\mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'}, \quad \forall \boldsymbol{\alpha}' \notin \mathcal{C}_j. \tag{3.2}$$

Additionally, it is assumed that there exists a universal "least capable" class $\boldsymbol{\alpha}_0$ such that $\theta_{j,\boldsymbol{\alpha}} \geq \theta_{j,\boldsymbol{\alpha}_0}$ for any $\boldsymbol{\alpha} \in \mathcal{A}$ and $j \in \mathcal{S}$. Note that a latent class $\boldsymbol{\alpha}'$ satisfying $\boldsymbol{\alpha}' \notin \mathcal{C}_j$ and $\theta_{j,\boldsymbol{\alpha}'} > \theta_{j,\boldsymbol{\alpha}_0}$ can be viewed as "partially capable".

An attribute profile $\boldsymbol{\alpha}$ also represents a latent class. Without loss of generality, assume there are m latent classes existing in the population denoted by $\mathcal{A} = \{\boldsymbol{\alpha}_0, \dots, \boldsymbol{\alpha}_{m-1}\}$, where $m > 1$ is assumed known in this chapter. For any $\boldsymbol{\alpha} \in \mathcal{A}$, $p_{\boldsymbol{\alpha}} = P(\mathbf{A} = \boldsymbol{\alpha})$ still denotes the proportion of subjects in the population that belong to class $\boldsymbol{\alpha}$. Under this specification, we have $p_{\boldsymbol{\alpha}} \in (0,1)$ and $\sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} = 1$. Specifically in a SLAM with K binary latent attributes, $\mathcal{A} = \{\boldsymbol{\alpha} \in \{0,1\}^K : p_{\boldsymbol{\alpha}} > 0\}$. So the latent pattern space $\mathcal{A}$ is a subset of $\{0,1\}^K$. If $\mathcal{A} = \{0,1\}^K$, we say $\mathcal{A}$ is *saturated*, which means the population contain subjects with all the possible configurations of attribute profiles. The universal least capable latent pattern $\boldsymbol{\alpha}_0$ corresponds to the all-zero attribute profile, that is, $\boldsymbol{\alpha}_0 = (0, \dots, 0)$.

When the latent pattern space $\mathcal{A}$ is saturated with $\mathcal{A} = \{0,1\}^K$, we have $m = |\mathcal{A}| = 2^K$. In practice, however, this may not always hold. For instance, researchers may assume there exist additional restrictions on the dependence structure among the latent attributes, such as an attribute hierarchy with some attributes being the prerequisite for some others (Leighton et al., 2004; Templin and Bradshaw, 2014). A hierarchical structure among the K attributes would reduce the number of possible attribute profiles from 2^K to m ($m < 2^K$), by excluding those not respecting the hierarchy. For example, consider a diagnostic test with $K = 2$ attributes. If it is scientifically reasonable to assume the first attribute is the prerequisite for the second one, then the latent pattern space is reduced to $\mathcal{A} = \{(0,0), (1,0), (1,1)\}$ with $m = |\mathcal{A}| = 3$, since the attribute profile $(0,1)$ does not respect this hierarchy. Note that as shown in (von Davier and Haberman, 2014), a cognitive diagnosis model with such a linear hierarchy can equivalently reduce to a located latent class model with $m < 2^K$ classes.

In this chapter, we assume the latent pattern set $\mathcal{A}$ is prespecified and known. This would be the case when practitioners have solid scientific reasons or prior knowledge from exploratory data analysis to assume certain structure among attributes. This chapter aims to answer the question that for an arbitrary $\mathcal{A} \subseteq \{0,1\}^K$, what kind of conditions would guarantee identifiability of $\boldsymbol{\Theta}$ and $\boldsymbol{p} = (p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A})$. Later in Chapter V, the latent pattern space $\mathcal{A}$ will not be assumed known and instead will be learned from the data with its own identifiability guarantees there.

This chapter proposes a general framework of strict and partial identifiability for restricted latent class models. Practical sufficient conditions for strict and partial identifiability are proposed and their necessity is discussed. In particular, depending on the two different types of algebraic structures of restricted latent class models, we introduce and study two useful notions of partial identifiability, respectively (see

Sections 3.2 and 3.3). The established identifiability results are widely applicable in practice, by relaxing most of the constraints imposed on the design matrix. Moreover, under correct model specification, all the identifiability conditions only depend on the design matrix and are easily checkable by practitioners. We apply the new theory to several existing designs and establish identifiability under them for the first time in the literature.

The rest of this chapter is organized as follows. Section 3.1 summarizes the issues with existing works on identifiability and discusses the open problems. Sections 3.2 and 3.3 present our main identifiability results. Section 3.4 includes extensions of the new theory to some more complicated models. Section 3.5 gives a further discussion, and proofs of the theoretical results are presented in the Appendix B.

3.1 Issues with Existing Works and Open Problems

Though widely used in various applications, the identifiability issue of SLAMs or restricted latent class models remains largely unaddressed. We next introduce the concept of identifiability and discuss the limitations of the exiting theory.

For a SLAM introduced in Chapter I, we restate the probability mass function of the response pattern $\boldsymbol{R}$:

$$P(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{r_j} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-r_j}, \quad \boldsymbol{r} \in \{0,1\}^J. \tag{3.3}$$

Following the definition of identifiablity in the literature (e.g., Casella and Berger, 2002), the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ of a SLAM are identifiable if for any $(\boldsymbol{\Theta}, \boldsymbol{p})$ in the parameter space $\mathcal{T}$, there is no $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \neq (\boldsymbol{\Theta}, \boldsymbol{p})$ such that

$$\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \text{ for all } \boldsymbol{r} \in \{0,1\}^J. \tag{3.4}$$

In the following, we also say that the model parameters are *strictly* identifiable if the above condition holds.

To establish model identifiability, a strong and often impractical assumption made by previous works is that the Q-matrix must contain at least one $K \times K$ identity submatrix I_K up to some row permutation, that is, the Q-matrix must contain all K distinct single-attribute $\boldsymbol{q}$-vectors (Chen et al., 2015; Xu and Zhang, 2016; Xu, 2017; Gu and Xu, 2019b). A Q-matrix satisfying this requirement is also said to be complete under the DINA model (Chiu et al., 2009). For general Q-restricted latent class models including the multiparameter models, Xu (2017) requires at least two disjoint $K \times K$ identity submatrices in Q to establish identifiability. However, in practice, in the existence of a large number of fine-grained attributes and complex cognitive process, a Q-matrix rarely satisfies such requirements. For the TOEFL data in Example I.5, in both Q-matrices, there does not exist any item that solely requires the fourth skill attribute. For the Q-matrix of the TIMSS data in Example I.6, only three attributes (1, 7 and 8) out of twelve are measured by some single-attribute items. For the Q-matrix in Example I.7, there are only two attributes (2 and 7) out of eight measured by some single-attribute items. Many other examples can be found in the literature (e.g., Jaeger et al., 2006; Henson et al., 2009; de la Torre, 2011; Lee et al., 2011). Moreover, another strong assumption made in existing works Xu (2017); Gu and Xu (2019b) is that $\mathcal{A} = \{0,1\}^K$, that is, $p_{\boldsymbol{\alpha}} > 0$ for any $\boldsymbol{\alpha} \in \{0,1\}^K$, which fails when some attribute profiles are deemed impossible to exist.

Such identifiability issues of cognitive diagnosis models have long been recognized (de la Torre and Douglas, 2004b; von Davier, 2008; Tatsuoka, 2009; DeCarlo, 2011; Maris and Bechger, 2009; Zhang et al., 2013; von Davier, 2014). For instance, von Davier (2008) pointed out in the study of the TOEFL data that larger numbers of skills (i.e., K) very likely pose problems with identifiability, unless the number of items per skill is "sufficiently" large. But given the complicated structure of constraints, how the number of items and the form of the design matrix influence identifiability is still an open problem in the literature.

This chapter addresses this open problem by developing a general theoretical framework based on a key technical tool, the *indicator matrix* Γ. Under an arbitrary SLAM, we define Γ to be a $J \times m$ matrix using the sets $\mathcal{C}_j$'s. The Γ-matrix has the same size as the matrix $\mathbf{\Theta}$, with rows indexed by items in $\mathcal{S}$, and columns by latent classes in $\mathcal{A}$. The $(j, \boldsymbol{\alpha})$th entry of Γ is

$$\Gamma_{j,\boldsymbol{\alpha}} = I(\boldsymbol{\alpha} \in \mathcal{C}_j), \quad j \in \mathcal{S},\ \boldsymbol{\alpha} \in \mathcal{A}, \tag{3.5}$$

which is a binary indicator of whether $\boldsymbol{\alpha}$ is "most capable" to give a positive response to j. For $\boldsymbol{\alpha} \in \mathcal{A}$, denote the $\boldsymbol{\alpha}$th column vector of Γ by $\Gamma_{\cdot,\boldsymbol{\alpha}}$. The Γ-matrix defined this way turns out to be a useful tool for developing the identifiability theory, and it helps to relax many of the existing strong assumptions, as shown later in Sections 3.2.1 and 3.3.1. Indeed, most of our identifiability conditions can be represented as requirements on the structure of Γ, since the information of which latent classes achieve the highest level of $\theta_{j,\boldsymbol{\alpha}}$ of item j is what our theoretical derivations essentially rely on.

The DINA and DINO models are restricted latent class models with appropriately defined constraint sets $\mathcal{C}_j$'s. Specifically, under the conjunctive DINA model, the $\mathcal{C}_j$ defined in (3.1) takes the form of

$$\mathcal{C}_j^{DINA} = \{\boldsymbol{\alpha} \in \mathcal{A} :\ \boldsymbol{\alpha} \succeq \boldsymbol{q}_j\}, \quad j \in \mathcal{S}; \tag{3.6}$$

while under the disjunctive DINO model, the $\mathcal{C}_j$ defined in (3.1) becomes $\mathcal{C}_j^{DINO} = \{\boldsymbol{\alpha} \in \mathcal{A} :\ \text{if } \exists k \text{ s.t. } \alpha_k = q_{j,k} = 1\}$ for $j \in \mathcal{S}$.

Depending on two different algebraic structures of the constrained parameter spaces, we next consider two types of restricted latent class models and present their identifiability results in Sections 3.2 and 3.3, respectively.

3.2 Identifiability Results for Two-Parameter Models

This section considers two-parameter restricted latent class models where each item j has two item parameters, that is, $|\{\theta_{j,\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}\}| = 2$. Specifically, a two-parameter model assumes that for each item j, the latent classes in $\mathcal{C}_j$ share a same positive response probability, denoted by θ_j^+, and the latent classes in the complement set $\mathcal{A} \setminus \mathcal{C}_j$ share another same positive response probability, denoted by θ_j^-. We assume $\theta_j^+ > \theta_j^-$. Note that the unique item parameters in $\boldsymbol{\Theta}$ reduce to $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$, where $\boldsymbol{\theta}^+ = (\theta_1^+, \ldots, \theta_J^+)^\top$ and $\boldsymbol{\theta}^- = (\theta_1^-, \ldots, \theta_J^-)^\top$. The motivation for studying the two-parameter models comes from the popular DINA and DINO models in cognitive diagnosis, as introduced in Example I.1. Moreover, the study of the two-parameter models provides insight into understanding other restricted latent class models, as they serve as submodels for many multiparameter models.

Under a two-parameter model, the Γ-matrix fully captures the model structure, in the sense that $\theta_{j,\boldsymbol{\alpha}} = \theta_j^+$ if $\Gamma_{j,\boldsymbol{\alpha}} = 1$ and $\theta_{j,\boldsymbol{\alpha}} = \theta_j^-$ if $\Gamma_{j,\boldsymbol{\alpha}} = 0$. So in this scenario, if Γ contains two identical columns, then the corresponding latent classes have the same item parameters across all items. Namely, if $\Gamma_{\bullet,\boldsymbol{\alpha}} = \Gamma_{\bullet,\boldsymbol{\alpha}'}$, then $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}} = \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}'}$. Thus from an identifiability perspective, these two latent classes are equivalent and cannot be distinguished based on their observed responses. This implies that in order to distinguish the latent classes, it is necessary that each latent class in $\mathcal{A}$ should correspond to a distinct column vector of Γ. We shall call such a Γ-matrix separable.

Definition III.1. A Γ-matrix is said to be separable, if any two column vectors of Γ are distinct. Otherwise, we say Γ is inseparable.

To see how the separability of the Γ-matrix influences model identifiability, we start with an ideal case with all the item parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ known. The following proposition characterizes the importance of Γ's separability.

Proposition III.1. Consider a two-parameter restricted latent class model with known $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$. Then the proportion parameters $\boldsymbol{p}$ are identifiable if and only if the Γ-matrix is separable.

We use the following example as an illustration.

Example III.1. Consider the Q-matrix in (3.7) with $K = 2$ attributes. Under the DINA model with $\mathcal{C}_j$ in the form of (3.6), if $\mathcal{A} = \{0,1\}^2 = \{\boldsymbol{\alpha}_0 = (0,0),\ \boldsymbol{\alpha}_1 = (1,0),\ \boldsymbol{\alpha}_2 = (0,1),\ \boldsymbol{\alpha}_3 = (1,1)\}$, then $\Gamma^{(1)}$ in (3.7) represents the corresponding Γ-matrix, which is inseparable. Specifically, we can see that $\Gamma_{\bullet,\boldsymbol{\alpha}_0} = \Gamma_{\bullet,\boldsymbol{\alpha}_2}$ and the two classes $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_2$ have the same item parameters, $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}_0} = \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}_2} = \boldsymbol{\theta}^-$. Thus $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_2$ are not distinguishable and equivalently, their proportion parameters $p_{\boldsymbol{\alpha}_0}$ and $p_{\boldsymbol{\alpha}_2}$ are not identifiable.

$$
Q = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}
\begin{array}{l}
\xRightarrow[\mathcal{A}=\{0,1\}^2]{DINA} \quad \Gamma^{(1)} = \begin{array}{cccc} \boldsymbol{\alpha}_0 & \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3 \\ \end{array} \\
\qquad\qquad\qquad\quad \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}; \\
\xRightarrow[\mathcal{A}=\{0,1\}^2\setminus\{0,1\}]{DINA} \quad \Gamma^{(2)} = \begin{array}{ccc} \boldsymbol{\alpha}_0 & \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_3 \\ \end{array} \\
\qquad\qquad\qquad\quad \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.
\end{array}
\tag{3.7}
$$

On the other hand, if prior knowledge shows that the first attribute is the prerequisite for the second, then $\mathcal{A}$ reduces to $\{0,1\}^2 \setminus \{(0,1)\}$ and the Γ-matrix becomes $\Gamma^{(2)}$ in (3.7). The $\Gamma^{(2)}$ is separable, with each $\boldsymbol{\alpha}$ having a distinct column vector in Γ and $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}_0} \neq \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}_1} \neq \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}_3}$. Therefore Proposition III.1 gives that $\boldsymbol{p}$ is identifiable in the ideal case with known $\boldsymbol{\Theta}$.

An inseparable Γ-matrix violates the necessary condition for identifying $\boldsymbol{p}$ under the two-parameter models. To study the "partial" identifiability of $\boldsymbol{p}$ when Γ is inseparable, we next define an equivalence relation "$\sim$" of latent classes induced by the column vectors of Γ. Specifically, we define $\boldsymbol{\alpha} \sim \boldsymbol{\alpha}'$ if and only if $\Gamma_{\bullet,\boldsymbol{\alpha}} = \Gamma_{\bullet,\boldsymbol{\alpha}'}$. Let C be the number of distinct column vectors of Γ and $\mathcal{A}_1, \ldots, \mathcal{A}_C$ be the C equivalence classes under $\sim$. Let $\boldsymbol{\alpha}_{\mathcal{A}_i}$ be a representative of $\mathcal{A}_i$ and we write $[\boldsymbol{\alpha}_{\mathcal{A}_i}] = \mathcal{A}_i$. We define the grouped population proportion parameters to be

$$\nu_{[\boldsymbol{\alpha}_{\mathcal{A}_i}]} := \sum_{\boldsymbol{\alpha}\colon \boldsymbol{\alpha}\in\mathcal{A}_i} p_{\boldsymbol{\alpha}}, \qquad \text{for } i = 1, \ldots, C, \tag{3.8}$$

and write $\boldsymbol{\nu} = (\nu_{[\boldsymbol{\alpha}_{\mathcal{A}_1}]}, \ldots, \nu_{[\boldsymbol{\alpha}_{\mathcal{A}_C}]})^\top$. When Γ is separable, we have $C = m$, $\boldsymbol{\nu} = \boldsymbol{p}$ and each $\boldsymbol{\alpha}$ represents a unique equivalence class.

The following result shows that under an inseparable Γ-matrix, though $\boldsymbol{p}$ are not identifiable, the parameters $\boldsymbol{\nu}$ are identifiable.

Proposition III.2. Consider a two-parameter model with known $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$. When the Γ-matrix is inseparable, $\boldsymbol{\nu}$ is identifiable. Moreover, the latent classes in the same equivalence class cannot be distinguished in the sense that for any model parameters $\boldsymbol{p} \neq \bar{\boldsymbol{p}}$, if $\nu_{[\boldsymbol{\alpha}_{\mathcal{A}_i}]} = \bar{\nu}_{[\boldsymbol{\alpha}_{\mathcal{A}_i}]}$, where $\bar{\nu}_{[\boldsymbol{\alpha}_{\mathcal{A}_i}]} = \sum_{\boldsymbol{\alpha}\colon \boldsymbol{\alpha}\in\mathcal{A}_i} \bar{p}_{\boldsymbol{\alpha}}$ for $i = 1, \ldots, C$, then $\mathbb{P}(\boldsymbol{R} \mid \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} \mid \boldsymbol{\Theta}, \bar{\boldsymbol{p}})$.

When Γ is inseparable, Proposition III.2 implies that even in the ideal case with known $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$, the identification of $\boldsymbol{\nu}$ is the strongest identifiability result one can obtain for two-parameter restricted latent class models. This therefore motivates us to introduce the following definition of the $\boldsymbol{p}$-partial identifiability when both $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ and $\boldsymbol{p}$ are unknown.

Definition III.2 ($\boldsymbol{p}$-partial identifiability)**.** For a two-parameter restricted latent class model with a given Γ-matrix, the model parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ are said to be $\boldsymbol{p}$-partially identifiable if $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$ are identifiable.

We point out that when the Γ-matrix is separable, the $\boldsymbol{p}$-partial identifiability exactly becomes the strict identifiability. When Γ is inseparable, the definition of

$\boldsymbol{p}$-partial identifiability here refers to partially identifying the proportion parameters $\boldsymbol{p}$, and strictly identifying all the item parameters. Such definition suits for the needs of cognitive diagnosis applications, by ensuring the identification of the equivalent attribute profiles of interest, and also ensuring the estimability of all item parameters so that the quality of the items can be accurately evaluated and validated.

In the framework of $\boldsymbol{p}$-partial identifiability, the following Section 3.2.1 presents a general identifiability result, allowing $\mathcal{A}$ to be arbitrary and Γ to be inseparable. Section 3.2.2 further focuses on the family of Q-restricted latent class models and discusses the necessity of the proposed conditions. Section 3.2.3 includes the applications of the new theory.

Remark III.1. For the family of two-parameter Q-restricted latent class models, the Γ-induced equivalence classes can be obtained as follows. We define two sets of attribute profiles under the conjunctive DINA and disjunctive DINO assumptions, respectively:

$$\mathcal{R}^{Q,conj} = \{\boldsymbol{\alpha} = \vee_{h\in S}\, \boldsymbol{q}_h : S \subseteq \mathcal{S}\}, \quad \mathcal{R}^{Q,disj} = \{\mathbf{1} - \boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathcal{R}^{Q,conj}\}, \tag{3.9}$$

where $\vee_{h\in S}\, \boldsymbol{q}_h = (\max_{h\in S}\{q_{h,1}\}, \ldots, \max_{h\in S}\{q_{h,K}\})$, and $\vee_{h\in\varnothing}\, \boldsymbol{q}_h$ is defined to be the all-zero vector. We claim that when $\mathcal{A} = \{0,1\}^K$, the $\mathcal{R}^{Q,conj}$ or $\mathcal{R}^{Q,disj}$ is a complete set of representatives of the conjunctive or disjunctive equivalence classes, respectively; the proof of this result is given in Section B of the Supplementary Material. Moreover, for any latent class space $\mathcal{A} \subseteq \{0,1\}^K$, define a map $f(\boldsymbol{\cdot}) : \mathcal{A} \to \mathcal{R}^{Q,conj}$ (or $R^{Q,disj}$) which sends each attribute pattern $\boldsymbol{\alpha} \in \mathcal{A}$ to the element in $\mathcal{R}^{Q,conj}$ (or $R^{Q,disj}$) equivalent to $\boldsymbol{\alpha}$. Then $f(\mathcal{A})$ forms a complete set of conjunctive or disjunctive representatives. A similar grouping operation in the saturated and conjunctive case was introduced in Zhang et al. (2013). Consider Example III.1 for an illustration. If $\mathcal{A} = \{0,1\}^2$, $\Gamma^{(1)}$ is inseparable. The equivalence class representatives are $\mathcal{R}^{Q,conj} = \{(0,0),(1,0),(1,1)\}$ by (3.9) and $\boldsymbol{\nu} = (\nu_{[0,0]}, \nu_{[1,0]}, \nu_{[1,1]})$ with $\nu_{[0,0]} =$

$p_{(0,0)} + p_{(0,1)}$, $\nu_{[1,0]} = p_{(1,0)}$, $\nu_{[1,1]} = p_{(1,1)}$. On the other hand, $\Gamma^{(2)}$ is separable with latent class space $\mathcal{A} = R^{Q,conj}$. This also illustrates that a separable Γ-matrix does not necessarily correspond to a Q-matrix containing an identity submatrix I_K. Therefore, compared with existing theory, the Γ-matrix provides a more suitable tool than the Q-matrix for studying identifiability of Q-restricted models.

3.2.1 Strict and Partial Identifiability

This subsection presents conditions depending on the Γ-matrix that lead to the $\boldsymbol{p}$-partial identifiability of a two-parameter restricted latent class model. We first introduce some notation. Based on the constraint sets $\mathcal{C}_j$'s, we categorize the entire set of items $\mathcal{S} = \{1, \ldots, J\}$ into two subsets, the set of *nonbasis* items S_{non} and that of *basis* items S_{basis} as follows,

$$S_{non} = \{j : \exists h \in \mathcal{S} \setminus \{j\}, \text{ s.t. } \mathcal{C}_h \supseteq \mathcal{C}_j\} \quad \text{and} \quad S_{basis} = \mathcal{S} \setminus S_{non}. \tag{3.10}$$

By this definition, an item j is a nonbasis item if the capability of item j implies capability of some other item, and a basis item otherwise. With a slight abuse of notation, for any subset of items $S \subseteq \mathcal{S}$, denote $\mathcal{C}_S = \cap_{j \in S} \mathcal{C}_j$. We introduce the next definition of *S-differentiable* to describe the relation between an item and a set of items.

Definition III.3. For an item j and a set of items S that does not contain j, item j is said to be *S-differentiable* if there exist two subsets S_j^+, S_j^- of S, which are not necessarily nonempty or disjoint, such that

$$\mathcal{C}_{S_j^+} \supsetneqq \mathcal{C}_{S_j^-} \quad \text{and} \quad \mathcal{C}_{S_j^-} \setminus \mathcal{C}_{S_j^+} \subseteq \mathcal{A} \setminus \mathcal{C}_j. \tag{3.11}$$

When j is S-differentiable, the set S is said to be a *separator set* of item j. An item j is S-differentiable indicates that the items in the separator set S can differentiate at least one incapable latent class of j (i.e., one latent class in $\mathcal{A}\backslash\mathcal{C}_j$) from the universal least capable class $\boldsymbol{\alpha}_0$.

We need the following two conditions to establish identifiability.

(C1) *Repeated measurement condition*: For each item j, there exist two disjoint sets of items S_j^1, $S_j^2 \subset \mathcal{S}\backslash\{j\}$ such that $\mathcal{C}_j \supseteq \mathcal{C}_{S_j^1}$ and $\mathcal{C}_j \supseteq \mathcal{C}_{S_j^2}$.

(C2) *Sequentially differentiable condition*: Start with the set $S_{sep} = S_{non}$. Expand S_{sep} by including all items in $\mathcal{S} \setminus S_{sep}$ that are S_{sep}-differentiable, and repeat the expanding procedure until no items can be added to S_{sep}. The sequentially expanding procedure ends up with $S_{sep} = \mathcal{S}$.

Before presenting the formal theorem, we first give a simple illustration of how condition (C2) can be checked.

Example III.2. Consider the following 3×4 Γ-matrix,

$$\Gamma = \begin{array}{c} \begin{array}{cccc} \boldsymbol{\alpha}_0 & \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array},$$

then $\mathcal{C}_1 = \{\boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3\}$, $\mathcal{C}_2 = \{\boldsymbol{\alpha}_3\}$ and $\mathcal{C}_3 = \{\boldsymbol{\alpha}_1\}$. By (3.10), $S_{non} = \{2, 3\}$ and $S_{basis} = \{1\}$. To check condition (C2), we start with the separator set $S_{sep} = S_{non} = \{2, 3\}$. For basis item 1, we define $S_1^+ = \varnothing$ and $S_1^- = \{3\}$. Then $\mathcal{C}_{S_1^+} = \{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3\}$ and $\mathcal{C}_{S_1^-} = \{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3\}$, so $\mathcal{C}_{S_1^+} \setminus \mathcal{C}_{S_1^-} = \{\boldsymbol{\alpha}_1\} \subseteq \mathcal{C}_1^c = \{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1\}$, which means (3.11) holds for $j = 1$. Besides, $S_1^+ \cup S_1^- \subseteq S_{non}$. So by Definition III.3, item 1 is S_{non}-differentiable. Now we can expand the separator set S_{sep} to be $S_{non} \cup \{1\} = \mathcal{S}$. So the sequentially expanding procedure described in condition (C2) ends in one step with $S_{sep} = \mathcal{S}$, and (C2) is satisfied.

Theorem III.1. *Under the two-parameter restricted latent class models, condition (C1) is sufficient for identifiability of* $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-_{non})$*, where* $\boldsymbol{\theta}^-_{non} = (\theta_j^-,\ j \in S_{non})$*. Moreover, conditions (C1) and (C2) are sufficient for* $\boldsymbol{p}$*-partial identifiability of the model parameters* $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$*.*

Theorem III.1 presents a general identifiability result with strict identifiability being a special case. For instance, in the case of $\mathcal{A} = \{0,1\}^K$, if the $J \times 2^K$ Γ-matrix is separable, then $\boldsymbol{\nu} = \boldsymbol{p}$ and the $\boldsymbol{p}$-partial identifiability in Theorem III.1 exactly ensures strict identifiability of all the parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$. Similarly, in the case of $\mathcal{A} \subset \{0,1\}^K$, if the $J \times |\mathcal{A}|$ Γ-matrix is separable, the $\boldsymbol{p}$-partial identifiability ensures $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ and $(p_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha} \in \mathcal{A})$ are strictly identifiable. Conditions (C1) and (C2) only depend on the structure of the Γ-matrix and are easily checkable. Condition (C1) implies that at least one capable class of each item is repeatedly measured by other items. Condition (C2) requires that for each basis item, at least one of its incapable classes should be differentiated from the universal least capable class through a sequential procedure. From the proof of Theorem III.1, (C1) suffices for identifiability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-_{non})$; furthermore, the sequential procedure in condition (C2) ensures that as S_{sep} sequentially expands its size, for any item h included in S_{sep}, the parameter θ_h^- is identifiable. If (C2) holds, that is, the sequential procedure ends up with $S_{sep} = \mathcal{S}$, we have the entire $\boldsymbol{\theta}^-$ identifiable, which further leads to identifiability of $\boldsymbol{\nu}$. The sequential statement of (C2) accurately characterizes the underlying structure of the Γ-matrix needed for identifiability. In particular, if there are no basis items, that is, $\mathcal{S} = S_{non}$, then (C2) automatically holds with zero expanding step; while if there do exist basis items and each basis item is S_{non}-differentiable, then (C2) holds with one expanding step.

The next proposition further extends the result in Theorem III.1 to the case where the Γ-matrix may not satisfy (C1) and (C2). For any subset of items $S \subseteq \mathcal{S}$, define the *S-adjusted Γ-matrix* $\Gamma(S)$ as follows, which has the same size as the original Γ.

Its jth row $\{\Gamma(S)\}_{j,\bullet}$ equals $\mathbf{1}_m^\top - \Gamma_{j,\bullet}$ if $j \in S$, and equals $\Gamma_{j,\bullet}$ if $j \notin S$. Here $\mathbf{1}_m^\top$ denotes an all-one row vector of length m.

Proposition III.3. Consider a two-parameter restricted latent class model associated with a Γ-matrix. If there exist a subset of items $S \subseteq \mathcal{S}$ such that the S-adjusted Γ-matrix $\Gamma(S)$ satisfies conditions (C1) and (C2), then the two-parameter model is $\boldsymbol{p}$-partially identifiable.

Proposition III.3 relaxes the conditions of Theorem III.1, by only requiring that (C1) and (C2) can be satisfied after switching the zeros and ones for some rows of in the Γ. The identifiability conditions in Theorem III.1 and Proposition III.3 allow for a nonsaturated latent class space $\mathcal{A}$ and inseparability of the Γ-matrix, which relaxes the existing identifiability conditions in the literature. Moreover, the proposed conditions (C1) and (C2) would become necessary and sufficient in certain scenarios to be discussed in the following subsection.

3.2.2 Results for Q-restricted Latent Class Models

To further illustrate the result in Theorem III.1, we focus on the two-parameter Q-restricted latent class model with a saturated latent class space $\mathcal{A} = \{0,1\}^K$. This includes the conjunctive DINA and disjunctive DINO models in Example I.1 as special cases. Without loss of generality, we next only consider the two-parameter conjunctive model. Nevertheless, all the $\boldsymbol{p}$-partial identifiability results presented in this subsection hold for both the conjunctive and the disjunctive models, due to the duality between them (Chen et al., 2015).

We introduce the following definitions adapted from Section 3.2.1. Under the conjunctive model assumption with $\mathcal{C}_j$ taking the form of (3.6), the non-basis and basis items defined earlier in (3.10) can be equivalently expressed in terms of the

$\boldsymbol{q}$-vectors as follows

$$S_{non} = \{j : \exists h \in \mathcal{S} \setminus \{j\} \text{ s.t. } \boldsymbol{q}_h \preceq \boldsymbol{q}_j\} \text{ and } S_{basis} = \mathcal{S} \setminus S_{non}. \tag{3.12}$$

Moreover, item j is set S-differentiable if there exist S^+, $S^- \subseteq S$ such that

$$\mathbf{0} \precneqq \vee_{h \in S^+} \boldsymbol{q}_h - \vee_{h \in S^-} \boldsymbol{q}_h \preceq \boldsymbol{q}_j. \tag{3.13}$$

In addition, conditions (C1) and (C2) are equivalent to:

(C1*) *Repeated measurement condition*: For each $j \in \mathcal{S}$, there exist two disjoint item sets S_j^1, $S_j^2 \subseteq \mathcal{S} \setminus \{j\}$ such that $\boldsymbol{q}_j \preceq \vee_{h \in S_j^1} \boldsymbol{q}_h$ and $\boldsymbol{q}_j \preceq \vee_{h \in S_j^2} \boldsymbol{q}_h$.

(C2*) *Sequentially differentiable condition*: The same as condition (C2), but using definition (3.13) of S-differentiable regarding the $\boldsymbol{q}$-vectors.

Following Theorem III.1, the next corollary shows that the derived conditions on the Q-matrix suffice for the $\boldsymbol{p}$-partial identifiability of both the conjunctive and disjunctive two-parameter models.

Corollary III.1. *Under the two-parameter Q-restricted latent class models, assuming $\nu_{[\boldsymbol{\alpha}]} > 0$ for any equivalence class $[\boldsymbol{\alpha}]$, (C1*) and (C2*) are sufficient for the $\boldsymbol{p}$-partial identifiability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$.*

We use the following example as an illustration of the identifiability result; see also real data examples in Section 3.2.3.

Example III.3. Under the DINA model, consider the following Q-matrix.

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hdashline 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \tag{3.14}$$

This Q-matrix lacks the single-attribute item $(0, 0, 1)$, and the corresponding Γ-matrix under $\mathcal{A} = \{0, 1\}^3$ is inseparable. In this case, we have the following 7 equivalence classes $\{[0, 0, 0],\ [1, 0, 0],\ [0, 1, 0],\ [1, 1, 0],\ [0, 1, 1],\ [1, 0, 1],\ [1, 1, 1]\}$, with the equivalence class $[0, 0, 0]$ containing attribute profiles (0, 0, 0) and (0, 0, 1), while each of the other equivalence classes contains one attribute profile. Following the definition in (3.12), items 1 and 2 are basis items, and items 3, 4 and 5 are non-basis items. For all the five items, condition (C1*) is satisfied by taking $(S_1^1,\ S_1^2) = (\{3\},\ \{5\})$, $(S_2^1,\ S_2^2) = (\{3\},\ \{4\})$, $(S_3^1,\ S_3^2) = (\{1, 4\},\ \{2, 5\})$, $(S_4^1,\ S_4^2) = (\{3\},\ \{2, 5\})$, and $(S_5^1,\ S_5^2) = (\{3\},\ \{1, 4\})$. In addition, condition (C2*) is also satisfied since the basis items 1 and 2 are $(S_1^+ \cup S_1^-)$- and $(S_2^+ \cup S_2^-)$-differentiable, respectively, where $(S_1^+, S_1^-) = (\{3\}, \{4\})$ and $(S_2^+, S_2^-) = (\{3\}, \{5\})$. By Corollary III.1, the DINA model parameters are $\boldsymbol{p}$-partially identifiable.

As shown above, conditions (C1*) and (C2*) are sufficient conditions to ensure $\boldsymbol{p}$-partial identifiability. In the following, we discuss the necessity of (C1*) and (C2*) and further provide procedures to establish identifiability in certain cases when these conditions fail to hold.

For a general Q-matrix, condition (C1*) implies that each attribute is required by at least three items. In the next theorem, we show that it is necessary for each attribute to be required by at least two items; in particular, if some attribute is required by only two items, the identifiability conclusion would depend on the structure of the $\boldsymbol{q}$-vectors of those two items.

Theorem III.2 (Discussion of C1*)**.** *Consider a two-parameter Q-restricted latent class model.*

(a) If some attribute is required by only one item, then the model is not $\boldsymbol{p}$-partially identifiable.

(b) If some attribute is required by only two items, without loss of generality, suppose the first attribute is required by the first two items and the Q-matrix takes the following form

$$Q = \begin{pmatrix} 1 & \boldsymbol{v}_1^\top \\ 1 & \boldsymbol{v}_2^\top \\ \mathbf{0} & Q' \end{pmatrix}_{J\times K}, \tag{3.15}$$

where Q' is a $(J-2)\times(K-1)$ sub-matrix of Q and $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ are $(K-1)$-dimensional vectors.

(b.1) If $\boldsymbol{v}_1 = \mathbf{0}$ or $\boldsymbol{v}_2 = \mathbf{0}$, the model is not $\boldsymbol{p}$-partially identifiable.

(b.2) If $\boldsymbol{v}_1 \neq \mathbf{0}$ and $\boldsymbol{v}_2 \neq \mathbf{0}$, the model is $\boldsymbol{p}$-partially identifiable if the sub-matrix Q' satisfies conditions (C1) and (C2*), and either (a) or (b) below holds for $i = 1$ and 2: (a) There exists some $j \geq 3$ such that $\boldsymbol{q}_{j,2:K} \not\succeq \boldsymbol{v}_i$; (b) There does not exist any $j \geq 3$ such that $\boldsymbol{q}_{j,2:K} \not\succeq \boldsymbol{v}_i$, and among the attributes required by $\boldsymbol{v}_i$, there exists at least one attribute k that is not required by every item $j \in \{3, \ldots, J\}$.*

Theorem III.2 characterizes the different situations when condition (C1*) fails to hold for some attribute, and provides sufficient conditions for identifiability when the Q-matrix falls in the scenario (B). In addition, the result in Theorem III.2 can be easily extended to the case where there are multiple attributes that are required by only two items.

The next theorem discusses the necessity of Condition (C2*) and states that if there exists some basis item that does not have any separator set, then the model parameters are not $\boldsymbol{p}$-partially identifiable.

Theorem III.3 (Discussion of C2*)**.** *Under the two-parameter Q-restricted models, the condition that each basis item j is $(\mathcal{S} \setminus \{j\})$-differentiable, is necessary for the $\boldsymbol{p}$-partial identifiability.*

Furthermore, under the two-parameter Q-restricted models with a separable Γ-matrix and a saturated latent class space $\mathcal{A}$, the following theorem shows conditions (C1*) and (C2*) are exactly the minimal requirement for strict identifiability of the model.

Theorem III.4 (Result on the Necessary and Sufficient Condition)**.** *Under the two-parameter Q-restricted models, if $\mathcal{A}$ is saturated and Γ is separable, then conditions (C1*) and (C2*) are necessary and sufficient for the strict identifiability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$.*

Under the assumptions of Theorem III.4, conditions (C1*) and (C2*) are equivalent to the following explicit conditions on the structure of the Q-matrix: (C1$'$) Each attribute is required by at least three items; (C2$'$) With Q in the form $Q = (I_K^\top, (Q')^\top)^\top$, any two different columns of the submatrix Q' are distinct. Please see the proof of Theorem III.4 for details.

3.2.3 Applications

One important implication of the established identifiability theory is the consistent estimability of the model parameters. Consider a sample of size N and denote the ith subject's multivariate binary responses by $\boldsymbol{R}_i = (R_{i,1}, \ldots, R_{i,J})^\top$. Assume $\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N$ identically and independently follow the categorical distribution with the probability mass function (3.3). The likelihood based on the sample can be written as $L(\boldsymbol{\Theta}, \boldsymbol{p} \mid \boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = \prod_{i=1}^N \mathbb{P}(\boldsymbol{R} = \boldsymbol{R}_i \mid \boldsymbol{\Theta}, \boldsymbol{p})$. We denote the true parameters by $(\mathbf{\Theta}^0, \boldsymbol{p}^0)$ and the maximum likelihood estimators (MLE) by $(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{p}})$, which may not be unique. We further define the corresponding parameters $\boldsymbol{\nu}^0$ and $\widehat{\boldsymbol{\nu}}$ as in (3.8). We have the following conclusion on the estimability of a two-parameter model.

Proposition III.4. If a two-parameter model is $\boldsymbol{p}$-partially identifiable, then $(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{\nu}}) \to (\mathbf{\Theta}^0, \boldsymbol{\nu}^0)$ almost surely as $N \to \infty$. In addition, if Γ-matrix is also separable, then $(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{p}}) \to (\mathbf{\Theta}^0, \boldsymbol{p}^0)$ almost surely. On the other hand, if Γ-matrix is inseparable, $\boldsymbol{p}$ cannot be consistently estimated.

With the consistency result, we can directly establish the asymptotic normality of $(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{\nu}})$ when the model is $\boldsymbol{p}$-partially identifiable, following a standard argument of asymptotic statistics Van der Vaart (2000).

We next apply the newly developed theory to the data examples introduced in Section 1.3, and establish the $\boldsymbol{p}$-partial identifiability of the two-parameter restricted latent class model under the Q-matrices.

For the TOEFL iBT data introduced in Example I.5, the two-parameter restricted latent class models associated with the Q-matrices corresponding to reading forms A and B, denoted by Q_A and Q_B, respectively, are both $\boldsymbol{p}$-partially identifiable. Specifically, under the conjunctive DINA model, the Q_A and Q_B specified in Table 1.1 induce 14 and 12 equivalence classes of attribute profiles respectively, for which the sets of representatives are $\mathcal{R}^{Q_A} = \{0,1\}^4 \backslash \{(0,0,0,1),\ (1,0,0,1)\}$ and $\mathcal{R}^{Q_B} = \{0,1\}^4 \backslash \{(0,0,0,1),\ (1,0,0,1),\ (0,1,0,1),\ (1,1,0,1)\}$. The $\mathcal{R}^{Q_A}$ and $\mathcal{R}^{Q_B}$ are calculated following the procedure introduced in Remark III.1. It is straightforward to check that for both Q_A and Q_B, condition (C1*) holds and there is no basis item, which further implies the satisfaction of condition (C2*). Therefore Corollary III.1 gives the $\boldsymbol{p}$-partial identifiability of the two-parameter models associated with both Q_A and Q_B. Furthermore, Proposition III.4 implies the consistent estimability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$. In particular, the proportion parameters of the equivalence classes $\boldsymbol{\nu} = (\nu_{[\boldsymbol{\alpha}]},\ \boldsymbol{\alpha} \in \mathcal{R}^{Q_A})$ can be consistently estimated, while for those attribute profiles in a same equivalent class, their proportions cannot be consistently estimated. For instance, under Q_A, attribute patterns $\boldsymbol{\alpha}^\star = (0,0,0,1)$ and $\boldsymbol{\alpha}^{\star\star} = (0,0,0,0)$ share the same equivalent class; so $p_{\boldsymbol{\alpha}^\star}$ and $p_{\boldsymbol{\alpha}^{\star\star}}$ are not estimable, and it is only possible and meaningful to estimate $\nu_{[\boldsymbol{\alpha}^\star]} = p_{\boldsymbol{\alpha}^\star} + p_{\boldsymbol{\alpha}^{\star\star}}$.

Other than the TOEFL data, our new results in Section 3.2.2 also guarantee the $\boldsymbol{p}$-partial identifiability of two-parameter models associated with the $Q_{43\times 12}$ for the TIMSS data, and the $Q_{20\times 8}$ for the fraction subtraction data. The details of checking our conditions for $Q_{43\times 12}$ and $Q_{20\times 8}$ are included in Section A of the Supplementary Material.

3.3 Identifiability Results for Multiparameter Models

This section considers multiparameter restricted latent class models where each item j allows for more than two item parameters, i.e., $|\{\theta_{j,\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}\}| \geq 2$. In a multiparameter model, those latent classes in $\mathcal{C}_j$ still have the same level of positive response probability, according to the definition of $\mathcal{C}_j$ in (3.1); however, the classes in $\mathcal{A} \setminus \mathcal{C}_j$ can have multiple levels of positive response probabilities, depending on the extents of their "partial" capability of item j. Examples of multiparameter models include the Main-Effect and the All-Effect models introduced in Examples I.2 and I.3, respectively.

We would like to point out that the Γ-matrix defined in (3.5) still provides a useful technical tool for studying identifiability of multiparameter models, despite the fact that the entry $\Gamma_{j,\boldsymbol{\alpha}}$ only indicates whether $\boldsymbol{\alpha}$ belongs to the most-capable-set $\mathcal{C}_j$ and it does not summarize all the structural assumptions in multiparameter models.

On the one hand, similar to the two-parameter case, under a multiparameter model, the separability of the Γ-matrix is still *necessary* for the *strict* identifiability of $(\boldsymbol{\Theta}, \boldsymbol{p})$. This is because a two-parameter model, such as DINA, can be viewed as a submodel of a multiparameter model, such as GDINA or GDM, by constraining certain parameters in the multiparameter model to zero. So in order to ensure identifiability of all possible model parameters in the parameter space of a multiparameter model, Proposition III.1 implies the Γ must be separable.

On the other hand, when the Γ-matrix is inseparable and contains identical columns, the item parameter vectors in the matrix $\mathbf{\Theta}$ may still be distinct. This is because under the general constraints (1.2), when $\Gamma_{j,\boldsymbol{\alpha}} = 0$ under a multiparameter model, $\boldsymbol{\alpha}$ could be either least capable or partially capable of item j, and hence the latent classes in the set $\mathcal{A} \setminus \mathcal{C}_j = \{\boldsymbol{\alpha} : \Gamma_{j,\boldsymbol{\alpha}} = 0\}$ can still have different positive response probabilities, as shown in Examples I.2 and I.3. Such a difference from the two-parameter models makes the $\boldsymbol{p}$-partial identifiability theory developed in Section 3.2 not applicable to multiparameter models. To study identifiability of multiparameter models when Γ is inseparable, we therefore need an alternative partial identifiability notion and technique. We use the next example to illustrate this and show how the separable requirement of the Γ-matrix in Proposition III.1 could be relaxed under multiparameter models.

Example III.4. Consider the Q-matrix in (3.7). Under a two-parameter conjunctive restricted latent class model, we have shown attribute profiles $\boldsymbol{\alpha}_0 = (0,0)$ and $\boldsymbol{\alpha}_2 = (0,1)$ are not distinguishable. However, a multiparameter model models the main effect of each required attribute for an item. Consider the Main-Effect model with the identity link function as introduced in Example I.2 (the ACDM), one has $\mathbf{\Theta}_{\bullet,\boldsymbol{\alpha}_0} = (\beta_{1,0},\ \beta_{2,0})^\top$ and $\mathbf{\Theta}_{\bullet,\boldsymbol{\alpha}_2} = (\beta_{1,0},\ \beta_{2,0} + \beta_{2,2})^\top$; then $\mathbf{\Theta}_{\bullet,\boldsymbol{\alpha}_0} \neq \mathbf{\Theta}_{\bullet,\boldsymbol{\alpha}_2}$ as long as $\beta_{2,2} \neq 0$. When this inequality constraint $\beta_{2,2} \neq 0$ holds, $\mathbf{\Theta}_{\bullet,\boldsymbol{\alpha}_0} \neq \mathbf{\Theta}_{\bullet,\boldsymbol{\alpha}_2}$ despite that $\Gamma_{\bullet,\boldsymbol{\alpha}_0} = \Gamma_{\bullet,\boldsymbol{\alpha}_2}$. In such scenarios, the grouping operation of the proportion parameters introduced in Section 3.2 is not appropriate, and one needs to treat these two latent classes $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_2$ separately. Consider any possible $\mathbf{\Theta}$ for which the inequality constraint $\beta_{2,2} \neq 0$ does not hold, then all such $\mathbf{\Theta}$ indeed fall into a subset of the parameter space $\mathcal{T}$ with smaller dimension than $\mathcal{T}$, characterized by

$\mathcal{V} = \{(\boldsymbol{\Theta}, \boldsymbol{p}) : \beta_{2,2} = 0\}$. This implies that for almost all valid model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ in $\mathcal{T}$, except a Lebesgue measure zero set $\mathcal{V}$, the $\boldsymbol{\Theta}$ satisfy $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}_0} \neq \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}_1}$. This observation naturally leads to the following notion of generic identifiability.

Motivated by Example III.4, when the Γ-matrix is inseparable, we shall study the *generic identifiability* of the restricted latent class model. Let $\mathcal{T}$ denote the restricted parameter space of $(\boldsymbol{\Theta}, \boldsymbol{p})$ under the general constraints (1.2), and let d denote the number of free parameters in $(\boldsymbol{\Theta}, \boldsymbol{p})$, so $\mathcal{T}$ is of full dimension in $\mathbb{R}^d$. Generic identifiability means that identifiability holds for almost all points except a subset of $\mathcal{T}$ that has Lebesgue measure zero. Generic identifiability is closely related to the concept of algebraic variety in algebraic geometry. Following the definition in Allman et al. (2009), an algebraic variety $\mathcal{V}$ is defined as the simultaneous zero-set of a finite collection of multivariate polynomials $\{f_i\}_{i=1}^n \subseteq \mathbb{R}[x_1, x_2, \ldots, x_d]$, $\mathcal{V} = \mathcal{V}(f_1, \ldots, f_n) = \{\boldsymbol{x} \in \mathbb{R}^d \mid f_i(\boldsymbol{x}) = 0,\ 1 \leq i \leq n.\}$ An algebraic variety $\mathcal{V}$ is all of $\mathbb{R}^d$ only when all the polynomials defining it are zero polynomials; otherwise, $\mathcal{V}$ is called a *proper subvariety* and is of dimension less than d, hence necessarily of Lebesgue measure zero in $\mathbb{R}^d$. The same argument holds when $\mathbb{R}^d$ is replaced by the parameter space $\mathcal{T} \subseteq \mathbb{R}^d$ that has full dimension in $\mathbb{R}^d$. We next present the definition of generic identifiability for restricted latent class models.

Definition III.4 (Generic Identifiability)**.** A restricted latent class model is said to be generically identifiable on the parameter space $\mathcal{T}$, if $(\boldsymbol{\Theta}, \boldsymbol{p})$ are strictly identifiable on $\mathcal{T} \setminus \mathcal{V}$ where $\mathcal{V}$ is a proper algebraic subvariety of $\mathcal{T}$.

Generic identifiability could be viewed as some "partial" identification of model parameters in the sense that, the nonidentifiable parameters fall in a subset of the parameter space that can be characterized as solutions to some nonzero polynomial equations. As can be seen from the form of (1.2), the constraints on the parameter space introduced by the Γ-matrix already force the parameters fall into a proper algebraic subvariety of the unrestricted parameter space, so previous results established in Allman et al. (2009) for unrestricted latent class models do not apply to the models considered in this chapter.

Remark III.2. Under multiparameter models, it is still possible that two latent classes $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ always have the same positive response probabilities, i.e., $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}} = \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}'}$ and $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$ are not distinguishable even generically. In this case one could have $\boldsymbol{p}$-partial identifiability of the model. However, this happens only when $\Gamma_{\bullet,\boldsymbol{\alpha}} = \Gamma_{\bullet,\boldsymbol{\alpha}'} = \mathbf{1}$; moreover, under Q-restricted models, this happens only if the Q-matrix contains an all-zero column, which is a trivial case with a redundant column in Q. Under such a Q-matrix, we can simply remove these all-zero columns and study the (generic) identifiability under the reduced Q-matrix. Therefore, without loss of generality, in the following discussion we assume the Q-matrix does not contain any all-zero column such that $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}} = \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}'}$ would not happen.

Based on the above discussions, to study identifiability of multiparameter restricted latent class models, we consider two situations in Section 3.3.1: first, when the Γ-matrix is separable, we study the *strict* identifiability of model parameters; second, when the Γ-matrix is inseparable, we study the *generic* identifiability of model parameters. Furthermore, in Section 3.3.2 we present sufficient conditions for generic identifiability of the family of Q-restricted latent class models, and discuss the necessity of the proposed conditions.

3.3.1 Strict and Generic Identifiability

First consider the case where the Γ-matrix is separable. For a subset of items S, denote the corresponding $|S| \times m$ indicator matrix by $\Gamma^S = (\Gamma_{j,\boldsymbol{\alpha}},\ j \in S,\ \boldsymbol{\alpha} \in \mathcal{A})$, which is a submatrix of the previously defined Γ-matrix. We say $\boldsymbol{\alpha}$ succeeds $\boldsymbol{\alpha}'$ with respect to S and denote it by $\boldsymbol{\alpha} \succeq_S \boldsymbol{\alpha}'$, if $\Gamma_{j,\boldsymbol{\alpha}} \geq \Gamma_{j,\boldsymbol{\alpha}'}$ for any $j \in S$; this means $\boldsymbol{\alpha}$ is at least as capable as $\boldsymbol{\alpha}'$ of items in set S. With this definition, any subset of items S induces a partial order "$\succeq_S$" on the set of latent classes $\mathcal{A}$. When two sets S_1 and S_2 induce the same partial order on $\mathcal{A}$, that is, for any $\boldsymbol{\alpha}'$ and $\boldsymbol{\alpha} \in \mathcal{A}$, $\boldsymbol{\alpha}' \succeq_{S_1} \boldsymbol{\alpha}$ if and only if $\boldsymbol{\alpha}' \succeq_{S_2} \boldsymbol{\alpha}$, we write " $\succeq_{S_1}$ " = " $\succeq_{S_2}$ ". The following theorem gives conditions that lead to strict identifiability of multiparameter restricted latent class models.

Theorem III.5. *For a multiparameter restricted latent class model, if the* Γ*-matrix satisfies the following conditions, then the parameters* $(\boldsymbol{\Theta}, \boldsymbol{p})$ *are strictly identifiable:*

(C3) There exist two disjoint item sets S_1 *and* S_2*, such that* Γ^{S_i} *is separable for* $i = 1, 2$ *and* " $\succeq_{S_1}$ " = " $\succeq_{S_2}$ ".

(C4) $\Gamma^{(S_1 \cup S_2)^c}_{\bullet,\boldsymbol{\alpha}} \neq \Gamma^{(S_1 \cup S_2)^c}_{\bullet,\boldsymbol{\alpha}'}$ *for any* $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'$ *such that* $\boldsymbol{\alpha}' \succeq_{S_i} \boldsymbol{\alpha}$ *for* $i = 1$ *or* 2.

Condition (C3) implies the entire Γ-matrix is separable, and it requires two disjoint sets of items S_1 and S_2 to have enough information to distinguish the latent classes, and it serves as a *repeated measurement condition* for the identifiability of multiparameter restricted latent class models. Condition (C4) states that, for those pairs of latent classes $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ such that $\boldsymbol{\alpha}$ is more capable than $\boldsymbol{\alpha}'$ uniformly on either S_1 or S_2, the remaining items in $(S_1 \cup S_2)^c$ should differentiate $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ by their column vectors in $\Gamma^{(S_1 \cup S_2)^c}$.

Strict identifiability can be achieved with a relaxation of condition (C4) together with a stronger version of condition (C3). Before presenting this result, we define a latent class $\boldsymbol{\alpha}$ as a basis latent class under an item set S, if there does not exist $\boldsymbol{\alpha}' \in \mathcal{A}$ such that $\boldsymbol{\alpha}' \preceq_S \boldsymbol{\alpha}$. Denote the set of all basis latent classes under S by $\mathcal{B}_S$. Then "$\succeq_{S_1}$ " = " $\succeq_{S_2}$ " implies $\mathcal{B}_{S_1} = \mathcal{B}_{S_2}$.

Proposition III.5. Under a multiparameter restricted latent class model, if the Γ-matrix satisfies the following conditions, then $(\boldsymbol{\Theta}, \boldsymbol{p})$ are identifiable.

(C3*) There exist two disjoint item sets S_1 and S_2, such that Γ^{S_i} is separable for $i = 1, 2$ and " $\succeq_{S_1}$ " = " $\succeq_{S_2}$ ". Moreover, for any $j \in S_1 \cup S_2$, there exists $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$ such that $\Gamma_{j,\boldsymbol{\alpha}} = 1$.

(C4*) $\Gamma^{(S_1\cup S_2)^c}_{\bullet,\boldsymbol{\alpha}} \neq \Gamma^{(S_1\cup S_2)^c}_{\bullet,\boldsymbol{\alpha}_0}$ for any $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$ and $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}_0$, where $\boldsymbol{\alpha}_0$ is the universal least capable class.

Remark III.3. Theorem III.5 and Proposition III.5 show the trade-off between the conditions on the separable submatrices part of Γ and on the remaining part. They establish identifiability for a wide range of restricted latent class models, with the Γ-matrix ranging in the spectrum of different extents of inseparability. Specifically, for a Q-restricted latent class model that lacks many single-attribute items, (C3) is easier to satisfy than (C3*) and Theorem III.5 would be more applicable; while for a Q-restricted model that lacks few single-attribute items, Proposition III.5 would become more applicable as (C4*) imposes a weaker condition on the set $(S_1 \cup S_2)^c$.

Remark III.4. Theorem III.5 and Proposition III.5 extend the existing work Xu (2017). Compared with the identifiability result in Xu (2017) that requires two copies of the identity submatrix I_K to be included in the Q-matrix, in the special case with $\mathcal{A} = \{0,1\}^K$, the proposed conditions (C3*) and (C4*) reduce to the conditions in Xu (2017). Furthermore, in general cases of an unsaturated latent class space with $|\mathcal{A}| < 2^K$, the conditions in Theorem III.5 and Proposition III.5 impose much weaker requirements than those in Xu (2017), because a Q-matrix lacking some single-attribute items may suffice for a separable Γ-matrix and further suffice for strict identifiability under the conditions in this chapter.

Next, we consider the case where the multiparameter restricted latent class model is associated with an inseparable Γ-matrix, which violates condition (C3). We study the generic identifiability of the model parameters.

Theorem III.6. *Consider a multiparameter restricted latent class model. If there exist two disjoint item sets S_1 and S_2, such that altering some entries of zero to one in $\Gamma^{S_1\cup S_2}$ can yield a $\widetilde{\Gamma}^{S_1\cup S_2}$ that satisfies condition (C3); and that the $\Gamma^{(S_1\cup S_2)^c}$ satisfies condition (C4), then the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ under the original Γ-matrix are generically identifiable.*

Theorem III.6 is established based on the theoretical development of Theorem III.5. By relaxing the condition (C3) and allowing Γ to be inseparable, we may not have strict identifiability, as discussed in Example III.4. We use the following example to further illustrate the results of Theorems III.5–III.6.

Example III.5. For a multiparameter restricted latent class model, if $\Gamma = ((\Gamma^{sub})^\top, (\Gamma^{sub})^\top, (\Gamma^{sub})^\top)^\top$ contains three copies of the following Γ^{sub}, then (C3) and (C4) are satisfied and $(\boldsymbol{\Theta}, \boldsymbol{p})$ under Γ are strictly identifiable.

$$\Gamma^{sub} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}; \quad \Gamma^{S_1} = \begin{pmatrix} 0 & \mathbf{0} & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Gamma^{S_2} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & \mathbf{0} & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Instead, consider $\Gamma_{\text{new}} = ((\Gamma^{S_1})^\top, (\Gamma^{S_2})^\top, (\Gamma^{sub})^\top)^\top$ with two submatrices in the forms of Γ^{S_1} and Γ^{S_2} above, then neither of Γ^{S_i} is separable. But by changing the $(1,2)$th entry of Γ^{S_1} and $(2,3)$th entry of Γ^{S_2} from zero to one, the resulting $\widetilde{\Gamma}^{S_1}$ and $\widetilde{\Gamma}^{S_2}$ are separable, so the conditions of Theorem III.6 are satisfied and $(\boldsymbol{\Theta}, \boldsymbol{p})$ under Γ_{new} are generically identifiable.

3.3.2 Results for Q-restricted Latent Class Models

In this subsection we characterize how the Q-matrix impacts the identifiability of multiparameter models. Similar to Section 3.2.2, we consider the case $\mathcal{A} = \{0,1\}^K$. For strict identifiability, the result of either Theorem III.5 or Proposition III.5 implies the result of Theorem 1 in Xu (2017), as discussed in Remark III.4. Our next result gives a flexible structural condition on Q that leads to generic identifiability.

Theorem III.7. *Under a multiparameter Q-restricted latent class model, if the Q-matrix satisfies the following conditions, then the model parameters are generically identifiable, up to label swapping among those latent classes that have identical column vectors in Γ.*

(C5) Q contains two $K \times K$ submatrices Q_1, Q_2, such that for $i = 1, 2$,

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q' \end{pmatrix}_{J\times K}; \quad Q_i = \begin{pmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & 1 \end{pmatrix}_{K\times K}, \quad i = 1, 2, \tag{3.16}$$

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

(C6) With the Q-matrix taking the form of (5.5)*, in the submatrix Q' each attribute is required by at least one item.*

The above identifiability result does not require the Q to contain an identity submatrix I_K and provides a flexible new condition for generic identifiability that are satisfied by various Q-matrix structures; see examples in Section 3.3.3. Under a multiparameter restricted latent class model with all entries of the Q-matrix being ones, conditions (C5) and (C6) in Theorem III.7 equivalently reduce to $J \geq 2K + 1$, which is consistent with the result in Allman et al. (2009) for unrestricted latent class models.

Next we discuss the necessity of the proposed sufficient conditions for generic identifiability. Conditions (C5) and (C6) imply that each attribute is required by at least three items. The next theorem shows that it is necessary for each attribute to be required by at least two items.

Theorem III.8. *Consider a multiparameter Q-restricted latent class model.*

(a) If some attribute is required by only one item, then the model is not generically identifiable.

(b) If some attribute is required by only two items, without loss of generality assume

Q takes the following form

$$Q = \begin{pmatrix} 1 & \boldsymbol{v}_1^\top \\ 1 & \boldsymbol{v}_2^\top \\ 0 & Q' \end{pmatrix}, \tag{3.17}$$

then as long as $\boldsymbol{v}_1 \vee \boldsymbol{v}_2 \neq \mathbf{1}_{K-1}$ and the submatrix Q' satisfies conditions (C5) and (C6), then the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ are generically identifiable, up to label swapping among those latent classes that have identical column vectors in Γ.

Remark III.5. As a notion of partial identification of model parameters, generic identifiability does not imply strict identifiability. For instance, if the Q-matrix is in the form of (3.17) and $\boldsymbol{v}_i = \mathbf{0}$ for $i = 1$ and 2, then the model is not strictly identifiable, but generic identifiability can still hold as stated in Theorem III.8. This is also an analogue to the situations discussed in Theorem III.2 for two-parameter restricted latent class models. Based on Theorems III.7 and III.8, we would recommend practitioners in diagnostic test designs to ensure each attribute is measured by at least three items.

3.3.3 Applications

Similar to the discussion in Section 3.2.3, our results of generic identifiability also lead to the estimability of the model parameters.

Proposition III.6. Suppose a restricted latent class model is generically identifiable on the parameter space $\mathcal{T}$ with a measure-zero nonidentifiable set $\mathcal{V}$. If the true parameters $(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) \in \mathcal{T} \setminus \mathcal{V}$, then $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}}) \to (\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ almost surely as $N \to \infty$.

We apply the new theory of generic identifiability to the designs introduced in Section 1.3, and establish generic identifiability of the multiparameter restricted latent class models. Consider the TOEFL iBT data. Both Q-matrices corresponding to TOEFL reading forms A and B can be transformed into the form of (5.5) through some row rearrangements, with the corresponding Q' requiring each attribute at least once. Therefore both Q-matrices satisfy conditions (C5) and (C6) and any multiparameter Q-restricted models associated with them are generically identifiable and estimable. Our results in this section also guarantee the generic identifiability of multiparameter models associated with the $Q_{43\times 12}$ for the TIMSS data, and the $Q_{20\times 8}$ for the fraction subtraction data; please see Section A in the Supplementary Material for details of checking the conditions.

3.4 Extensions to More Complex Models

In this section, we extend our identifiability theory to some more complicated latent variable models.

3.4.1 Mixed-items Restricted Latent Class Models

Our identifiability theory based on Γ directly applies to the case of mixed types of items, where the J items can conform to different models, including two-parameter conjunctive, two-parameter disjunctive, or multiparameter.

First consider the *two-parameter-mixed* restricted latent class model, where each item is either two-parameter conjunctive or disjunctive. Let $I(\cdot)$ denote the binary indicator function. For any Q-matrix and latent class space $\mathcal{A}$, denote the Γ-matrix under the two-parameter conjunctive model by $\Gamma^{conj}(Q,\mathcal{A})$ with the $(j,\boldsymbol{\alpha})$th entry being $I(\boldsymbol{\alpha}\succeq \boldsymbol{q}_j)$, and denote the Γ-matrix under the two-parameter disjunctive model by $\Gamma^{disj}(Q,\mathcal{A})$ with the $(j,\boldsymbol{\alpha})$th entry being $I(\exists k \text{ s.t. } \alpha_k = q_{j,k} = 1)$. The following is a corollary of Theorem III.1.

Corollary III.2. *Consider a two-parameter-mixed restricted latent class model with $Q = (Q_{disj}^\top, Q_{conj}^\top)^\top$, where Q_{disj} and Q_{conj} correspond to disjunctive and conjunctive* *items, respectively. If the following condition (E1) holds, then* $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ *are* $\boldsymbol{p}$*-partially identifiable.*

(E1) The $J \times |\mathcal{A}|$ *matrix* $\Gamma = (\Gamma^{disj}(Q_{disj}, \mathcal{A})^\top,\ \Gamma^{conj}(Q_{conj}, \mathcal{A})^\top)^\top$ *satisfies conditions (C1) and (C2) in Theorem III.1.*

In particular, if $\mathcal{A} = \{0,1\}^K$ *and the* Γ *defined in (E1) is separable, then* $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ *are strictly identifiable.*

One implication of Corollary III.2 is that when a diagnostic test contains both conjunctive and disjunctive items, the underlying Q-matrix does not need to include a submatrix I_K for $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ to be strictly identifiable. This is in contrary to the case of a purely conjunctive or purely disjunctive two-parameter model, where this requirement is indeed necessary Xu and Zhang (2016); Gu and Xu (2019b). The following application of Corollary III.2 illustrates this point.

Example III.6. Consider a diagnostic test with 4 conjunctive items and 2 disjunctive items with the following Q-matrix

$$Q = \begin{pmatrix} Q^{conj}_{4\times 2} \\ \hline Q^{disj}_{2\times 2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ \hline 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \Longrightarrow \quad \Gamma = \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \left(\begin{array}{cccc} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ \hline 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{array}\right) \end{array}.$$

Then if $\mathcal{A} = \{0,1\}^2$, the corresponding Γ-matrix as shown above is separable, and conditions (C1*) and (C2*) are satisfied. So $\boldsymbol{\theta}^+ = (\theta_1^+, \ldots, \theta_6^+)^\top$, $\boldsymbol{\theta}^- = (\theta_1^-, \ldots, \theta_6^-)^\top$ and $\boldsymbol{p} = (p_{(0,0)},\ p_{(0,1)},\ p_{(1,0)},\ p_{(1,1)})^\top$ are strictly identifiable, despite that Q does not contain a submatrix I_2.

If there exist both two-parameter items and multiparameter items in the model, we have the following identifiability result, the part (a) of which directly results from Theorem III.5 and Proposition III.5. Please see Section D in the Supplementary Material for details.

Corollary III.3. *Assume $Q = (Q_{disj}^\top, Q_{conj}^\top, Q_{mult}^\top)^\top$ where Q_{disj}, Q_{conj} and Q_{multi} correspond to the two-parameter disjunctive, two-parameter conjunctive, and multiparameter items, respectively.*

(a) If $\Gamma = (\Gamma^{disj}(Q_{disj}, \mathcal{A})^\top,\ \Gamma^{conj}(Q_{conj}, \mathcal{A})^\top,\ \Gamma^{conj}(Q_{mult}, \mathcal{A})^\top)^\top$ satisfies conditions (C3) and (C4) in Theorem III.5; or conditions (C3) and (C4*) in Proposition III.5, then $(\boldsymbol{\Theta}, \boldsymbol{p})$ are strictly identifiable.*

(b) If Γ satisfies condition (E2) in Section D of the Supplementary Material, then $(\boldsymbol{\Theta}, \boldsymbol{p})$ are generically identifiable.

3.4.2 Restricted Latent Class Models with Categorical Responses

We next study restricted latent class models with multiple levels of responses per item, that is, categorical responses, instead of binary responses considered in previous sections. These models have been considered in von Davier (2008), Ma and de la Torre (2016) and Chen and de la Torre (2018). We consider the setting in Chen and de la Torre (2018). Suppose for each item j out of the J items in a diagnostic test, there are L_j categories of responses. For each item j and each category of response $l \in \{0, \ldots, L_j - 1\}$, there are a set of positive response parameters of the latent classes $\boldsymbol{\theta}_j^{(l)} = \{\theta_{j,\boldsymbol{\alpha}}^{(l)} : \boldsymbol{\alpha} \in \mathcal{A}\}$ with $\boldsymbol{\theta}_j^{(0)} = \mathbf{1} - \sum_{l>0} \boldsymbol{\theta}_j^{(l)}$. Further, for each item j, the $\boldsymbol{q}$-vector $\boldsymbol{q}_j$ constrains the vector $\boldsymbol{\theta}_j^{(l)}$ based on (1.2) for each category $l \in \{1, \ldots, L_j - 1\}$ independently, other than the basic level $l = 0$. Namely, for any $j \in \mathcal{S}$,

$$\max_{\boldsymbol{\alpha} \in \mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}}^{(l)} = \min_{\boldsymbol{\alpha} \in \mathcal{C}_j} \theta_{j,\boldsymbol{\alpha}}^{(l)} > \theta_{j,\boldsymbol{\alpha}'}^{(l)}, \quad \forall l \in \{1, \ldots, L_j - 1\} \text{ and } \forall \boldsymbol{\alpha}' \notin \mathcal{C}_j.$$

We collect all the model parameters in $(\mathbf{\Theta}^{\text{cat}}, \boldsymbol{p})$ with $\mathbf{\Theta}^{\text{cat}} = \{\boldsymbol{\theta}_j^{(l)} : j = 1, \ldots, J;\ l = 0, \ldots, L_j - 1\}$. Then we have the following identifiability result.

Proposition III.7. For a given Q-matrix, consider the following cases.

(a) If for any $j \in \mathcal{S}$ and $l \in \{1, \ldots, L_j\}$, item parameters $\{\theta^l_{j,\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}\}$ follow the two-parameter assumption, and Q satisfies (C1*) and (C2*) in Corollary III.1, then $(\mathbf{\Theta}^{\text{cat}}, \boldsymbol{p})$ are $\boldsymbol{p}$-partially identifiable.

(b) If for any $j \in \mathcal{S}$ and $l \in \{1, \ldots, L_j\}$, item parameters $\{\theta^l_{j,\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}\}$ follow the multiparameter assumption, and Q satisfies conditions (C5) and (C6) in Theorem III.7, then $(\mathbf{\Theta}^{\text{cat}}, \boldsymbol{p})$ are generically identifiable.

3.4.3 Deep Restricted Boltzmann Machines

As mentioned in Example I.4 in Chapter I, structured latent attribute models share great similarities with Restricted Boltzmann Machines (RBM) (Goodfellow et al., 2016). Here we restate how the RBM architecture can be used as a special restricted latent class model for cognitive diagnosis. The RBM on the right panel of Figure 3.1 consists of two latent layers $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ and one observed layer $\boldsymbol{R}$. In a diagnostic test, the $\boldsymbol{R}$ represents multivariate binary responses to test items, the first latent layer $\boldsymbol{\alpha}^{(1)}$ represents the fine-grained binary skill attributes measured by the items, while the second binary latent layer $\boldsymbol{\alpha}^{(2)}$ helps to model the dependence among $\boldsymbol{\alpha}^{(1)}$ and may be interpreted as more general skill domains. Denote the lengths of vectors $\boldsymbol{R}$, $\boldsymbol{\alpha}^{(1)}$ and $\boldsymbol{\alpha}^{(2)}$ by J, K_1 and K_2. Under RBM assumptions, the probability distribution of all the observed and latent variables is

$$\mathbb{P}(\boldsymbol{R}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}) = \frac{1}{Z} \exp\Big(- \boldsymbol{R}^\top \boldsymbol{W}^Q \boldsymbol{\alpha}^{(1)} - (\boldsymbol{\alpha}^{(1)})^\top \boldsymbol{U} \boldsymbol{\alpha}^{(2)} \Big), \tag{3.18}$$

where Z is the normalization constant, and $\boldsymbol{W}^Q$, $\boldsymbol{U}$ are parameter matrices, of size $J \times K_1$ and $K_1 \times K_2$, respectively. We drop the bias terms in the above energy function without loss of generality (Goodfellow et al., 2016). We can impose a Q-matrix of size $J \times K_1$ to restrict the parameters $\boldsymbol{W}^Q$ in (3.18). Specifically, Q specifies which entries of $\boldsymbol{W}^Q = (w_{j,k})$ are zero, that is, $w_{j,k} = 0$ if $q_{j,k} = 0$. The form of Q underlying the $\boldsymbol{W}^Q$ in Figure 3.1 is on the left panel of the figure.

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

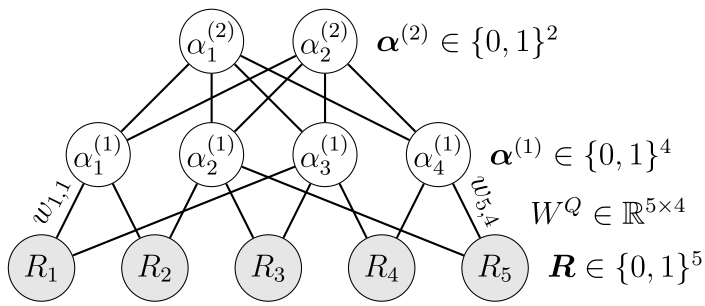


Figure 3.1: (Deep) Restricted Boltzmann machine

We call $\boldsymbol{W}^Q$ the *item parameters* of a RBM, since these parameters relate to the observed responses to items; and call a RBM with a Q-matrix structure an *item-parameter-restricted* RBM. Then an item-parameter-restricted RBM can be viewed as a multiparameter main-effect restricted latent class model, with $\boldsymbol{\alpha}^{(1)}$ belonging to the latent class space $\{0,1\}^{K_1}$. The next proposition establishes identifiability of the item parameters $\boldsymbol{W}^Q$.

Proposition III.8. For a given Q-matrix, consider the following cases.

(a) If there is no sparsity structure in $\boldsymbol{W}^Q$ (i.e., $Q = \mathbf{1}_{J\times K}$), then as long as $J \geq 2K_1 + 1$, the item parameters $\boldsymbol{W}^Q$ are generically identifiable.

(b) If the Q-matrix satisfies the sufficient conditions for strict or generic identifiability in Section 3.3, then $\boldsymbol{W}^Q$ are strictly or generically identifiable, respectively.

Proposition III.8 establishes identifiability of the item parameters $\boldsymbol{W}^Q$, which provides the theoretical guarantee in the application of item calibration to assess the quality of the items. It would also be interesting to further investigate identifiability of other parameters besides the item parameters in a deep restricted Boltzmann machine, which we leave for future study.

3.5 Discussion

This chapter proposes a general framework of strict and partial identifiability of restricted latent class models.

We provide a flowchart in Figure 3.2 to summarize our main theoretical results in Sections 3.2 and 3.3. The flowchart illustrates how to apply the new theory in cognitive diagnosis. Specifically, given the specification of the Q-matrix, the latent class space $\mathcal{A} \subseteq \{0,1\}^K$, and the diagnostic model assumptions, one can construct the corresponding $J \times |\mathcal{A}|$ Γ-matrix based on the $\mathcal{C}_j$'s defined in (3.1). Then in the case of a separable Γ-matrix, if the model is two-parameter, the $\boldsymbol{p}$-partial identifiability exactly reduces to strict identifiability and one can use results in Section 3.2 to establish strict identifiability; and if the model is multiparameter, one can use results Theorem III.5 and Proposition III.5 in Section 3.3.1 for strict identifiability. On the other hand, if the Γ-matrix is inseparable, depending on whether the model is two-parameter or multiparameter, one can use the results in Section 3.2.2 or those in Section 3.3 to check whether the model is $\boldsymbol{p}$-partially identifiable or generically identifiable, respectively. Note that in the special case of $\mathcal{A} = \{0,1\}^K$, the Γ-matrix with 2^K columns is separable if and only if the Q-matrix contains an identity submatrix I_K, a key condition assumed in previous works (e.g., Xu, 2017; Xu and Shang, 2018). Hence, this chapter not only largely relaxes these existing conditions for strict identifiability by allowing more flexible attribute structures with an arbitrary $\mathcal{A}$, but also provides the first study on partial identifiability when the Q-matrix does not include an I_K (the Γ-matrix is inseparable). We give easily-checkable identifiability conditions to ensure estimability of the model parameters, and these conditions serve as practical guidelines for designing statistically valid diagnostic tests.

We point out that the strict identifiability results in Section 3.3.1 (Theorem III.5 and Proposition III.5) apply to the general family of restricted latent class models satisfying constraints (1.2), including not only multiparameter but also two-parameter

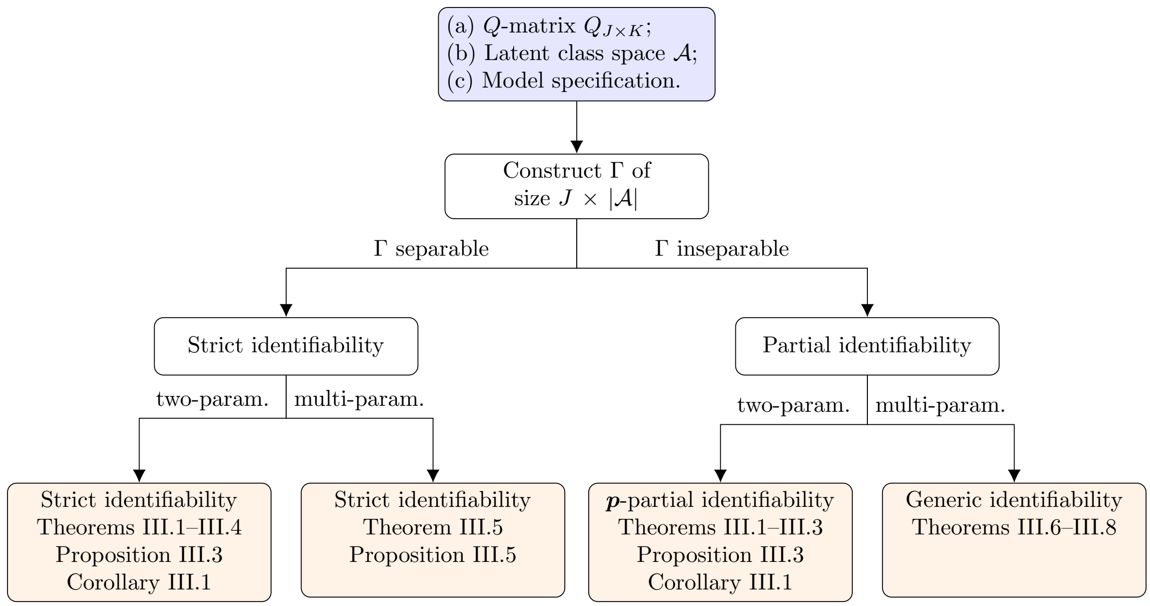


Figure 3.2: Flowchart of the results in Sections 3.2 and 3.3

models; on the other hand, since these results are established under the general constraints (1.2), their conditions are stronger than those in Section 3.2 under two-parameter models. In contrast, the generic identifiability results in Sections 3.3.1 and 3.3.2 (Theorems III.6–III.8) only apply to multiparameter models. This is because under generic identifiability, the nonidentifiable measure-zero subset of a multiparameter model's parameter space (such as GDINA), could still contain the parameter space corresponding to a two-parameter submodel (such as DINA), making these generic identifiability results not applicable to two-parameter models. Nevertheless, generic identifiability is a general concept not just restricted to the multiparameter models. An interesting future direction to study is the generic identifiability of two-parameter models under the introduced $\boldsymbol{p}$-partial identifiability framework; that is, one can study what conditions lead to the generic identifiability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$. We also point that a multiparameter model can also be $\boldsymbol{p}$-partially identifiable, as discussed in Remark III.2.

For the $\boldsymbol{p}$-partial identifiability and generic identifiability results in Sections 3.2–3.4, we assume that the model specification for each item, the design matrix and latent class space $\mathcal{A}$ are available as prior knowledge. In practice, there can be scenarios where not all of such information is available. As pointed out by one reviewer, in applications of cognitive diagnostic modeling, both the advances in modeling capacity and computing flexibility, and the recent real-data examples provide ground for adopting a model with mixed type of items, which are determined in a data-driven way. To this end, our strict identifiability results in Section 3.3.1 and those in Section 3.4.1 for mixed-items models can be applied to assess identifiability *a posteriori*. In practice, when deciding which model to adopt, one can use the response data to determine the number of latent classes and determine which diagnostic model an item conforms to. For instance, one may employ the popular information criteria such as AIC and BIC to perform model selection; or one may first fit a general cognitive diagnostic model, such as GDINA or GDM, then use the Wald test to determine which submodel an item follows de la Torre (2011). Alternatively, one may use a penalized likelihood method Xu and Shang (2018) or Bayesian method Chen et al. (2018a) to directly estimate the structure of the item parameters for each item; such structure informs the model specification of the item. For the selected candidate models, we would recommend further applying our identifiability theory to assess their identifiability and validity. The general theoretical framework developed in this chapter would be a useful tool to develop the identifiability and estimability conditions for learning the item-level model structure and the population-level latent class space $\mathcal{A}$. This is an interesting and important direction that we plan to pursue in the future.

CHAPTER IV

Necessary and Sufficient Conditions for the Identifiability of the Q-matrix

The previous two chapters study the identifiability of model parameters of SLAMs given a fixed and known Q-matrix. This chapter addresses a further question, which goes beyond merely identifying the model parameters. Rather, the main goal here is to identify the key latent structure, that is, the Q-matrix itself. In practice, the Q-matrix, specified by scientific experts when constructing the diagnostic items, can be misspecified. Moreover, in an exploratory analysis of newly designed items, much or all of the Q-matrix may not be available. Here, a misspecification of the Q-matrix could lead to a serious lack of fit for the model, and thus inaccurate inferences on the latent attribute profiles of the individuals. Therefore, it is desirable to estimate the Q-matrix and the model parameters jointly from the response data (e.g., de la Torre, 2008; DeCarlo, 2012; Liu et al., 2012; de la Torre and Chiu, 2016; Chen et al., 2018a). A reliable and valid estimation and inference on the Q-matrix requires that we ensure the joint identifiability of the Q-matrix and the associated model parameters. Such joint identifiability has been studied recently by Liu et al. (2013) and Chen et al. (2015) under the DINA model, and by Xu and Shang (2018) under general

This chapter contains the main part of Gu and Xu (2020b), accepted by *Statistica Sinica*.

RLCMs. Nevertheless, most of these works focus on developing sufficient conditions for joint identifiability, and thus often impose stronger than needed constraints on the experimental design of a cognitive diagnosis.

Therefore, the necessary and sufficient conditions (or minimal requirements) for the joint identifiability of the Q-matrix and the model parameters remains an open problem. This study addresses this problem, making the following contributions to the literature.

First, under the DINA model, we derive the necessary and sufficient conditions for the joint identifiability of the Q-matrix and the associated DINA model parameters. Our necessary and sufficient conditions are succinctly and neatly written as three algebraic properties of the Q-matrix, which we summarize as *completeness* (Condition A), *distinctness* (Condition B), and *repetition* (Condition C); please see Theorem IV.1 for details. These three conditions require that the binary Q-matrix is *complete* by containing an identity submatrix, has all columns *distinct* other than the part of the identity submatrix, and *repeatedly* contains at least three entries of one in each column. In addition to guaranteeing identifiability, these conditions give the minimal requirements for the Q-matrix and DINA model parameters to be estimable from the observed responses. The identifiability result can be applied directly to the deterministic input noisy output "Or" gate (DINO) model (Templin and Henson, 2006), owing to the duality of the DINA and DINO models (Chen et al., 2015). The derived identifiability conditions also serve as necessary requirements for joint identifiability under general RLCMs, which include the DINA model as a submodel.

Second, we propose sufficient and necessary conditions for a weaker notation of identifiability, the so-called generic identifiability, under both the DINA model and general RLCMs. Generic identifiability implies that those parameters for which identifiability does not hold live in a set of Lebesgue measure zero (Allman et al., 2009). The motivation for studying generic identifiability is that the strict identifiability conditions are sometimes too restrictive in practice. For instance, it is known that unrestricted latent class models are not strictly identifiable (Gyllenberg et al., 1994), but are generically identifiable under certain conditions (Allman et al., 2009). In RLCMs, the model parameters are forced by the Q-matrix-induced constraints to fall in a measure-zero subset of the parameter space, and, thus, existing results for unrestricted models cannot be applied directly. Moreover, the generic identifiability conditions needed to jointly identify the Q-matrix and the model parameters are unknown. Therefore, in this chapter, we propose sufficient and necessary conditions for generic identifiability, and explicitly characterize the nonidentifiable measure-zero subset. Our mild sufficient conditions for generic identifiability under general RLCMs can be summarized as the following properties of the Q-matrix: *double generic completeness* (Condition D), and *generic repetition* (Condition E); see Theorem IV.4 for details. These two conditions require that the binary Q-matrix contains two *generically complete* square submatrices with all diagonal elements equal to one, and (*repeatedly*) contains at least one entry of "1" other than the part comprising these two submatrices.

The rest of this chapter is organized as follows. Section 4.1 defines strict and generic identifiability for RLCMs, and presents an illustrative example. Sections 4.2 and 4.3 contain our main theoretical results for strict and generic identifiability for the DINA model and multiparameter RLCMs, respectively. Section 4.4 concludes the chapter. The proofs of the theoretical results and additional simulation studies that verify the developed theory are included in Appendix C. The Matlab code used to check the proposed conditions is available at `https://github.com/yuqigu/Identify_Q`.

4.1 Definitions and Examples of Strict and Generic Identifiability

This section introduces the definitions of joint strict identifiability and joint generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ for SLAMs, and gives an illustrative example.

Note that the monotonicity assumption stated in (1.2), is necessary for the identifiability of the Q-matrix, because, without it, $Q \neq \mathbf{1}_{J\times K}$ with parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ is distinguished from $\bar{Q} = \mathbf{1}_{J\times K}$ with the same parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ under the multi-parameter SLAM. The monotonicity constraints ensure that the constraints induced by $Q \neq \mathbf{1}_{J\times K}$ and $\bar{Q} = \mathbf{1}_{J\times K}$ cannot be the same and, therefore, Q can be identified under additional conditions; see Sections 4.2 and 4.3. In the following we assume the monotonicity assumption introduced in Section 2 is satisfied.

Another common issue with the identifiability of the Q-matrix is label swapping. In an RLCM setting, arbitrarily reordering the columns of a Q-matrix does not change the distribution of the responses. As a result, it is only possible to identify Q up to column permutation; thus, we write $\bar{Q} \sim Q$ if $\bar{Q}$ and Q have an identical set of column vectors, and write $(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \sim (Q, \boldsymbol{\Theta}, \boldsymbol{p})$ if $\bar{Q} \sim Q$ and $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) = (\boldsymbol{\Theta}, \boldsymbol{p})$.

We first define the identifiability of the Q-matrix and the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$. We refer to this as *joint strict identifiability.*

Definition IV.1 (Joint Strict Identifiability)**.** Under an RLCM, the design matrix Q joint with the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ are said to be strictly identifiable if for any $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$, there is no $(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \nsim (Q, \boldsymbol{\Theta}, \boldsymbol{p})$ such that

$$\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \quad \text{for all} \quad \boldsymbol{r} \in \{0,1\}^J. \tag{4.1}$$

In the following discussion, we write (5.9) simply as $\mathbb{P}(\boldsymbol{R} \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$.

Despite being the most stringent criterion for identifiability, strict identifiability can be too restrictive, ruling out many cases where $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ are "almost surely" identifiable. In the literature on unrestricted latent class models, Allman et al. (2009) proposed and studied the so-called *generic identifiability* of such models. Here, we introduce the concept of generic identifiability for RLCMs as follows.

Definition IV.2 (Joint Generic Identifiability)**.** Consider an RLCM with parameter space $\boldsymbol{\vartheta}_Q$, which is of full dimension in $\mathbb{R}^m$, with m corresponding to the number of free parameters in the model. The matrix Q joint with the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ are said to be generically identifiable if the following set has Lebesgue measure zero in $\mathbb{R}^m$: $\boldsymbol{\vartheta}_{non} = \{(\boldsymbol{\Theta}, \boldsymbol{p}) : \exists(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \nsim (Q, \boldsymbol{\Theta}, \boldsymbol{p}),\ \text{such that}\ \mathbb{P}(\boldsymbol{R} \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})\}$.

4.1.1 Example of the Generic Identifiability Phenomenon with $Q_{4\times 2}$

Here, we use an example to explain the difference between generic identifiability and strict identifiability. Consider the Q-matrix $Q_{4\times 2}$ in (5.2). Under the DINA model, we prove that this Q-matrix, joint with the associated model parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$, is generically identifiable (by part (b.2) of Theorem IV.2), but not strictly identifiable (by Theorem IV.1).

$$Q_{4\times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.2}$$

In particular, as long as the true proportions $\boldsymbol{p} = (p_{(00)}, p_{(01)}, p_{(10)}, p_{(11)})$ satisfy the following inequality constraint, $(Q_{4\times 2}, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is identifiable (see the proof of Theorem IV.2 (b.2)):

$$p_{(01)}p_{(10)} \neq p_{(00)}p_{(11)}. \tag{4.3}$$

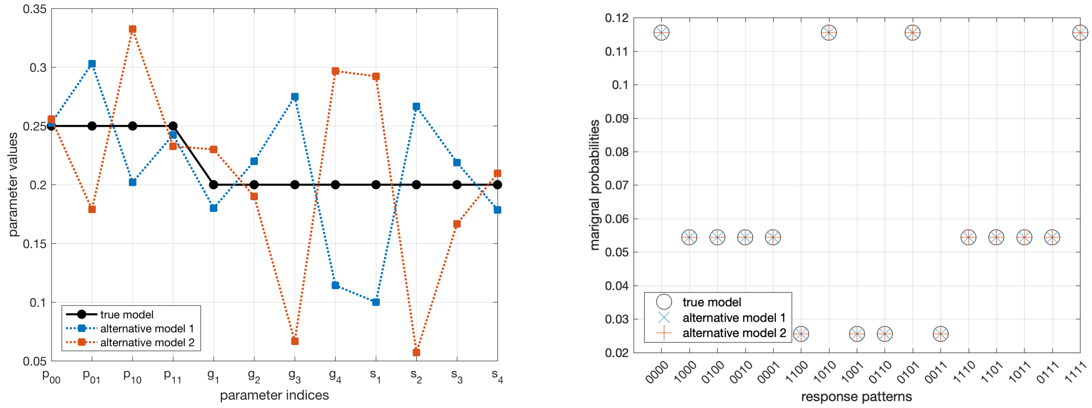
On the other hand, when $p_{(01)}p_{(10)} = p_{(00)}p_{(11)}$, the model parameters are not identifiable, and there exist infinitely many sets of parameters that provide the same distribution of the observed response vector. Here, the parameter space $\boldsymbol{\vartheta}_Q = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \mathbf{1} - \boldsymbol{s} \succ \boldsymbol{g},\ \boldsymbol{p} \succ \mathbf{0},\ \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = 1\}$ is of full dimension in $\mathbb{R}^{11}$, where the nonidentifiable subset $\boldsymbol{\vartheta}_{non} = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : p_{(01)}p_{(10)} = p_{(00)}p_{(11)}\}$ has Lebesgue measure zero in $\mathbb{R}^{11}$. We use a simulation study to illustrate the generic identifiability phenomenon. Under the $Q_{4\times 2}$ in (5.2), consider the following two simulation scenarios:

(a) the true model parameters are set as $g_j = s_j = 0.2$ for $j = 1, 2, 3, 4$, and $p_{(00)} = p_{(01)} = p_{(10)} = p_{(11)} = 0.25$, which violates (4.3);

(b) the true model parameters are generated randomly, which almost always satisfies (4.3). Specifically, we randomly generate 100 true parameter sets $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ using the following generating mechanism: $s_j \sim \mathcal{U}(0.1, 0.3)$, $g_j \sim \mathcal{U}(0.1, 0.3)$ for $j = 1, 2, 3, 4$, and $\boldsymbol{p} \sim \text{Dirichlet}(3, 3, 3, 3)$. Here $\mathcal{U}(0.1, 0.3)$ denotes the uniform distribution on $[0.1, 0.3]$, and $\text{Dirichlet}(3, 3, 3, 3)$ denotes the Dirichlet distribution with parameter vector $(3, 3, 3, 3)$.

We show numerically that in scenario (a), there exist multiple sets of valid DINA parameters that give the same distribution of $\boldsymbol{R}$; in scenario (b), the model $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is almost surely identifiable and estimable. In particular, corresponding to scenario (a), Figure 4.1 (a) plots the true model parameters and the other two sets of valid DINA model parameters (constructed based on the derivations in the proof of Theorem IV.2 (b.2)), and Figure 4.1 (b) plots the marginal probabilities of all $2^4 = 16$ response patterns under the three sets of model parameters. We can see that despite these three sets of parameters being quite different, they give the identical distribution of the four-dimensional binary response vector.

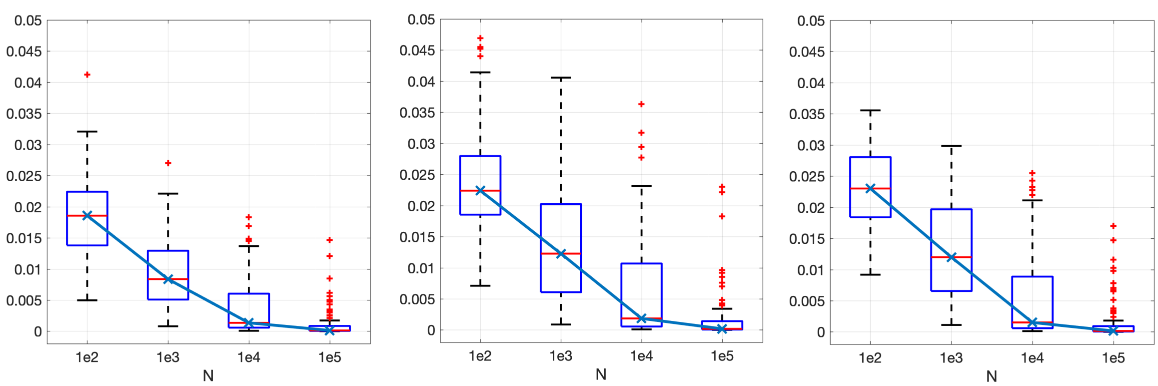
Corresponding to scenario (b), we randomly generate $B = 100$ sets of true parameters $(\boldsymbol{s}^i, \boldsymbol{g}^i, \boldsymbol{p}^i)$, for $i = 1, \ldots, 100$. Then, for each $(\boldsymbol{s}^i, \boldsymbol{g}^i, \boldsymbol{p}^i)$, we generate 200



(a) (b)

Figure 4.1: Illustration of nonidentifiability under $Q_{4\times 2}$ in scenario (a).

independent data sets of size N, with $N = 10^2$, 10^3, 10^4, and 10^5, and then compute the mean square errors (MSEs) of the maximum likelihood estimators (MLEs) of the slipping, guessing and proportion parameters. To compute the MLEs of the model parameters for each simulated data set, we run the EM algorithm with 10 random initializations, and choose the estimators that achieve the largest log-likelihood value of the 10 runs. Figure 4.2 shows the box plots of MSEs associated with the $B = 100$ true parameter sets for each sample size N. As N increases, we observe that the MSEs decrease to zero, indicating the (generic) identifiability of these randomly generated parameters.



(a) MSE of $\boldsymbol{p}$ (b) MSE of $\boldsymbol{s}$ (c) MSE of $\boldsymbol{g}$

Figure 4.2: Illustration of generic identifiability under $Q_{4\times 2}$, which corresponds to simulation scenario (b).

On the other hand, Figure 4.2 also shows that several parameter sets have MSEs that are "outliers" that converge to zero more slowly than others do as N increases. This happens because these sets of parameters fall near the nonidentifiability set $\mathcal{V}_{non} = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \; p_{(01)}p_{(10)} - p_{(00)}p_{(11)} = 0\}$, making it more difficult to identify them. To illustrate this point, consider the scenario corresponding to the rightmost box plot in Figure 4.2(a), with sample size $N = 10^5$. For each of the 100 sets of true parameters $(\boldsymbol{s}^i, \boldsymbol{g}^i, \boldsymbol{p}^i)$, we plot $p^i_{(00)} \cdot p^i_{(11)}$ and $p^i_{(01)} \cdot p^i_{(01)}$ as the x-axis and y-axis coordinates, respectively (see Figure 4.3). Then, each point represents one set of true parameters used to generate the data. Specifically, we plot these parameter sets using a red "$*$" if their corresponding MSEs are the 20% largest outliers in the rightmost box plot in Figure 4.2(a); we plot the remaining 80% of the parameter sets using a blue "+". One can clearly see that as the true parameters become closer to the nonidentifiability set $\mathcal{V}_{non} = \{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \; p_{(01)}p_{(10)} - p_{(00)}p_{(11)} = 0\}$ (represented by the straight reference line drawn from $(0, 0)$ to $(0.17, 0.17)$), the MSEs increase, and the MSEs converge more slowly. Thus, under generic identifiability, when the true model is close to the nonidentifiable set, the convergence of their MLEs becomes slow.

Interestingly, the generic identifiability constraint (4.3) is equivalent to the statement that the two latent attributes are *not independent* of each other. To see this, view each subject's two-dimensional attribute profile as a random vector taking values in a 2×2 contingency table. Then, (4.3) states that the 2×2 matrix of joint probabilities of attributes mastery,

$$\begin{pmatrix} p_{(00)} & p_{(01)} \\ p_{(10)} & p_{(11)} \end{pmatrix},$$

has full rank, with nonzero determinant $p_{(00)}p_{(11)} - p_{(01)}p_{(10)}$. Therefore, one row (resp. column) of the matrix cannot be a multiple of the other row (resp. column), and hence the two binary attributes can not be independent. Intuitively, this implies that the

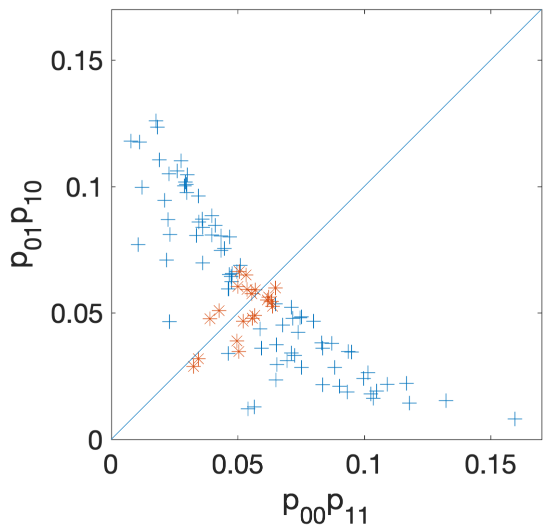
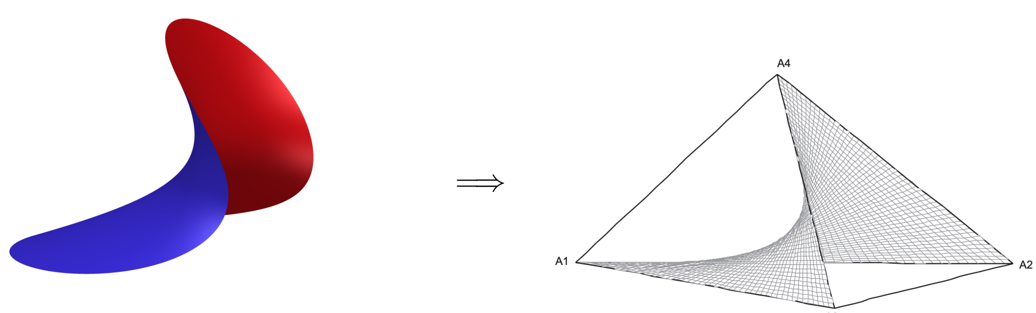


Figure 4.3: The effect of the generic identifiability constraint (4.3). Red "$*$"s represent parameter sets with the 20% largest MSEs in Figure 4.2(a), with $N = 10^5$; blue "+"s represent the remaining parameter sets.

DINA model essentially requires that each attribute is measured at least three times for identifiability (as shown in Condition B in Theorem IV.1). In particular, consider those attributes that are measured by only two items in the Q-matrix. If these attributes are independent, then, intuitively, they provide an independent source of information in which case the model is not identifiable. However, if these attributes are dependent, then the dependency instead helps to identify the model structure.



(a) measure-zero nonidentifiable subset, independent attributes

(b) parameter space for the proportion parameters $\boldsymbol{p}$

Figure 4.4: geometry of generic identifiability with $Q_{4\times 2} = (I_2;\, I_2)$.

Before stating the strict and generic identifiability results on $(Q, \mathbf{\Theta}, \boldsymbol{p})$, we show in the next proposition that any all-zero row vector in the Q-matrix can be dropped without affecting the identifiability conclusion.

Proposition IV.1. Suppose the Q-matrix of size $J \times K$ takes the form $Q = ((Q')^\top, \mathbf{0}^\top)^\top$, where Q' is a $J' \times K$ submatrix containing J' nonzero $\boldsymbol{q}$-vectors, and $\mathbf{0}$ denotes a $(J - J') \times K$ submatrix containing these zero $\boldsymbol{q}$-vectors. Let $\mathbf{\Theta}'$ be the submatrix of $\mathbf{\Theta}$ containing its first J' rows. Then, for any SLAM, $(Q, \mathbf{\Theta}, \boldsymbol{p})$ are jointly strictly (generically) identifiable if and only if $(Q', \mathbf{\Theta}', \boldsymbol{p})$ are jointly strictly (generically) identifiable.

Therefore, without loss of generality, from now on, we only consider Q-matrices without any zero $\boldsymbol{q}$-vectors when discussing joint identifiability. We examine various SLAMs that are popular in cognitive diagnosis assessment. In particular, in Section 4.2, we present the sufficient and necessary conditions for the strict and generic identifiability of $(Q, \mathbf{\Theta}, \boldsymbol{p})$ under the basic DINA model. These identifiability results can also be applied to the DINO model (Templin and Henson, 2006), owing to the duality between the two models (Chen et al., 2015). Section 4.3 presents the sufficient and necessary conditions for the generic identifiability of $(Q, \mathbf{\Theta}, \boldsymbol{p})$ under multiparameter SLAMs, which include the popular GDINA and LCDM models.

4.2 Identifiability of $(Q, \mathbf{\Theta}, \boldsymbol{p})$ under the DINA Model

Under the DINA model, Liu et al. (2013) first studied the identifiability of the Q-matrix under the assumption that the guessing parameters $\boldsymbol{g}$ are known. Chen et al. (2015) and Xu and Shang (2018) proposed a further set of sufficient conditions without needing to assume known item parameters. An important requirement in these identifiability studies is the completeness of the Q-matrix (Chiu et al., 2009). Under the DINA model, the Q-matrix is said to be complete if it contains a $K \times$

K identity submatrix I_K up to column permutation. Chen et al. (2015) and Xu and Shang (2018) require Q to contain at least two complete submatrices I_K for identifiability.

However, determining the minimal requirements on the Q-matrix for identifiability remains an open problem. In the next theorem, we solve this problem by providing the necessary and sufficient condition for the identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ under the earlier assumption that $p_{\boldsymbol{\alpha}} > 0$, for all $\boldsymbol{\alpha} \in \{0,1\}^K$ (Xu and Zhang, 2016; Gu and Xu, 2019b).

Theorem IV.1. *Under the DINA model, the combination of Conditions A, B, and C is necessary and sufficient for the strict identifiability of* $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$*:*

A. The true Q-matrix is complete. Without loss of generality, assume the Q-matrix takes the following form:

$$Q = \begin{pmatrix} I_K \\ Q^\star \end{pmatrix}. \tag{4.4}$$

B. The column vectors of the submatrix $Q^\star$ in (4.4) are distinct.

C. Each column in Q contains at least three entries equal to one.

In the Supplementary Material, we provide simulations that verify Theorem IV.1. In particular, see simulation study I for the sufficiency of Conditions A, B, and C for joint identifiability; also see simulation studies III and IV for the necessity of the proposed conditions. Next, we compare our Theorem 1 with several existing results. First, although the same set of conditions is proposed in Gu and Xu (2019b), they assumed *a known* Q when examining the identifiability of the parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$. In contrast, Theorem 1 studies the joint identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$, which is theoretically much more challenging, owing to the unknown Q-matrix, and therefore provides a much stronger result than that in Chapter II (Gu and Xu, 2019b). In terms of estimation, Theorem IV.1 implies that we can consistently estimate both Q

and $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$, without worrying that an incorrect Q-matrix is indistinguishable from the true Q. Second, Theorem IV.1 has much weaker requirements than those of the well-known identifiability conditions resulting from a three-way tensor decomposition (Kruskal, 1977; Allman et al., 2009). Specifically, these classical results require that the number of items $J \geq 2K + 1$ for (generic) identifiability. In contrast, the conditions in Theorem IV.1 imply that we need the number of items J to be at least $K + \lceil \log_2(K) \rceil + 1$ under the DINA model. This is because, other than the identity submatrix I_K, in order to satisfy Condition B of *distinctness*, the Q-matrix needs only contain a further $\log_2(K)$ items whose K-dimensional $\boldsymbol{q}$-vectors form a matrix with K distinct columns. For example, for $K = 8$, the conditions in Allman et al. (2009) require at least $2K + 1 = 17$ items, whereas our Theorem IV.1 guarantees that the following Q with $K + \log_2(K) + 1 = 12$ items suffices for the strict identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ under DINA:

$$Q = \begin{pmatrix} & & & I_8 & & & & \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Conditions A, B, and C are the minimal requirements for joint strict identifiability. When the true Q fails to satisfy one or more of these, Theorem 1 implies that there must exist $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) \not\sim (\bar{Q}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}})$ such that (5.9) holds. In this scenario, there are still cases where the model is "almost surely" identifiable, though not strictly identifiable, as illustrated by the example under $Q_{4\times 2}$ in (5.2). On the other hand, there are also cases where the entire model is never identifiable, as shown in simulation studies III and IV in the Supplementary Material. Therefore, it is desirable to determine which conditions guarantee the generic identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$.

In the following, we discuss the necessity of Conditions A, B, and C under the weaker notion of generic identifiability. First, Condition A is necessary for the joint generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$. If the true Q-matrix does not satisfy Condition A, then under the DINA model, certain latent classes would be equivalent given Q, and their separate proportion parameters can never be identified, not even generically (Gu and Xu, 2020a). In certain scenarios where Condition A fails, one can find a different $\bar{Q}$ that is not distinguishable from Q. Simulation study IV in the Supplementary Material illustrates the necessity of Condition A.

Second, Condition B is also difficult to relax, and serves as a necessary condition for generic identifiability when $K = 2$. Specifically, as shown in Gu and Xu (2019b), when $K = 2$, the only possible structure of the Q-matrix that violates Condition B while satisfying Conditions A and C is

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

In addition, in Chapter II we prove that for *any* valid DINA parameters associated with this Q, there exist infinitely many different sets of DINA parameters that lead to the same distribution of the responses. Therefore, the model is not generically identifiable.

Third, in contrast to Conditions A and B, for generic identifiability, Condition C can be relaxed to a certain extent. The next theorem characterizes how the Q-matrix structure in this case affects generic identifiability. For an empirical verification of Theorem IV.2, see simulation study II in the Supplementary Material.

Theorem IV.2. *Under the DINA model,* $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ *is not generically identifiable if* *some attribute is required by only one item.*

If some attribute is required by only two items, suppose the Q-matrix takes the following form, after some column and row permutations:

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \boldsymbol{v}^\top \\ \mathbf{0} & Q^\star \end{pmatrix}, \tag{4.5}$$

where $\boldsymbol{v}$ is a vector of length $K-1$, and $Q^\star$ is a $(J-2)\times(K-1)$ submatrix.

(a) If $\boldsymbol{v} = \mathbf{1}$, $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not locally generically identifiable.

(b) If $\boldsymbol{v} = \mathbf{0}$, $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is globally generically identifiable if either

(b.1) the submatrix $Q^\star$ satisfies Conditions A, B, and C in Theorem IV.1; or

(b.2) the submatrix $Q^\star$ has two submatrices I_{K-1}.

(c) If $\boldsymbol{v} \neq \mathbf{0}, \mathbf{1}$, $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is locally generically identifiable if $Q^\star$ satisfies Conditions A, B, and C in Theorem IV.1.

Remark IV.1. We say $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is locally identifiable if, in a neighborhood of the true parameters, there does not exist a different set of parameters that gives the same distribution of the responses. Local generic identifiability is a weaker notion than (global) generic identifiability. Therefore, the statement in part (a) of Theorem IV.2 also implies that $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not globally generically identifiable.

Remark IV.2. In scenario (b.1) of Theorem IV.2, the identifiable subset of the parameter space is $\big\{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \exists \boldsymbol{\alpha}^1 = (0, \alpha_2^1, \ldots, \alpha_K^1), \boldsymbol{\alpha}^2 = (0, \alpha_2^2, \ldots, \alpha_K^2) \in \{0\} \times \{0,1\}^{K-1}$, such that $p_{\boldsymbol{\alpha}^1} p_{\boldsymbol{\alpha}^2 + \boldsymbol{e}_1} \neq p_{\boldsymbol{\alpha}^2} p_{\boldsymbol{\alpha}^1 + \boldsymbol{e}_1}\big\}$, where $\boldsymbol{e}_j$ is a J-dimensional unit vector, with the jth element equal to one and all the others zero. In scenario (b.2) of Theorem IV.2, we can write $Q = (I_K, I_K, (Q^{\star\star})^\top)^\top$, in which case, the identifiable subset is $\big\{(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p}) : \forall k \in \{1, \ldots, K\}, \exists \boldsymbol{\alpha}^{k,1}, \boldsymbol{\alpha}^{k,2} \in \{0,1\}^{k-1} \times \{0\} \times \{0,1\}^{K-k-1}$, such that

$p_{\boldsymbol{\alpha}^{k,1}} p_{\boldsymbol{\alpha}^{k,2}+\boldsymbol{e}_k} \neq p_{\boldsymbol{\alpha}^{k,2}} p_{\boldsymbol{\alpha}^{k,1}+\boldsymbol{e}_k}\}$. The complements of these identifiable subsets in the parameter space give the nonidentifiable subsets, which are both of measure zero in the DINA model parameter space.

Next we discuss the generic identifiability of the DINA model in the special case of $K = 2$. We have the following proposition.

Proposition IV.2. Under the DINA model with $K = 2$ attributes, $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is generically identifiable if and only if the conditions in Theorem IV.1 or IV.2(b) hold.

Proposition IV.2 gives a full characterization of joint generic identifiability when $K = 2$, showing that the proposed generic identifiability conditions are necessary and sufficient in this case. The following example discusses all possible Q-matrices with $K = 2$, such that $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not strictly identifiable, which proves Proposition IV.2 automatically.

Example IV.1. When $K = 2$, the discussions on Conditions A and B before Theorem IV.2 show that $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not generically identifiable when A or B is violated. Therefore, we need only focus on cases where Condition C is violated and Conditions A and B are satisfied. Specifically, when $J \leq 5$, the Q-matrix can only take the following forms up to column and row permutations:

$$Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad Q_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

By Theorem IV.2, Q_1 falls in scenario (a); therefore, $(Q_1, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ is not locally generically identifiable; that is, even in a small neighborhood of the true parameters, there exist infinitely many different sets of parameters that give the same distribution of the responses. On the other hand, Q_2 falls in scenario (b.2) and Q_3 falls in scenario (b.1). Therefore, $(Q_2, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ and $(Q_3, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are both generically identifiable. In the case of $J > 5$, any Q satisfying A and B while violating C must contain one of the above Q_i as a submatrix and include additional row vectors of $(0, 1)$. By Theorem IV.2, any such Q extended from Q_1 is still not locally generically identifiable, and any such Q extended from Q_2 or Q_3 is globally generically identifiable.

4.3 Identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under General SLAMs

Because the DINA model is a submodel of multiparameter SLAMs, Conditions A, B, and C in Theorem IV.1 are also necessary for the strict identifiability of multiparameter SLAMs. For instance, our proposed Conditions A, B, and C are weaker than the sufficient conditions proposed by Xu and Shang (2018) for the strict identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under multiparameter SLAMs; and if their conditions are satisfied, the current conditions A, B, and C are also satisfied. However, these necessary requirements may be strong in practice, and cannot be applied to identify any Q that lacks some single-attribute items (i.e., lacks some unit vector as a row vector). A natural question is whether Conditions A, B, and C can be relaxed under the weaker notation of of generic identifiability. This section addresses this question.

Under multiparameter SLAMs, the next theorem shows that Condition C (each attribute is required by at least three items) is necessary for the generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$, contrary to the results for the DINA model, where Conditions A and B cannot be relaxed, but Condition C can. Simulation studies VI and VII in the Supplementary Material verify Theorem IV.3.

Theorem IV.3. *Under a multiparameter SLAM, Condition C in Theorem 1 is necessary for the generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$. Specifically, when the true Q-matrix violates C, for any model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ associated with Q, there exist infinitely* *many sets of* $(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) \nsim (Q, \boldsymbol{\Theta}, \boldsymbol{p})$ *such that equation* (5.9) *holds. Thus,* $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ *is not generically identifiable.*

Whereas Condition C is necessary, we next show that the other two conditions, A and B, can be relaxed further for the generic identifiability of multiparameter SLAMs. Before stating the result, we first introduce a new concept about the Q-matrix, called *generic completeness.*

Definition IV.3 (Generic Completeness)**.** A Q-matrix with K attributes is said to be generically complete if, after some column and row permutations, it has a $K \times K$ submatrix with all diagonal entries equal to one.

Generic completeness is a relaxation of the concept of completeness. In particular, a Q-matrix is generically complete if, up to column and row permutations, it contains a submatrix as follows:

$$\begin{pmatrix} 1 & * & \dots & * \\ * & 1 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & 1 \end{pmatrix},$$

where the off-diagonal entries "$*$" are left unspecified. Note that any complete Q-matrix is also generically complete, whereas a generically complete Q-matrix may not have any single-attribute items.

Using the concept of generic completeness, the next theorem gives sufficient conditions for joint generic identifiability, and shows that under multiparameter SLAMs, the necessary conditions A and B for strict identifiability are no longer necessary in the current setting.

Theorem IV.4. *Under a general SLAM, if the true* Q*-matrix satisfies the following Conditions D and E, then* $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ *is generically identifiable.*

D. The Q-matrix has two nonoverlapping generically complete $K \times K$ submatrices Q_1 and Q_2. Without loss of generality, assume the Q-matrix is in the following form:

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q^\star \end{pmatrix}_{J \times K}. \tag{4.6}$$

E. Each column of the submatrix $Q^\star$ in (5.5) *contains at least one entry of one.*

Remark IV.3. Under Theorem IV.4, the identifiable subset of the parameter space is $\{(\boldsymbol{\Theta}, \boldsymbol{p}) : \det(T(Q_1, \boldsymbol{\Theta}_{Q_1})) \neq 0,\ \det(T(Q_2, \boldsymbol{\Theta}_{Q_2})) \neq 0$, and $T(Q^\star, \boldsymbol{\Theta}_{Q^\star}) \cdot \mathrm{Diag}(\boldsymbol{p})$ has distinct column vectors$\}$. Its complement is the nonidentifiable subset, and it has measure zero in the parameter space $\boldsymbol{\vartheta}_Q$ when Q satisfies Conditions D and E. Please see the supplementary materials for the definition of the T-matrices ($T(Q_1, \boldsymbol{\Theta}_{Q_1})$, etc.).

Remark IV.4. The proof of Theorem IV.4 is based on the proof of Theorem 7 in Gu and Xu (2020a), who proposed the same Conditions D and E as sufficient conditions for the generic identifiability of model parameters, given a known Q. We point out that though D and E serve as sufficient conditions for generic identifiability, both when Q is known and when Q is unknown, the generic identifiability results in these two scenarios are different. In particular, Theorem 8 in Gu and Xu (2020a) shows that when Q is known, some attribute can be required by only two items for generic identifiability to hold (i.e., Condition C can be relaxed); in contrast, our current Theorem IV.3 shows that when Q is unknown, Condition C indeed becomes necessary.

The proposed sufficient Conditions D and E weaken the strong requirement of Conditions A and B, especially the identity submatrix requirement that may be difficult to satisfy in practice. Simulation study V in the Supplementary Material verifies Theorem IV.4. Note that Conditions D and E imply the necessary Condition C that each attribute is required by at least three items.

We next discuss the necessity of Conditions D and E. As shown in Section 3.2, under DINA, the completeness of Q is necessary for the joint strict identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$. For multiparameter SLAMs, we have an analogous conclusion that the generic completeness of Q, which is part of Condition D, is necessary for the joint generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$. This is stated in the next theorem.

Theorem IV.5. *Under a general SLAM, generic completeness of the Q-matrix is necessary for the joint generic identifiability of* $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$.

Furthermore, we show that Conditions D and E themselves are in fact necessary when $K = 2$, indicating the difficulty of relaxing these further.

Proposition IV.3. For a general SLAM with $K = 2$, Conditions D and E are necessary and sufficient for the generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$.

We use the following example to illustrate the result of Proposition IV.3, which also gives a natural proof of the proposition.

Example IV.2. When $K = 2$, a Q-matrix that satisfies the necessary Condition C, but not Conditions D or E, can only take the following form Q_1 or Q_2, up to row permutations:

$$Q_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} 1 & * \\ * & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}; \qquad \bar{Q}_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The "$*$"s in Q_2 are unspecified values, and can be either zero or one. For Q_1 with $J = 3$, $K = 2$, and any parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$, there are $2^J = 8$ constraints in (5.9) for solving $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ under Q_1 itself, whereas the number of free parameters of $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ is $|\{p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^2\} \cup \{\theta_{j,\boldsymbol{\alpha}} : j \in \{1,2\}, \boldsymbol{\alpha} \in \{0,1\}^2\}| = 2^K + 2^K \times J = 16 > 8$. For

Q_2 with $J = 4$, $K = 2$, and any associated $(\mathbf{\Theta}, \boldsymbol{p})$, there are $2^J = 16$ constraints in (5.9) for solving $(\bar{\mathbf{\Theta}}, \bar{\boldsymbol{p}})$, whereas the number of free parameters of $(\bar{\mathbf{\Theta}}, \bar{\boldsymbol{p}})$ under the alternative $\bar{Q}_2$ is $2^K + J \times 2^K = 20 > 2^J = 16$. In both cases, there are infinitely many sets of solutions of (5.9) as alternative model parameters. Therefore, neither $(Q_1, \mathbf{\Theta}, \boldsymbol{p})$ nor $(Q_2, \mathbf{\Theta}, \boldsymbol{p})$ are generically identifiable.

4.4 Discussion

In this chapter, we study the identifiability issue of SLAMs with unknown Q-matrices. For the basic DINA model, we derive the necessary and sufficient conditions for the strict joint identifiability of the Q-matrix and the associated model parameters. We also study a slightly weaker identifiability notion, called generic identifiability, and propose sufficient and necessary conditions for it under the DINA model and multiparameter SLAMs.

Statistical consequences of identifiability. In the setting of SLAMs, identifiability naturally leads to estimability, in different senses, under strict and generic identifiability. If the Q-matrix and the associated model parameters are strictly identifiable, then Q and the model parameters can consistently be jointly estimated from the data. If the Q-matrix and the model parameters are generically identifiable, then for true parameters ranging almost everywhere in the parameter space with respect to the Lebesgue measure, the Q-matrix and the model parameters can consistently be jointly estimated from the data.

As pointed out by one reviewer, the analysis of identifiability is under an ideal situation with an infinite sample size. Indeed, general identification problems assume the hypothetical exact knowledge of the distribution of the observed variables, and ask under what conditions one can recover the underlying parameters (Allman et al., 2009). Next, we discuss the finite-sample estimation issue under the proposed identifiability conditions for strict identifiability, following a similar argument to that in Proposition 1 in Xu and Shang (2018). Denote the true Q-matrix and model parameters by Q^0 and $\boldsymbol{\eta}^0 = (\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$, respectively. Consider a sample with N independent and identically distributed (i.i.d.) response vectors $\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_N$, and denote the log-likelihood of the sample by $\ell(\boldsymbol{\Theta}, \boldsymbol{p}) = \sum_{i=1}^{N} \log \mathbb{P}(\boldsymbol{R}_i \mid Q, \boldsymbol{\Theta}, \boldsymbol{p})$. Under a specified SLAM, a Q-matrix determines the structure of the item parameter matrix $\boldsymbol{\Theta}$ by specifying which entries are equal. For a given $\boldsymbol{\Theta}$, we can define an equivalent formulation of it, a sparse matrix $\boldsymbol{B}$, with the same size as $\boldsymbol{\Theta}$, as follows. Under a general SLAM, such as the GDINA model in Example I.3, the item parameters can be parameterized as $\theta_{j,\boldsymbol{\alpha}} = \sum_{\mathcal{S} \subseteq \{1,\ldots,K\}} \beta_{j,\mathcal{S}} \prod_{k \in \mathcal{S}} \alpha_k$. Based on this, we define the jth row of $\boldsymbol{B}$ as a 2^K-dimensional vector collecting all of these β-coefficients; that is, $\boldsymbol{B}_j = (\beta_{j,0}, \beta_{j,1}, \ldots, \beta_{j,K}, \ldots, \beta_{j,12\cdots K})$. Then, as long as the $\boldsymbol{q}$-vector $\boldsymbol{q}_j \neq \mathbf{1}_K$, the vector $\boldsymbol{B}_j$ and the matrix $\boldsymbol{B}$ are both "sparse". For the true Q^0, we denote the corresponding $\boldsymbol{B}$-matrix by $\boldsymbol{B}^0$. Under a specified SLAM (e.g., DINA or GDINA), the identification of Q^0 is then implied by the identification of the indices of nonzero elements of $\boldsymbol{B}^0$. Denote the support of the true $\boldsymbol{B}^0$ and any candidate $\boldsymbol{B}$ by S_0 and S, respectively. Define $C_{\min}(\boldsymbol{\eta}^0) = \inf_{\{S \neq S_0,\, |S| \leq |S_0|\}} (|S_0 \setminus S|)^{-1} h^2(\boldsymbol{\eta}^0, \boldsymbol{\eta})$, where $h^2(\boldsymbol{\eta}^0, \boldsymbol{\eta})$ denotes the Hellinger distance between the two distributions of $\boldsymbol{R}$, indexed by parameters $\boldsymbol{\eta}^0$ under the true $\boldsymbol{B}^0$, and by $\boldsymbol{\eta}$ under the candidate $\boldsymbol{B}$. Denote the Q-matrix and the model parameters that maximize the log-likelihood $\ell(\boldsymbol{\Theta}, \boldsymbol{p})$ subject to the L_0 constraint $|S| \leq |S_0|$ by $\widehat{\boldsymbol{\eta}} = (\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$, and denote the "oracle" MLEs of the model parameters obtained, assuming Q^0 is known, by $\widehat{\boldsymbol{\eta}}^0 = (\widehat{\boldsymbol{\Theta}}^0, \widehat{\boldsymbol{p}}^0)$. Then, we have the following finite-sample error bound for the estimated Q-matrix and model parameters.

Proposition IV.4. Suppose Q^0 satisfies the proposed sufficient conditions for joint strict identifiability; then, $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) \geq c_0$, for some positive constant c_0. Furthermore,

$$\mathbb{P}(\widehat{Q} \not\sim Q^0) \le \mathbb{P}(\widehat{\boldsymbol{\eta}} \ne \widehat{\boldsymbol{\eta}}^0) \le c_2 \exp\{-c_1 N C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)\}, \tag{4.7}$$

where $c_1, c_2 > 0$ are some constants. That is, when the joint strict identifiability conditions hold, the finite-sample estimation error has an exponential bound.

Proposition IV.4 shows that the estimation error decreases exponentially in N if the model is identifiable. On the other hand, when the identifiability conditions fail to hold, there exist alternative models that are close to the true model in terms of the Hellinger distance. This would make the $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ in (4.7) equal to zero, instead of being bounded away from zero, as shown in Proposition IV.4. Therefore, the finite-sample error bound in (4.7) becomes $O(1)$ in this nonidentifiable scenario. In particular, when the generic identifiability conditions are satisfied, $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ depends on the distance between the true parameters and the nonidentifiable measure-zero subset of the parameter space; as the true parameters become closer to this measure-zero set, $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ decreases to zero, and a larger sample size may be needed to achieve a prespecified level of estimation accuracy.

Potential extensions to other latent variable models. We briefly discuss potential extensions of the proposed theory to other latent variable models, such as SLAMs with ordinal polytomous attributes (von Davier, 2008; Ma and de la Torre, 2016; Chen and de la Torre, 2018), and multidimensional latent trait models (Embretson, 1991). First, an SLAM with ordinal polytomous attributes can be viewed as an SLAM with binary attributes and a constrained relationship among the binary attributes. For instance, consider an ordinal attribute γ that can take C different values $\{0, 1, \dots, C-1\}$; then, γ can be equivalently viewed as a collection of $C-1$ binary random variables $\boldsymbol{\alpha}^\gamma := (\alpha_1, \dots, \alpha_{C-1})$ with the following constraints. If $\alpha_i = 0$ for some $i < C-1$, then $\alpha_j = 0$, for all $j = i+1, \dots, C-1$. In other words, any pattern $\boldsymbol{\alpha}^\gamma$ with $\alpha_i = 0$ and $\alpha_j = 1$, for some $i < j$ is "forbidden" and constrained to have proportion zero. The vector of polytomous attributes can be augmented to a longer vector of binary attributes using constraints in this fashion. Then, we can consider the SLAM with the augmented proportion parameters by constraining the proportions of the "forbidden" binary attribute patterns to zero. In this scenario, it might be possible to extend the current theory and develop identifiability conditions for the case of polytomous attributes.

Second, if a multidimensional latent trait model includes both continuous and discrete latent traits, then the techniques used to establish the identifiability of the latent class models in this study would also be useful when treating discrete latent variables. For continuous latent variables, the techniques developed in Bai and Li (2012) for the identifiability of the factor analysis model and those developed for traditional multivariate analyses (Anderson, 2009) would be helpful.

In practice, the proposed identifiability theory can serve as a foundation for designing statistically guaranteed estimation procedures. Specifically, consider the set of all Q-matrices that satisfy our identifiability conditions (A, B, and C under the DINA model, or D and E under multiparameter SLAMs), and call it the "identifiable Q-set." Then, we can use likelihood-based approaches, such as that in Xu and Shang (2018), to jointly estimate Q and the model parameters by constraining Q to the identifiable Q-set; alternatively we can use Bayesian approaches to estimate Q, as in Chen et al. (2018a). Additionally, if under the DINA model, the Q-matrix does not contain a submatrix I_K, then according to Chapter III, certain attribute profiles would be equivalent and the strongest possible identifiability argument therein is the so-called $\boldsymbol{p}$-partial identifiability. In this scenario, it would be interesting to study the identifiability of the incomplete Q-matrix under the notion of $\boldsymbol{p}$-partial identifiability. We leave this to future research.

CHAPTER V

Learning Attribute Patterns in High-Dimensional Structured Latent Attribute Models

The previous Chapters II–IV on the identifiability theory provide easily checkable conditions that guarantee the identifiability and estimability of model parameters and latent structures. These lay the solid foundation for performing subsequent estimation tasks of SLAMs. In the modern data science era, data exhibits an increasingly large volume and complex structure. To rise to these challenges, the remaining part of this dissertation is devoted to developing novel statistical methods and efficient algorithms to tackle combinatorial estimation problems of SLAMs in high-dimensional settings.

One challenge in modern applications of SLAMs is that the number of discrete latent attributes could be large, leading to a high-dimensional space for all the possible configurations of the attributes, i.e., a high-dimensional space for latent attribute patterns. In many applications, the number of potential patterns is much larger than the sample size. For scientific interpretability and practical use, it is often assumed that not all the possible attribute patterns exist in the population. Examples with a large number of potential latent patterns and a moderate sample size can be found in educational assessments (Lee et al., 2011; Choi et al., 2015; Yamaguchi and

This chapter contains the main part of Gu and Xu (2019a), accepted by *Journal of Machine Learning Research*.

Okada, 2018) and the epidemiological diagnosis of disease etiology (Wu et al., 2016, 2017; O'Brien et al., 2019). For instance, a dataset from Trends in International Mathematics and Science Study (TIMSS), which has 13 binary latent attributes (i.e., $2^{13} = 8192$ possible latent attribute patterns) while only 757 students' responses are observed; see Example V.1 in Section 5.1 for details. In cognitive diagnosis, it is of interest to select the significant attribute patterns among these $2^{13} = 8192$ ones. In such high-dimensional scenarios, existing estimation methods often tend to over select the number of latent patterns, and may not scale to datasets with a huge number of patterns. Moreover, theoretical questions remain open on whether and when the "sparse" latent attribute patterns are identifiable and can be consistently learned from data.

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

Our contributions in this chapter contain the following aspects. First, we characterize the identifiability requirement needed for a SLAM with an arbitrary subset of attribute patterns to be learnable, and establish mild identifiability conditions.

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

The rest of the chapter is organized as follows. Section 5.1 investigates the learnability requirement and proposes mild sufficient conditions for learnability. Section 5.3 proposes the estimation methodology and establishes theoretical guarantee for the proposed methods. Section 5.4 and Section 5.5 include simulations and real data analysis, respectively. The proofs of all the theoretical results and additional experimental results are included in Appendix D.

5.1 Motivation for Latent Pattern Selection

One challenge in modern applications of SLAMs is that the number of potential latent attribute patterns 2^K increases exponentially with K and could be much larger than the sample size N. It is often assumed that a relatively small portion of attribute patterns exist in the population. We give a specific example as follows.

Example V.1. Trends in International Mathematics and Science Study (TIMSS) is a large scale cross-country educational assessment. TIMSS evaluates the mathematics and science abilities of fourth and eighth graders every four years since 1995. Researchers have used SLAMs to analyze the TIMSS data (e.g., Lee et al., 2011; Choi et al., 2015; Yamaguchi and Okada, 2018). For example, a 23×13 Q-matrix constructed by mathematics educators was specified for the TIMSS 2003 eighth grade mathematics assessment (Su et al., 2013). Table 5.1 presents the Q-matrix.

Table 5.1: Q-matrix in Su et al. (2013) for TIMSS 2003 8th Grade Data

Item ID	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}	α_{13}
1	1	0	0	0	0	0	0	0	0	0	1	0	1
2	0	0	0	0	0	1	0	0	0	0	0	0	0
3	0	1	0	0	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	1	0	0	0	0
5	0	0	0	0	0	1	0	0	0	1	0	1	0
6	0	0	0	0	0	1	1	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	1	0	0	0	0	0	1	0	0	0	0	0
12	0	0	1	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	1	0	0	0	0	0	0	0	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0	0	0	0	1	0
16	0	0	0	0	1	0	0	0	0	0	0	0	0
17	0	0	0	1	0	0	0	0	0	0	0	0	0
18	0	0	1	0	0	0	0	0	1	0	1	0	1
19	0	1	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	1	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	1	0	0	0	0	1	0	0	0	0

Example V.1 has $2^K = 2^{13} = 8192$ different configurations of attribute patterns; for the limited sample size 757 there, it is desirable to learn the potentially small set of significant attribute patterns from data.

Another motivation for assuming a small number of attribute patterns exist in the population results from the possible hierarchical structure among the targeted attributes. For instance, in an educational assessment of a set of underlying latent skill attributes, some attributes often serve as prerequisites for some others (Leighton et al., 2004; Templin and Bradshaw, 2014). Specifically, the prerequisite relationship depicts the different level of difficulty of the skill attributes, and also reveals the order in which these skills are learned in the population of students. For instance, if attribute α_1 is a prerequisite for attribute α_2, then the attribute pattern $(\alpha_1 = 0, \alpha_2 = 1)$ does not exist in the population, naturally resulting in a sparsity structure of the existence of attribute patterns. When the number of attributes is large and the underlying hierarchy structure is complex and unknown, it is desirable to learn the hierarchy of attributes directly from data. In such cases with attribute hierarchy, the number of patterns respecting the hierarchy could be far fewer than 2^K.

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

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

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

In the following sections, we first investigate the learnability requirement of learning a SLAM with an arbitrary set of true latent patterns, and provide identifiability conditions in Section 5.2. Then in Section 5.3, we propose a penalized likelihood method to select the latent attribute patterns, and establish theoretical guarantee for the proposed method.

5.2 Learnability Requirement and Conditions

To facilitate the discussion on identifiability of SLAMs, we need to introduce a new notation, the Γ-matrix. We first introduce the $J \times 2^K$ constraint matrix Γ^{all} that is entirely determined by the Q-matrix. The rows of Γ^{all} are indexed by the J items, and columns by the 2^K latent attribute patterns in $\{0,1\}^K$. The $(j, \boldsymbol{\alpha})$th entry of $\Gamma^{\text{all}}_{j,\boldsymbol{\alpha}}$ is defined as

$$\Gamma^{\text{all}}_{j,\boldsymbol{\alpha}} = I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j) = I(\boldsymbol{\alpha} \in \mathcal{C}_j), \quad j \in \{1, \ldots, J\},\ \boldsymbol{\alpha} \in \{0,1\}^K, \tag{5.1}$$

which is a binary indicator of whether attribute pattern $\boldsymbol{\alpha}$ possess all the required attributes of item j. We will also call Γ^{all} the *constraint matrix*, since its entries indicate what latent patterns are constrained to have the highest level of Bernoulli parameters for each item. For example, consider the 2×2 Q-matrix in the following (5.2). Then its corresponding Γ-matrix Γ^{all} with a saturated set of attribute patterns takes the following form.

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \implies \Gamma^{\text{all}} = \begin{array}{c} \begin{array}{cccc} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3 & \boldsymbol{\alpha}_4 \\ (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} \quad 0 \quad & \quad 1 \quad & \quad 0 \quad & \quad 1 \quad \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}. \tag{5.2}$$

More generally, we generalize the definition of the constraint matrix Γ^{all} in (5.1) to an arbitrary subset of latent patterns $\mathcal{A} \subseteq \{0,1\}^K$, and an arbitrary set of items $S \subseteq [J]$. For $S \subseteq [J]$ and $\mathcal{A} \subseteq \{0,1\}^K$, we simply denote by $\Gamma^{(S,\mathcal{A})}$ the $|S| \times |\mathcal{A}|$ submatrix of Γ^{all} with row indices from S and column indices from $\mathcal{A}$. When $S = \{1, \ldots, J\}$, we will sometimes just denote $\Gamma^{(S,\mathcal{A})}$ by $\Gamma^{\mathcal{A}}$ for simplicity. Then $\Gamma^{\mathcal{A}}$ itself can be viewed as the constraint matrix for a SLAM with attribute pattern space $\mathcal{A}$, and $\Gamma^{\mathcal{A}}$ directly characterizes how the items constrain the positive response probabilities of latent attribute patterns in $\mathcal{A}$.

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

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

Example V.2. Consider the 2×2 Q-matrix together with its corresponding 2×4 Γ-matrix in Equation (5.2). Consider two attribute pattern sets, the true set $\mathcal{A}_0 = \{\boldsymbol{\alpha}_1 = (0,0), \boldsymbol{\alpha}_2 = (0,1)\}$ and an alternative set $\mathcal{A}_1 = \{\boldsymbol{\alpha}_2 = (0,1), \boldsymbol{\alpha}_3 = (1,0)\}$. Under the two-parameter SLAM, for any valid item parameters $\boldsymbol{\Theta}$ restricted by Γ and any proportion parameters $\boldsymbol{p} = (p_{\boldsymbol{\alpha}_1}, p_{\boldsymbol{\alpha}_2}, p_{\boldsymbol{\alpha}_3}, p_{\boldsymbol{\alpha}_4})$ such that $p_{\boldsymbol{\alpha}_1} = p_{\boldsymbol{\alpha}_3}$, we have $\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}^{\mathcal{A}_0},\, (p_{\boldsymbol{\alpha}_1}, p_{\boldsymbol{\alpha}_2})) = \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{\Theta}^{\mathcal{A}_1},\, (p_{\boldsymbol{\alpha}_3}, p_{\boldsymbol{\alpha}_2}))$. This is because $\Gamma^{\mathcal{A}_0} = \Gamma^{\mathcal{A}_1}$ from (5.2) and hence $\boldsymbol{\Theta}^{\mathcal{A}_0} = \boldsymbol{\Theta}^{\mathcal{A}_1}$; and also $(p_{\boldsymbol{\alpha}_1}, p_{\boldsymbol{\alpha}_2}) = (p_{\boldsymbol{\alpha}_3}, p_{\boldsymbol{\alpha}_2})$ by our construction that $p_{\boldsymbol{\alpha}_1} = p_{\boldsymbol{\alpha}_3}$. This implies even if one knows exactly there are two latent attribute patterns in the population, one can never tell which two patterns those are based on the likelihood function. In this sense, $\mathcal{A}_0$ is not identifiable, due to the fact that $\Gamma^{\mathcal{A}_0}$ and $\Gamma^{\mathcal{A}_1}$ do not lead to distinguishable distributions of responses under the two-parameter SLAM.

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

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

$$\mathbb{P}(\boldsymbol{R} \mid \boldsymbol{\Theta}^{\mathcal{A}_0},\, \boldsymbol{p}^{\mathcal{A}_0}) = \mathbb{P}(\boldsymbol{R} \mid \boldsymbol{\Theta}^{\mathcal{A}},\, \boldsymbol{p}^{\mathcal{A}}) \tag{5.3}$$

implies $\mathcal{A} = \mathcal{A}_0$. Moreover, if (5.3) implies $(\boldsymbol{\Theta}^{\mathcal{A}},\, \boldsymbol{p}^{\mathcal{A}}) = (\boldsymbol{\Theta}^{\mathcal{A}_0},\, \boldsymbol{p}^{\mathcal{A}_0})$, then we say the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0},\, \boldsymbol{p}^{\mathcal{A}_0})$ are (strictly) identifiable.

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

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

A. There exist two disjoint item sets S_1 and S_2, such that $\Gamma^{(S_i,\mathcal{A}_0)}$ has distinct column vectors for $i = 1, 2$ and "$\succeq_{S_1} = \succeq_{S_2}$" under $\Gamma^{\mathcal{A}_0}$.

B. For any $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}' \in \mathcal{A}_0$ where $\boldsymbol{\alpha}' \succeq_{S_i} \boldsymbol{\alpha}$ under $\Gamma^{\mathcal{A}_0}$ for $i = 1$ or 2, there exists some $j \in (S_1 \cup S_2)^c$ such that $\Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}} \neq \Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}'}$.

C. Any column vector of $\Gamma^{\mathcal{A}_0}$ is different from any column vector of $\Gamma^{\mathcal{A}_0^c}$, where $\mathcal{A}_0^c = \{0,1\}^K \setminus \mathcal{A}_0$.

Recall that each column in the Γ-matrix corresponds to a latent attribute pattern, then Conditions A and B help ensure the Γ-matrix of the true patterns $\Gamma^{\mathcal{A}_0}$ contains enough information to distinguish between these true patterns. Specifically, Condition A requires $\Gamma^{\mathcal{A}_0}$ to contain two vertically stacked submatrices corresponding to item sets S_1 and S_2, each having distinct columns, i.e., each being able to distinguish between the true patterns; and Condition B requires the remaining submatrix of $\Gamma^{\mathcal{A}_0}$

to distinguish those pairs of true patterns that have some order ($\boldsymbol{\alpha}' \succeq_{S_i} \boldsymbol{\alpha}$) based on the first two item sets S_1 or S_2. Condition C is necessary for identifiability of $\mathcal{A}_0$ by ensuring that any true pattern would have a different column vector in Γ^{all} from that of any false pattern. Condition C is satisfied for any $\mathcal{A}_0 \subseteq \{0,1\}^K$ if the Q-matrix contains an identity submatrix I_K, because such a Q-matrix will give a Γ^{all} that has all the 2^K columns distinct.

We would like to point out that our identifiability conditions in Theorem V.1 do not depend on the unknown parameters (e.g., $\boldsymbol{\Theta}$ and $\boldsymbol{p}$), but only rely on the structure of the constraint matrix Γ. The Γ-matrix with respect to the true set of patterns $\mathcal{A}_0$ is the key quantity that defines the latent structure of a SLAM. Generally, it is hard to establish identifiability conditions that only depend on the cardinality of $\mathcal{A}_0$ but not on $\Gamma^{\mathcal{A}_0}$. For instance, in Example V.2, the two sets $\mathcal{A}_0$ and $\mathcal{A}_1$ have the same cardinality but can not be distinguished under the conditions there; indeed further conditions on Q (and the resulting Γ) are needed to guarantee identifiability.

The developed identifiability conditions generally apply to any SLAM satisfying the constraints (1.2) and (1.3) introduced in Chapter I. If one makes further assumptions on $\boldsymbol{\Theta}$, such as assuming each item $j \in [J]$ has exactly two item parameters to make it a two-parameter model, then the conditions in Theorem V.1 may be further relaxed. For example, in the saturated case with $\mathcal{A}_0 = \{0,1\}^K$, the sufficient identifiability conditions developed in Xu (2017) for a general SLAM require Q to contain two copies of I_K as submatrices, while the necessary and sufficient conditions established in Gu and Xu (2019b) for the two-parameter SLAM require Q to have just one submatrix I_K. We expect that in the current case with an arbitrary $\mathcal{A}_0 \subseteq \{0,1\}^K$, the conditions in Theorem V.1 can also be relaxed under the two-parameter model in a technically nontrivial way. For the reason of generality, we focus on SLAMs under the general constraints (1.2) and (1.3) in this work.

When the conditions in Theorem V.1 are satisfied, $\mathcal{A}_0$ is identifiable; and from

Theorem 4.1 in Gu and Xu (2020a), the model parameters $(\mathbf{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ associated with $\mathcal{A}_0$ are also identifiable.

Corollary V.1. *Under the conditions in Theorem V.1, the model parameters* $(\mathbf{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ *associated with* $\mathcal{A}_0$ *are identifiable.*

Note that the result of Theorem V.1 differs from the existing works Xu (2017), Xu and Shang (2018) and Gu and Xu (2020a) in that those works assume $\mathcal{A}_0$ is known *a priori* and study the identifiability of $(\mathbf{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$, while in the current work $\mathcal{A}_0$ is unknown and we focus on the identifiability of $\mathcal{A}_0$ itself. This is crucially needed in order to guarantee that we can learn the set of true attribute patterns.

Remark V.1. The identifiability results in Theorem V.1 and Corollary V.1 are related to the uniqueness of tensor decomposition. As shown in (1.5), the probability mass function of the multivariate responses of each subject can be viewed as a higher order tensor with constraints on entries of the tensor, and unique decomposition of the tensor correspond to identification of the constraint matrix as well as the model parameters. The identifiability conditions in Theorem V.1 are weaker than the general conditions for uniqueness of three-way tensor decomposition in Kruskal (1977), which is a celebrated result in the literature. Kruskal's conditions require the tensor can be decomposed as a Khatri-Rao product of three matrices, two having full-rank and the other having Kruskal rank at least two (Kruskal rank of a matrix is the largest number T such that every set of T columns of it are linearly independent). Consider an example with $J = 5$, $K = 2$, $\mathcal{A}_0 = \{\boldsymbol{\alpha}_2 = (0, 1),\ \boldsymbol{\alpha}_3 = (1, 0)\}$, and the corresponding $\Gamma^{\mathcal{A}_0}$ in the form of (5.4). Then we can set $S_1 = \{1, 2\}$, $S_2 = \{3, 4\}$ and Condition A in Theorem V.1 is satisfied. Further, Condition B is also satisfied since $\boldsymbol{\alpha}_2 \not\succeq_{S_i} \boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_3 \not\succeq_{S_i} \boldsymbol{\alpha}_2$ under $\Gamma^{\mathcal{A}_0}$. Therefore, Theorem 1 guarantees the set $\mathcal{A}_0$ is identifiable, and further guarantees the parameters $(\mathbf{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ are identifiable. On the contrary, results based on Kruskal's conditions for unique three-way tensor decomposition can not guarantee identifiability, because other than two full rank structures given by the items in S_1 and S_2, the remaining item 5 in $(S_1 \cup S_2)^c$ corresponds to a structure with Kruskal rank only one.

$$
Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \hline 1 & 0 \\ 0 & 1 \\ \hline 1 & 1 \end{pmatrix} \quad \Longrightarrow \quad \Gamma^{\mathcal{A}_0} = \begin{array}{cc} \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3 \\ (0,1) & (1,0) \end{array} \atop \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \hline 1 & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{pmatrix}. \tag{5.4}
$$

We next discuss two extensions of the developed identifiability theory. First, Theorem V.1 guarantees the strict learnability of $\mathcal{A}_0$. Under a multi-parameter SLAM, these conditions can be relaxed if the aim is to obtain the so-called generic joint identifiability of $\mathcal{A}_0$, which means that $\mathcal{A}_0$ is learnable with the true model parameters ranging almost everywhere in the constrained parameter space except a set with Lebesgue measure zero. Specifically, we have the following definition.

Definition V.2 (generic learnability of the true model)**.** Denote the parameter space of $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ constrained by $\Gamma^{\mathcal{A}_0}$ by Ω. We say $\mathcal{A}_0$ is generically identifiable, if there exists a subset $\mathcal{V}$ of Ω that has Lebesgue measure zero, such that for any $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0}) \in \Omega \setminus \mathcal{V}$, Equation (5.3) implies $\mathcal{A} = \mathcal{A}_0$. Moreover, if for any $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0}) \in \Omega \setminus \mathcal{V}$, Equation (5.3) implies $(\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) = (\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$, we say the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ are generically identifiable.

The generic learnability result is presented in the next theorem.

Theorem V.2 (conditions for generic learnability)**.** *Consider a multi-parameter SLAM with the set of true attribute patterns $\mathcal{A}_0$ and the $J \times |\mathcal{A}_0|$ constraint matrix $\Gamma^{\mathcal{A}_0}$. If*

$\Gamma^{\mathcal{A}_0}$ *satisfies Condition C and also the following conditions, then* $\mathcal{A}_0$ *is generically identifiable.*

$A^\star$. There exist two disjoint item sets S_1 and S_2, such that altering some entries from 0 to 1 in $\Gamma^{(S_1\cup S_2,\mathcal{A}_0)}$ can yield a $\widetilde{\Gamma}^{(S_1\cup S_2,\mathcal{A}_0)}$ satisfying Condition A. That is, $\widetilde{\Gamma}^{(S_i,\mathcal{A}_0)}$ has distinct columns for $i=1,2$ and " $\succeq_{S_1}$ " = " $\succeq_{S_2}$ " under $\widetilde{\Gamma}^{(S_1\cup S_2,\mathcal{A}_0)}$.

$B^\star$. For any $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}'\in\mathcal{A}_0$ where $\boldsymbol{\alpha}'\succeq_{S_i}\boldsymbol{\alpha}$ under $\widetilde{\Gamma}^{(S_1\cup S_2,\mathcal{A}_0)}$ for $i=1$ or 2, there exists some $j\in(S_1\cup S_2)^c$ such that $\Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}}\neq\Gamma^{\mathcal{A}_0}_{j,\boldsymbol{\alpha}'}$.

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

Corollary V.2. *If the Q-matrix satisfies the following conditions, then for any true set of attribute patterns $\mathcal{A}_0\subseteq\{0,1\}^K$ such that $\Gamma^{\mathcal{A}_0}$ satisfies Condition C, the set $\mathcal{A}_0$ is generically identifiable.*

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

$$Q=\begin{pmatrix}Q_1\\Q_2\\Q'\end{pmatrix}_{J\times K};\quad Q_i=\begin{pmatrix}1&*&\dots&*\\ *&1&\dots&*\\ \vdots&\vdots&\ddots&\vdots\\ *&*&\dots&1\end{pmatrix}_{K\times K},\quad i=1,2,\tag{5.5}$$

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

($B^{\star\star}$) With Q in the form of (5.5), there is $\sum_{j=2K+1}^{J}q_{j,k}\geq 1$ for each $k\in\{1,\dots,K\}$.

Remark V.2. When the conditions in Theorem V.2 are satisfied, $\mathcal{A}_0$ is generically identifiable and from Theorem 4.3 in Gu and Xu (2020a), the model parameters

$(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$ are also generically identifiable. Corollary V.2 differs from Theorem 4.3 in Gu and Xu (2020a) in that, here we allow the true set of attribute patterns $\mathcal{A}_0$ to be unknown and arbitrary, and study its identifiability, while Gu and Xu (2020a) assumes $\mathcal{A}_0$ is pre-specified and studies the identifiability of the model parameters $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0})$.

Remark V.3. Under the conditions for generic identifiability in Theorem V.2 or Corollary V.2, we can obtain the explicit forms of the measure zero set $\mathcal{V}$ ($\mathcal{V} \subseteq \Omega$) where the non-identifiability may occur. Under either Theorem V.2 or Corollary V.2, the set $\mathcal{V}$ is characterized by the zero set of certain polynomials about the parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ (see the proofs for details). The zero set of these polynomials indeed defines a lower-dimensional manifold in the parameter space. Therefore, Theorem V.2 and Corollary V.2 supplement Theorem V.1 by relaxing the original conditions and establishing identifiability when $(\boldsymbol{\Theta}, \boldsymbol{p})$ satisfy certain shape constraints, i.e., $(\boldsymbol{\Theta}, \boldsymbol{p})$ do not fall on that manifold $\mathcal{V}$ in the parameter space.

The above generic identifiability results of $\mathcal{A}_0$ ensure that nonidentifiability happens only in a measure zero set in the parameter space. Next, we develop a second extension of Theorem V.1 for scenarios where nonidentifiability cases occupy a positive measure set in the parameter space. This situation happens when certain latent attribute patterns always have the same item parameters across all the items, i.e., $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}} = \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}'}$ for some $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$. We define $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ to be in the same equivalence class if $\boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}} = \boldsymbol{\Theta}_{\bullet,\boldsymbol{\alpha}'}$. For instance, still consider the following 2×2 Q-matrix under the two-parameter SLAM introduced in Example I.1,

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \tag{5.6}$$

then attribute patterns $\boldsymbol{\alpha}_1 = (0,0)$ and $\boldsymbol{\alpha}_3 = (1,0)$ are equivalent under the two-parameter SLAM, as can be seen from the Γ^{all} in (5.2). Therefore the two latent patterns $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$ are not identifiable, no matter which values the true model parameters take.

In this case where both strict and generic identifiability do not hold, we study the $\boldsymbol{p}$-partial identifiability, a concept introduced in Gu and Xu (2020a). Specifically, when some attribute patterns have the same item parameters across all items, we define the set of these attribute patterns as an equivalence class, and aim to identify the proportion of this equivalence class, instead of the separate proportions of these equivalent patterns, in the population. For instance, in the above example in (5.6), because $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$ are equivalent, there are three equivalence classes: $\{\boldsymbol{\alpha}_1 = (0,0), \boldsymbol{\alpha}_3 = (1,0)\}$, $\{\boldsymbol{\alpha}_2 = (0,1)\}$, and $\{\boldsymbol{\alpha}_4 = (1,1)\}$. We denote these three equivalence classes by $[\boldsymbol{\alpha}_1]$ (or $[\boldsymbol{\alpha}_3]$, since $[\boldsymbol{\alpha}_1] = [\boldsymbol{\alpha}_3]$), $[\boldsymbol{\alpha}_2]$ and $[\boldsymbol{\alpha}_4]$, since $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_4$ form a complete set of representatives of the equivalence classes. For any Q, we denote the induced set of equivalence classes by $\mathcal{A}^{\text{equiv}} = \{[\boldsymbol{\alpha}_1], \ldots, [\boldsymbol{\alpha}_C]\}$, where $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C$ form a complete set of representatives of the equivalence classes. In this case, the pattern selection problem of interest is to learn which equivalence classes in $\mathcal{A}^{\text{equiv}}$ are significant.

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

$$\mathcal{A}_Q = \{\vee_{j \in S}\, \boldsymbol{q}_j : S \subseteq \{1, \ldots, J\}\}, \tag{5.7}$$

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

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

We have the following corollary of Theorem V.1 on identifiability, when certain attribute patterns are not distinguishable. Denote the set of significant equivalence classes by $\mathcal{A}_0^{\text{equiv}} = \{[\boldsymbol{\alpha}_{\ell_1}], \ldots, [\boldsymbol{\alpha}_{\ell_m}]\}$, which is a subset of the saturated set $\mathcal{A}^{\text{equiv}} = \{[\boldsymbol{\alpha}_1], \ldots, [\boldsymbol{\alpha}_C]\}$. Denote the set of representative patterns of the significant equivalence classes by $\{\boldsymbol{\alpha}_{\ell_1}, \ldots, \boldsymbol{\alpha}_{\ell_m}\} = \mathcal{A}^{\text{rep}}$.

Corollary V.3. *If the matrix $\Gamma^{\mathcal{A}^{\text{rep}}}$ satisfies Conditions A, B and C, $\mathcal{A}_0^{\text{equiv}}$ is identifiable.*

Remark V.4. Under the two-parameter SLAM with $\mathcal{A}^{\text{equiv}} = \{[\boldsymbol{\alpha}_1], \ldots, [\boldsymbol{\alpha}_C]\}$, the Γ-matrix $\Gamma^{\{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_C\}}$ by definition would have distinct column vectors. Therefore any column vector of $\Gamma^{\mathcal{A}^{\text{rep}}}$ in Corollary V.3 must be different form any column vector of $\Gamma^{\{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_C\}\setminus\mathcal{A}^{\text{rep}}}$. In this case, Condition C is automatically satisfied. And in order to identify $\mathcal{A}_0^{\text{equiv}}$, one only needs to check if $\Gamma^{\mathcal{A}^{\text{rep}}}$ satisfies Conditions A and B.

5.3 Penalized Likelihood Approach to Pattern Selection

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

5.3.1 Shrinkage Estimation

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

In general, we denote by $\mathcal{A}_{\text{input}}$ the set of candidate attribute patterns given to the shrinkage estimation method as input. If the saturated space of all the possible attribute patterns are considered, then $\mathcal{A}_{\text{input}} = \{0,1\}^K$ and it contains all the 2^K possible configurations of attributes. When $2^K \gg N$, we propose to use a preprocessing step that returns a proper subset $\mathcal{A}_{\text{input}}$ of the saturated set $\{0,1\}^K$ as candidate attribute patterns, and then perform the shrinkage estimation (please see Section 5.3.2 for the preprocessing procedure).

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

$(\boldsymbol{R}_1^\top, \dots, \boldsymbol{R}_N^\top)^\top$. The marginal likelihood can be written as

$$L(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R}) = \prod_{i=1}^{N} \Big[\sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{R_{i,j}} (1 - \theta_{j,\boldsymbol{\alpha}})^{1-R_{i,j}} \Big], \tag{5.8}$$

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

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

$$\ell^{\lambda}(\boldsymbol{\Theta}, \boldsymbol{p}) \; = \ell(\boldsymbol{\Theta}, \boldsymbol{p}) + \lambda \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} \log_{\rho_N}(p_{\boldsymbol{\alpha}}), \quad \lambda \in (-\infty, 0), \tag{5.9}$$

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

$$\rho_N \asymp N^{-d} \tag{5.10}$$

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

Remark V.5. In the literature, Chen et al. (2001) and Chen et al. (2004) used a similar form of penalty as the summation term in our (5.9), but instead imposed $\lambda > 0$ to avoid sparse solutions of the proportion parameters. These works used that penalty in order to avoid singularity when performing restricted likelihood ratio test. While our goal here is to encourage sparsity of $\boldsymbol{p}$ so that significant attribute patterns can be selected.

The formulation of (5.9) can also be interpreted in a Bayesian way, where the penalty term regarding the proportions $\boldsymbol{p}$ is the logarithm of the Dirichlet prior density with hyperparameter $\beta = \lambda + 1$ over the proportions. But note that when $\beta < 0$, the penalty term is not a proper prior density. Our later Proposition V.1 reveals that, under nonstandard convergence rate of the mixture model, the traditional Bayesian way of imposing a proper Dirichlet prior over proportions is not sufficient for selecting significant attribute patterns consistently. Instead, this classical procedure will yield too many false patterns being selected. Therefore, our novelty of allowing λ in (5.9) to be negative with arbitrarily large magnitude is crucial to selection consistency.

Other than the nice connection to the Dirichlet prior density in the Bayesian literature, the log-type penalty in (5.9) also facilitates the computation based on modified EM and variational EM algorithms, as shown in our Algorithms 1 and 2. For such reasons, this work uses the log-type penalty. There are also alternative ways of imposing penalty on the proportion parameters $\boldsymbol{p}$ that would lead to selection consistency, such as the truncated L_1 penalty used in Shen et al. (2012a) for high-dimensional feature selection.

We denote the MLE obtained from directly maximizing $L(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R})$ in (5.8) by $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{p}}$, and denote the "oracle" MLE of the parameters obtained by maximizing the likelihood constrained to the true set of attribute patterns by $(\widehat{\boldsymbol{\Theta}}^{\mathcal{A}_0}, \widehat{\boldsymbol{p}}^{\mathcal{A}_0})$. We denote the rate of convergence of $\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$ to $\ell(\widehat{\boldsymbol{\Theta}}^{\mathcal{A}_0}, \widehat{\boldsymbol{p}}^{\mathcal{A}_0})$ by $\delta \in (0, 1]$, that is,

$$\big[\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}}) - \ell(\widehat{\boldsymbol{\Theta}}^{\mathcal{A}_0}, \widehat{\boldsymbol{p}}^{\mathcal{A}_0})\big]/N = O_P(N^{-\delta}). \tag{5.11}$$

When $\delta = 1$, (5.11) implies $\ell(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}})$ converges with the usual root-N rate, and $\delta < 1$

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

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

Theorem V.3 (selection consistency)**.** *Suppose the true constraint matrix $\Gamma^{\mathcal{A}_0}$ associated with $\mathcal{A}_0$ satisfies conditions A, B and C in Theorem V.1. The true parameters satisfy*

$$\min_{\boldsymbol{\alpha}\in\mathcal{A}_0} p_{\boldsymbol{\alpha}} > c_0; \quad \theta_{j,\boldsymbol{\alpha}^\star} - \max_{\boldsymbol{\alpha}:\,\Gamma_{j,\boldsymbol{\alpha}}=0} \theta_{j,\boldsymbol{\alpha}} \geq c_1, \ \forall\ j = 1, \ldots, J \ \textit{and}\ \boldsymbol{\alpha}^\star \in \mathcal{C}_j, \tag{5.12}$$

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

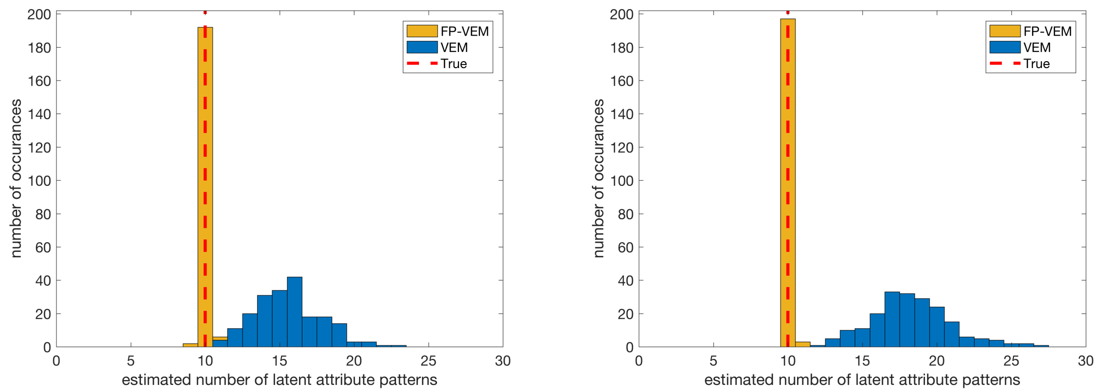
Remark V.6. Together with our identifiability result in Theorem V.1, the assumption (5.12) helps distinguish the true patterns from any alternative set of patterns with no larger cardinality, and further helps establish selection consistency. It is possible to further extend the current result and relax the constant lower bound assumption, though identifiability conditions would need to be adapted carefully to the case with a growing number of significant patterns and a shrinking magnitude of the proportions; we leave this for future work.

The proof of Theorem V.3 also reveals that if the convergence rate of U_N are slower than $\sqrt{N}$ with $\delta < 1$ in (5.11), then the tuning parameter λ in (5.9) has to satisfy $\lambda < -1$ in order to have pattern selection consistency; otherwise the issue of over selecting exists. Under the Bayesian interpretation as discussed in Remark V.5, this result implies that imposing the popular Dirichlet prior with a proper hyperparameter $\beta = \lambda + 1 \in (0, 1)$ is not sufficient for consistent selection of the significant mixture components (i.e., latent attribute patterns). Therefore, the approach proposed by Rousseau and Mengersen (2011) would not yield frequentist selection consistency in this considered scenario. We state this in the following proposition.

Proposition V.1 (selection inconsistency of Dirichlet prior)**.** Suppose $\delta < 1$ in (5.11), i.e., the rate of convergence of $\ell(\widehat{\mathbf{\Theta}}, \widehat{\boldsymbol{p}})$ is slower than the usual $\sqrt{N}$-rate. Then there does not exist a sequence of $\{\lambda_N,\ N = 1, 2, ...\} \subseteq [-1, 0)$ such that $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} = \mathcal{A}_0) \to 1$ as $N \to \infty$.

Example V.3. To visualize how the numbers of selected patterns differ for our proposed method based on maximizing (5.9) with $\beta = \lambda + 1 \in (-\infty, 1)$, and the variational EM algorithm resulting from imposing a proper Dirichlet prior over the proportions, we conduct a simulation study. In a simulation setting of $K = 10$ and $J = 30$, for each sample size $N = 500$ and 1000, we carry out 200 independent runs and in each run record the number of selected attribute patterns given by the proposed method, and that by the variational EM algorithm. We plot the histogram corresponding to the proposed method (FP-VEM, see Section 4 for details), together with that corresponding to Variational EM (VEM) with a small Dirichlet parameter $\beta = 0.01$. For both algorithms, we use the same threshold $\rho_N = 1/(2N)$ for selecting attribute patterns in the end of the algorithm, by only keeping patterns whose posterior means exceeds ρ_N. Here we did not plot the results corresponding to VEM with β smaller than 0.01, because we found the VEM algorithm with smaller β values can have convergence issues and in many cases it fails to converge but just jumps between several solutions. One can see from Figure 5.1 that the proposed method selects 10 patterns for most of datasets, which are indeed the 10 true patterns; while VEM over selects the patterns.

We next propose two algorithms to perform pattern selection, one being a modifi-



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

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

Figure 5.1: Histograms of estimated number of latent attribute patterns. VEM represents Variational EM with $\beta = \lambda + 1 = 0.01$, and FP-VEM represents the proposed Algorithm 2 in Section 4. The true number of latent attribute patterns is $|\mathcal{A}_0| = 10$.

cation of an EM algorithm, and the other being a variational EM algorithm resulting from an alternative formulation of the problem.

5.3.1.1 Modified EM algorithm.

We first consider using an EM algorithm with a slight modification in the E step to maximize (5.9). For each subject $i = 1, \ldots, N$, denote his/her latent attribute pattern by $\mathcal{A}_i = (A_{i,1}, \ldots, A_{i,K})$, then $\mathcal{A}_i \in \{0,1\}^K$. The complete log likelihood corresponding to (5.9) is

$$\ell^{\lambda}_{\text{comp}}(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R}, \mathcal{A}) = \sum_{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}}} \Big(\sum_i I(\mathcal{A}_i = \boldsymbol{\alpha}_l) + \lambda \Big) \log_{\rho_N}(p_{\boldsymbol{\alpha}_l}) \qquad (5.13)$$
$$+ \sum_{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}}} \sum_i I(\mathcal{A}_i = \boldsymbol{\alpha}_l) \sum_j \Big[R_{i,j} \log(\theta_{j,\boldsymbol{\alpha}_l}) + (1 - R_{i,j}) \log(1 - \theta_{j,\boldsymbol{\alpha}_l}) \Big],$$

where $I(\cdot)$ denotes the binary indicator function. Following the standard formulation of the EM algorithm (Dempster et al., 1977), in the E step of the $(t+1)$-th iteration, conditional expectations of $\ell^{\lambda}_{\text{comp}}(\boldsymbol{\Theta}, \boldsymbol{p} \mid \mathcal{R}, \mathcal{A})$ is evaluated with respect to the posterior distribution of latent variables $\mathcal{A}_i$'s given the current iterates of parameters $\boldsymbol{\Theta}^{(t)}$

and $\boldsymbol{p}^{(t)}$. Specifically, in the E step we replace the indicator $I(\mathcal{A}_i = \boldsymbol{\alpha}_l)$ in (5.13) by the probability $\varphi_{i,l} = \mathbb{P}(\mathcal{A}_i = \boldsymbol{\alpha}_l \mid \boldsymbol{\Theta}^{(t)}, \boldsymbol{p}^{(t)})$; and this is equivalent to updating

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

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

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

Data: Q, responses $\mathcal{R}$, and candidate attribute patterns $\mathcal{A}_{\text{input}}$.
Initialize $\boldsymbol{\Delta} = (\Delta_1^{(0)}, \ldots, \Delta^{(0)}_{|\mathcal{A}_{\text{input}}|})$.
while *not converged* **do**
  In the $(t+1)$th iteration,
  for $(i, l) \in [N] \times [|\mathcal{A}_{\text{input}}|]$ **do**

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

  for $l \in [|\mathcal{A}_{\text{input}}|]$ **do**
    $\Delta_l^{(t+1)} = \max\{c,\ \lambda + \sum_{i=1}^N \varphi^{(t+1)}_{i,\boldsymbol{\alpha}_l}\}$; ($c > 0$ is pre-specified);
  $\boldsymbol{p}^{(t+1)} \leftarrow \boldsymbol{\Delta}^{(t+1)} / (\sum_l \Delta_l^{(t+1)})$;
  for $j \in [J]$ **do**
    $\boldsymbol{\Theta}^{(t+1)} =$
    $\arg\max_{\boldsymbol{\Theta}} \Big\{\sum_{\boldsymbol{\alpha}_l}\sum_i \varphi^{(t+1)}_{i,\boldsymbol{\alpha}_l} \sum_j \Big[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}_l}) + (1 - R_{i,j})\log(1 - \theta_{j,\boldsymbol{\alpha}_l})\Big]\Big\}$;
After the total T iterations,
Output: $\{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{input}} :\ p^{(T)}_{\boldsymbol{\alpha}_l} > \rho_N\}$.

Remark V.7. Under the two-parameter SLAM, the DINA model in Example I.1, or the identity-link multi-parameter all-effect SLAM, the GDINA model in Example I.3, the M-step of updating the item parameters $\{\theta_{j,\boldsymbol{\alpha}}\}$'s in Algorithm 1 has closed forms. Specifically, under DINA, for any item j the update for the unique parameters (θ_j^+, θ_j^-) takes the form

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

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

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

In addition, when certain latent patterns are not distinguishable as discussed earlier in Corollary V.3, we can easily modify Algorithm 1 from selecting attribute patterns to selecting equivalence classes of attribute patterns. For instance, under a two-parameter SLAM, given the row vectors $\{\boldsymbol{q}_j, j \in [J]\}$ of Q, we first obtain the representatives of the Q-induced equivalence classes: $\mathcal{A}_Q = \{\vee_{j\in S}\, \boldsymbol{q}_j : S \subseteq \{1,\dots,J\}\}$, then get the ideal response matrix of $\mathcal{A}_Q$, namely $\Gamma(\cdot, \mathcal{A}_Q) = (\gamma_{j,l})_{J\times|\mathcal{A}_Q|}$ where $\gamma_{j,l} = I(\boldsymbol{\alpha}_l \succeq \boldsymbol{q}_j)$ for $\boldsymbol{\alpha}_l \in \mathcal{A}_Q$ and $j \in [J]$. After initializing $\boldsymbol{\Delta} = (\Delta_1,\dots,\Delta_{|\mathcal{A}_Q|})$, we just follow the same iterative procedure as that of Algorithm 1 for the two-parameter SLAM. In the end of the algorithm, after calculating $\nu_{[\boldsymbol{\alpha}_l]} = \Delta_l/(\sum_m \Delta_m)$, we select those $[\boldsymbol{\alpha}_l]$ with proportion $\nu_{[\boldsymbol{\alpha}_l]}$ above a pre-specified threshold. From the selected equivalence classes of attribute profiles, we can go back to obtain their representatives which are combinations of the $\boldsymbol{q}$-vectors from $\mathcal{A}_Q$ defined in Equation (5.7).

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

Given the large model space, we propose to use the Extended Bayesian Information Criterion (EBIC) introduced in Chen and Chen (2008) to select the tuning parameter. Recall that we denote by $\mathcal{A}^\lambda$ the selected set of attribute patterns obtained by maximizing the penalized likelihood function (5.9) with the specific tuning parameter λ. And we denote the item parameters and proportion parameters defined on this $\mathcal{A}^\lambda$ by $\boldsymbol{\Theta}^{\mathcal{A}^\lambda}$ and $\boldsymbol{p}^{\mathcal{A}^\lambda}$, respectively. The EBIC family have the following information criterion

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

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

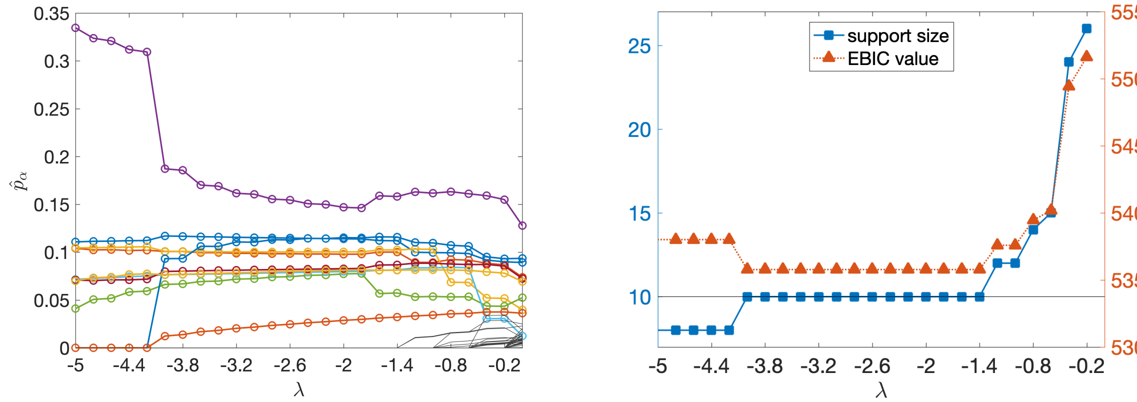
Example V.4. Figure 5.2 presents an illustration of the solution paths of the estimated proportions versus λ based on a simulated dataset with $N = 150$, $K = 10$, and $J = 30$. The Q-matrix $Q = (Q_1^\top, Q_2^\top, Q_3^\top)^\top$ with Q_i in the following form,

$$Q_1 = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 1 \end{pmatrix}, \quad Q_3 = \begin{pmatrix} 1 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 1 \end{pmatrix}. \tag{5.14}$$

When generating the data, 10 attribute patterns are randomly selected from the $2^{10} = 1024$ possible ones as true patterns, and the proportion of each of them is set to be 0.1. The item parameters are set to $1 - \theta_j^+ = \theta_j^- = 0.2$ for each j under a two-parameter SLAM. In the current setting with $K = 10$, we take the set of patterns as input to the PEM algorithm to be $\mathcal{A}_{\text{input}} = \{0, 1\}^K$. Figure 5.2(a) plots the solution paths of the estimated proportions of all the $2^{10} = 1024$ attribute patterns as λ varies in $\{-0.2, -0.4, \cdots, -4.8, -5.0\}$. The 10 true attribute patterns are plotted with colored lines with circles while the remaining $2^{10} - 10$ attribute patterns are plotted with black solid lines. Figure 5.2(b) plots the estimated support size of $\boldsymbol{p}$ versus λ, and the EBIC value versus λ. We observe that when $\lambda \in [-4.4, -1.4]$, Algorithm 1 selects the correct model with 10 true attribute patterns. This interval of λ corresponds to a "stable window" of the estimation algorithm that gives the correct selection and also has the smallest EBIC value. For this specific dataset, the proposed method along with EBIC succeeds in selecting the true model. Please see Section 5.4 for more simulation results which show that the proposed methods combined with EBIC indeed have good performance in general.

5.3.1.2 Variational EM algorithm from an alternative formulation.

In the following, we discuss an alternative formulation of the objective function (5.9) and propose a variational EM algorithm for estimation, by treating the proportion parameters $\boldsymbol{p}$ as latent random variables. As discussed in Remark V.5, for the



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

Figure 5.2: PEM solution paths and EBIC values in one trial, $N = 150$.

objective function (5.9) with $\lambda \in (-\infty, -1]$, the penalty term $\prod_{l=1}^{2^K} p_{\boldsymbol{\alpha}_l}^{\lambda}$ does not correspond to a proper Dirichlet distribution density. However, for any arbitrarily small λ value, the objective function (5.9) can be replaced by the following alternative formulation:

$$\ell_{\text{pseudo}}^{\lambda,\Upsilon}(\boldsymbol{\Theta}, \boldsymbol{p}) = \Upsilon \cdot \ell(\boldsymbol{\Theta}, \boldsymbol{p}) + (\beta - 1) \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}}} \log_{\rho_N}(p_{\boldsymbol{\alpha}}) \quad \text{for } \beta \in (0,1),\ \Upsilon \in (0,1]. \tag{5.15}$$

where we introduce a new parameter $\Upsilon \in (0, 1]$ and replace λ with $\beta - 1$ to respect the convectional notation of a Dirichlet distribution with hyperparameter $\beta \in (0, 1)$ to encourage sparsity. With $\beta \in (0, 1)$ and $\Upsilon \in (0, 1]$, the ratio $(1 - \beta)/\Upsilon$ can be arbitrarily large when Υ is arbitrarily close to zero, therefore making (5.15) equivalent to (5.9).

In the new objective function (5.15), the penalty term $\prod_{l=1}^{2^K} p_{\boldsymbol{\alpha}_l}^{\beta-1}$, $\beta \in (0, 1)$, can be viewed as a well-defined Dirichlet density function for the latent variables $\boldsymbol{p}$. In (5.15), the first term is the logarithm of the likelihood function raised to a fractional power $\Upsilon \in (0, 1]$. One intuition behind (5.15) is that given a moderate sample size and a large number of potential latent patterns, one needs to downweight the influence of the data likelihood and magnify the prior information encoded by the Dirichlet prior, in order to have the sufficient extent of shrinkage. The fractional-powered likelihood multiplied by the Dirichlet density can then be treated as a loss function to minimize. The idea of assigning a fractional power to the likelihood was also used in the Bayesian literature, such as Bissiri et al. (2016) and Holmes and Walker (2017) for Bayesian learning under model misspecification, and Yang et al. (2018) and Chérief-Abdellatif and Alquier (2018) for variational Bayesian inference. Different from these works, here we use the alternative formulation (5.15) of the original objective function (5.9) in order to consistently select the significant latent attribute patterns.

The formulation (5.15) allows for a variational EM algorithm for obtaining the item parameters $\mathbf{\Theta}$ and the posterior means of the latent variables $\boldsymbol{p}$. Here we treat $\mathbf{\Theta}$ still as model parameters, then we follow the general derivation of variational algorithms in Blei et al. (2017) to derive Algorithm 2. We denote the digamma function by $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$ for $x \in (0, \infty)$. In particular, the complete log likelihood is

$$\begin{aligned}\ell^{\lambda,\Upsilon}_{\text{comp}}(\mathbf{\Theta} \mid \mathcal{R}, \mathcal{A}, \boldsymbol{p}) = &\sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}}} \Big\{\Upsilon \cdot \Big[\sum_i I(\mathcal{A}_i = \boldsymbol{\alpha})\Big] + \beta - 1\Big\} \log_{\rho_N}(p_{\boldsymbol{\alpha}}) \\ &+ \Upsilon \cdot \Big\{ \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}}} \sum_i I(\mathcal{A}_i = \boldsymbol{\alpha}) \sum_j \Big[R_{i,j}\log(\theta_{j,\boldsymbol{\alpha}}) + (1 - R_{i,j})\log(1 - \theta_{j,\boldsymbol{\alpha}})\Big]\Big\}.\end{aligned} \tag{5.16}$$

In the variational E step, we first obtain the conditional probability of $I(\mathcal{A}_i = \boldsymbol{\alpha}_l)$ for each individual i and each input attribute pattern $\boldsymbol{\alpha}_l$, which we denote by $\varphi_{i,\boldsymbol{\alpha}_l}$. In updating this $\varphi_{i,\boldsymbol{\alpha}_l}$, the variational posterior distribution of the $p_{\boldsymbol{\alpha}}$'s are used, which is still a Dirichlet distribution with mean parameters $(\Delta_1, \dots, \Delta_{|\mathcal{A}_{\text{input}}|})$ updated in the previous E step (or from initializations if in the first iteration). Then we update the mean parameters for the variational posterior distribution of $p_{\boldsymbol{\alpha}_l}$'s based on the obtained $\varphi_{i,\boldsymbol{\alpha}_l}$, following the conventional derivation in variational inference. After finishing this E step, in the M step we maximize the complete likelihood with respect to $\mathbf{\Theta}$, by substituting the $I(\mathcal{A}_i = \boldsymbol{\alpha}_l)$'s with $\varphi_{i,\boldsymbol{\alpha}_l}$'s. Note that taking the derivatives of (5.16) with respect to $\theta_{j,\boldsymbol{\alpha}_l}$'s does not involve either terms of $p_{\boldsymbol{\alpha}_l}$ or terms of Υ and β, so only $\varphi_{i,\boldsymbol{\alpha}_l}$ are used in the M step for updating $\boldsymbol{\Theta}$. Indeed, the M step of updating $\boldsymbol{\Theta}$ in the current Algorithm 2 takes the same form as that of Algorithm 1.

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

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

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

while *not converged* **do**

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

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

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

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

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

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

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

After the total T iterations,

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

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

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

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

to a relatively small value, say $\beta = 0.01$, and let the fractional power $\Upsilon \in (0, 1]$ vary to control the sparsity level of the proportion parameters.

5.3.2 Screening as a Preprocessing Step When $2^K \gg N$

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

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

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

We next derive an algorithm with a stochastic EM flavor to estimate the posterior mean of each latent variable $A_{i,k}$, denoted by a matrix $(\widehat{a}_{i,k})$ of size $N \times K$, where $\widehat{a}_{i,k} = \mathbb{E}[A_{i,k} \mid \cdot]$. In the end of the algorithm, we obtain the binary matrix W containing the candidate attribute patterns by defining $W = (w_{i,k})_{N\times K}$ with $w_{i,k} = I(\widehat{a}_{i,k} > 1/2)$. In such a screening procedure, we first use the dependency among the K attributes in iterative updates, then partly ignore the dependency in the last step through applying Bayes' rule to each subject i's each single attribute k. This results in fast and valid screening of attribute patterns. Viewing the ith row vector of W as the estimated attribute pattern of subject i, the unique row vectors in W are the roughly selected attribute patterns output by the screening stage. We denote this set of candidate patterns by $\widehat{\mathcal{A}}_{\text{screen}}$. As long as the screening has the nice property of "no false exclusion", meaning the rows in W contain all the true attribute patterns, then the screening stage is considered successful. The selected candidate patterns are passed along to the shrinkage estimation stage as input patterns.

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

Theorem V.4 (sure screening property). *Suppose the identifiability conditions in Theorem V.1 and the constraints* (5.12) *are satisfied. The screening procedure applied to a SLAM that covers the two-parameter SLAM as a submodel has the sure screening property. Specifically, there exists a constant* $\beta_{\min} > 0$ *such that* $\mathbb{P}(\widehat{\mathcal{A}}_{\text{screen}} \supseteq \mathcal{A}_0) \geq 1 - |\mathcal{A}_0| \exp(-N\beta_{\min}) \to 1$ *as* $N \to \infty$.

Theorem V.4 shows that the probability of the screening procedure failing to include all true patterns has an exponential decay with the sample size N. We point out that despite having the nice property of sure screening, the screening procedure does not guarantee consistency in selecting exactly the set $\mathcal{A}_0$ of true patterns, if the number of observed variables per subject J is not large enough. Generally speaking, as N goes large but J does not, the set $\widehat{\mathcal{A}}_{\text{screen}}$ will include many false attribute patterns, although it will contain the true set $\mathcal{A}_0$ with probability tending to one. Therefore the shrinkage estimation approach in Section 5.3.1 is still essential to performing pattern selection.

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

Algorithm 3: Stochastic Approximation Gibbs Screening

Data: Q, $\boldsymbol{R}$

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

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

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

while *not converged* **do**

- $\mathcal{A}^{\text{s}} \leftarrow \mathbf{0}, \quad \mathbf{I}^{\text{s}} \leftarrow \mathbf{0}, \quad M_{\text{eff}} \leftarrow 0.$
- **for** $r \in [M_{\max}]$ **do**
  - **for** $(i,k) \in [N] \times [K]$ **do**
    - Draw $A_{i,k} \sim \text{Bernoulli}\Big(\text{logit}^{-1}\Big(\sum_j q_{j,k} \prod_{m\neq k} A_{i,m}^{q_{j,m}} \Big[R_{i,j} \log \frac{\theta_j^+}{\theta_j^-} + (1 - R_{i,j}) \log \frac{1-\theta_j^+}{1-\theta_j^-}\Big]\Big)\Big).$
  - **if** $r \geq M_{\max} - M_{\text{eff}}$ **then**
    - $\mathcal{A}^{\text{s}} \leftarrow \mathcal{A}^{\text{s}} + \mathcal{A}, \qquad \mathbf{I}^{\text{s}} \leftarrow \mathbf{I}^{\text{s}} + \Big(\prod_k A_{i,k}^{q_{j,k}}\Big)_{N\times J}.$
- $\mathcal{A}^{\text{ave}} \leftarrow \frac{1}{t}\mathcal{A}^{\text{s}}/M_{\text{eff}} + \Big(1 - \frac{1}{t}\Big)\mathcal{A}^{\text{ave}}, \quad \mathbf{I}^{\text{ave}} \leftarrow \frac{1}{t}\mathbf{I}^{\text{s}}/M_{\text{eff}} + \Big(1-\frac{1}{t}\Big)\mathbf{I}^{\text{ave}}, \quad t = t+1.$
- **for** $j \in [J]$ **do**
  - $\theta_j^+ \leftarrow (\sum_i R_{i,j} I_{i,j}^{\text{ave}})/(\sum_i I_{i,j}^{\text{ave}}), \qquad \theta_j^- \leftarrow (\sum_i R_{i,j}(1-I_{i,j}^{\text{ave}}))(\sum_i(1-I_{i,j}^{\text{ave}})).$

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

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

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

Remark V.8. The screening algorithm can be modified to be more conservative in order to reduce the risk of excluding true patterns. In particular, after each stochastic E step in the screening algorithm, based on the current iterate of $\mathbf{A}^{\text{ave}}$ we can obtain a $N \times K$ binary matrix with the (i,k)th entry being $I(A^{\text{ave}}_{i,k}) > 1/2$. The unique row vectors of this intermidiate binary matrix can be viewed as the current candidate latent patterns. To make the screening procedure more conservative, we recommend saving this set of candidate patterns after every M stochastic EM iterations (M is a positive integer), and take the union of these saved sets in the end of the algorithm to form $\widehat{\mathcal{A}}_{\text{screen}}$ as the output. We call this strategy "*screening enhanced by Gibbs exploration*", since it takes advantage of the latent patterns that the Gibbs sampling explores along the stochastic EM iterations.

In Figure 5.3, we present an estimation pipeline summarizing the proposed screening and shrinkage procedures. In practice, when the number of potential latent patterns 2^K is of too high dimensions, we recommend to first perform screening by using Algorithm 4 or "screening enhanced by Gibbs exploration" to bring down the number of candidate patterns. The cardinality of the set of candidate patterns is usually at the order of the sample size N. Then over a set of $O(N)$ number of candidate latent patterns, one can proceed to apply the shrinkage estimation methods Algorithm 1 or 2 to select the final set of latent attribute patterns.

5.4 Simulation Studies

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

Two-parameter SLAM (DINA Model). Consider the two-parameter SLAM with a $3K \times K$ Q-matrix $Q = (Q_1^\top, Q_2^\top, Q_3^\top)^\top$, where the three submatrices Q_1, Q_2 and Q_3 are specified in (5.14). We consider three dimensions of possible attribute patterns

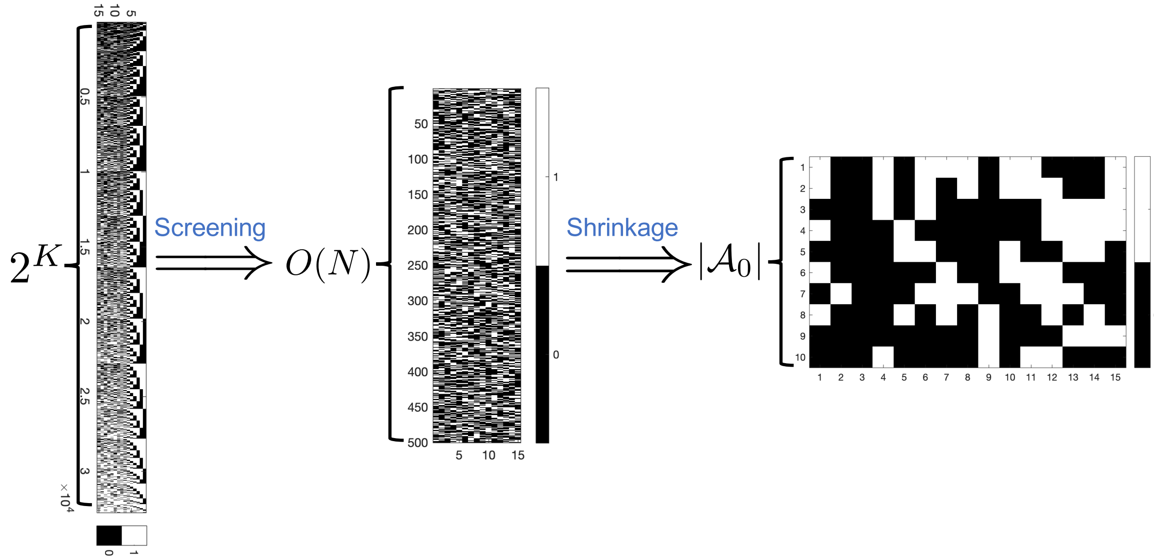


Figure 5.3: Estimation pipeline combining the proposed screening and shrinkage procedures.

with $2^K = 2^{10}, 2^{15}$, and 2^{20}, three sample sizes with $N = 150, 500$ and 1000, and two different signal levels with true item parameters: $\{\theta_j^+ = 0.8,\ \theta_j^- = 0.2;\ j \in [J]\}$, the relatively weak signals; and $\{\theta_j^+ = 0.9,\ \theta_j^- = 0.1;\ j \in [J]\}$, the relatively strong signals. We randomly generate the set of true attribute patterns $\mathcal{A}_0 \subseteq \{0,1\}^K$ with cardinality $|\mathcal{A}_0| = 10$ and set $p_{\boldsymbol{\alpha}} = 0.1$ for all $\boldsymbol{\alpha} \in \mathcal{A}_0$. In the simulations, for $K = 10$ the $\mathcal{A}_{\text{input}}$ is taken to be $\{0,1\}^K$; while for $K = 15$ and 20, the $\mathcal{A}_{\text{input}}$ is taken to be $\widehat{\mathcal{A}}_{\text{screen}}$, i.e., the set of candidate patterns output by the screening method.

In each scenario we perform 200 independent replications. For shrinkage estimation, we apply the proposed Algorithm 1 "Penalized EM (PEM)" and Algorithm 2 "Fractional Power Variational EM (FP-VEM)", and also apply the plain EM algorithm with thresholding for comparison. When running PEM we compute a solution path by varying λ in the range of $\lambda \in \{-0.2, -0.4, \ldots, -3.8, -4.0\}$, and select the λ that gives the smallest EBIC. When running FP-VEM we fix $\beta = \lambda + 1 = 0.01$ and compute a solution path by varying Υ in $\{1.0, 0.9, \ldots, 0.4, 0.3\}$ and also select Υ using EBIC. We use the threshold value $\rho_N = 1/(2N)$ for the estimated proportions in the last step for all three shrinkage algorithms to select patterns (other smaller ρ_N

values give similar results).

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

Table 5.2: Pattern selection accuracies for two-parameter SLAM. Tuning parameter $\lambda \in \{-0.2, -0.4, \ldots, -3.8, -4.0\}$ in PEM (Algorithm 1) and $\Upsilon \in \{1.0, 0.9, \ldots, 0.4, 0.3\}$ in FP-VEM (Algorithm 2) are selected based on EBIC.

The simulation results on selection accuracies are presented in Table 5.2. The "TPR" stands for True Positive Rate, which denotes the proportion of true patterns that are selected. The "1-FDR" stands for "1−False Discovery Rate (FDR)", which denotes the proportion of selected patterns that are true patterns. Table 5.2 shows the proposed PEM and FP-VEM yield good selection results in various scenarios, while the EM algorithm with direct thresholding at ρ_N suffers from high FDR, i.e., selecting too many non-existing attribute patterns. We would like to point out that the plain VEM as presented in Example V.3 is a special case of the proposed FP-VEM, by just taking the fractional power Υ to be $\Upsilon = 1$. So in each simulation run, the result given by VEM is included in the solution path given by FP-VEM with $\Upsilon \in \{1.0, 0.9, \dots, 0.4, 0.3\}$, and in the final step EBIC selects the best Υ from the entire solution path. Indeed, in all our simulations about FP-VEM, the result given by $\Upsilon = 1$ is never selected by EBIC, which means the selection result given by plain VEM is never favored over the proposed FP-VEM. We also remark here that the proposed methods are computationally efficient. All the algorithms are implemented in Matlab. In particular, in the case of relatively strong signal with $1-\theta_j^+ = \theta_j^- = 0.10$, screening and computing an entire solution path for $(2^K, N) = (2^{20}, 1000)$ takes < 2 minutes on average on a laptop with a 2.8 GHz processor, and yields almost perfect pattern selection results, as shown in the last row of Table 5.2.

We give some discussions on the comparison of the PEM and the FP-VEM algorithms. The estimation accuracies presented in Table 5.2 generally show the two algorithms have comparable performance on pattern selection. In terms of selecting the tuning parameter, the FP-VEM can be easier to tune because the fractional power Υ is always between 0 and 1, while the PEM algorithm has a negative tuning parameter $\lambda \in (-\infty, 0)$ that can have an arbitrarily large magnitude. Specifically, the scenario of an increasing sparsity corresponds to $\Upsilon \to 0$ and $\lambda \to -\infty$, and when extremal sparsity exists, the FP-VEM needs to choose Υ close to zero with a small magnitude and the PEM needs to choose λ with a large magnitude. Therefore, in such cases the tuning of PEM may take more time, since $\lambda < 0$ needs to be searched over a relatively large interval; an exponential grid search might be of help in this case, while further investigation into how to best specify the grid for searching tuning parameters would be needed. Meanwhile, we find in simulation studies that choosing a small Υ in FP-VEM too close to zero may result in the algorithm to be less stable in some cases. In practice, if the computation time is not a primary concern, we recommend first considering the PEM algorithm for the better stability.

We further conduct a simulation study to investigate how the threshold value

ρ_N for the estimated proportions impact the pattern selection results of different methods. In the setting with $1 - \theta_j^+ = \theta_j^- = 0.2$ and $N = 150$ (the same setting as the first line in Table 5.2), we simulate 200 independent datasets, and apply the proposed PEM (Algorithm 1), FP-VEM (Algorithm 2) and the usual EM algorithm with various thresholds $\rho_N \in \{1/(50N)\} \cup \{i/(2N),\ i = 1, 3, 5, \ldots, 15\}$. Figure 5.4 plots the average "TPR" and average "1−FDR" versus the threshold values. It can be seen that directly thresholding the MLE of the proportions (corresponding to the thresholding after EM) does not yield good selection results. For a small threshold $\rho_N = 1/(2N)$, the FDR of thresholded EM is quite high. When further decreasing the threshold ρ_N from $1/(2N)$ to $1/(50N)$, the FDR of thresholded EM becomes worse while the proposed methods have stable performance. On the other hand, as the threshold ρ_N increases from $1/(2N)$ to larger values, the TPR of EM quickly decreases. In contrast, the proposed methods PEM and FP-VEM give reasonably good selection results across all the threshold values, and have slightly better performance for smaller thresholds. Even the best selection result given by thresholding EM corresponding to the threshold $\rho_N = 7/(2N)$ is not comparable to those given by the proposed methods.

We next evaluate the performance of the screening procedure. We find that the screening procedure drastically reduces the computational cost in the subsequent shrinkage estimation stage. For instance, in the setting $(N, K) = (150, 15)$ when noise rate is $1 - \theta_j^+ = \theta_j^- = 20\%$, based on 200 runs, the variational screening procedure takes 1.55 seconds on average, and the subsequent PEM algorithm takes 6.42 seconds on average; while if no screening is performed, the PEM algorithm takes 7.96×10^3 seconds on average.

As described earlier, the screening is considered successful if all true patterns are included in the candidate set $\widehat{\mathcal{A}}_{\text{screen}}$. Under each simulation scenario in Table 5.2 corresponding to $K = 15$ or $K = 20$, we record the coverage probabilities of the

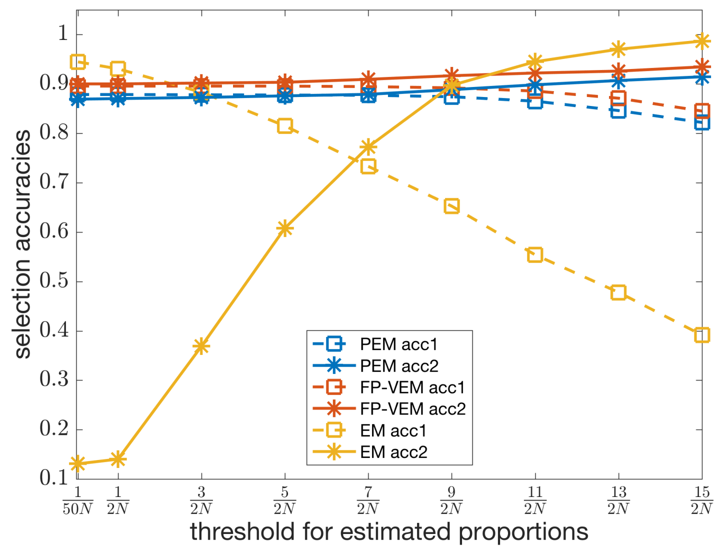
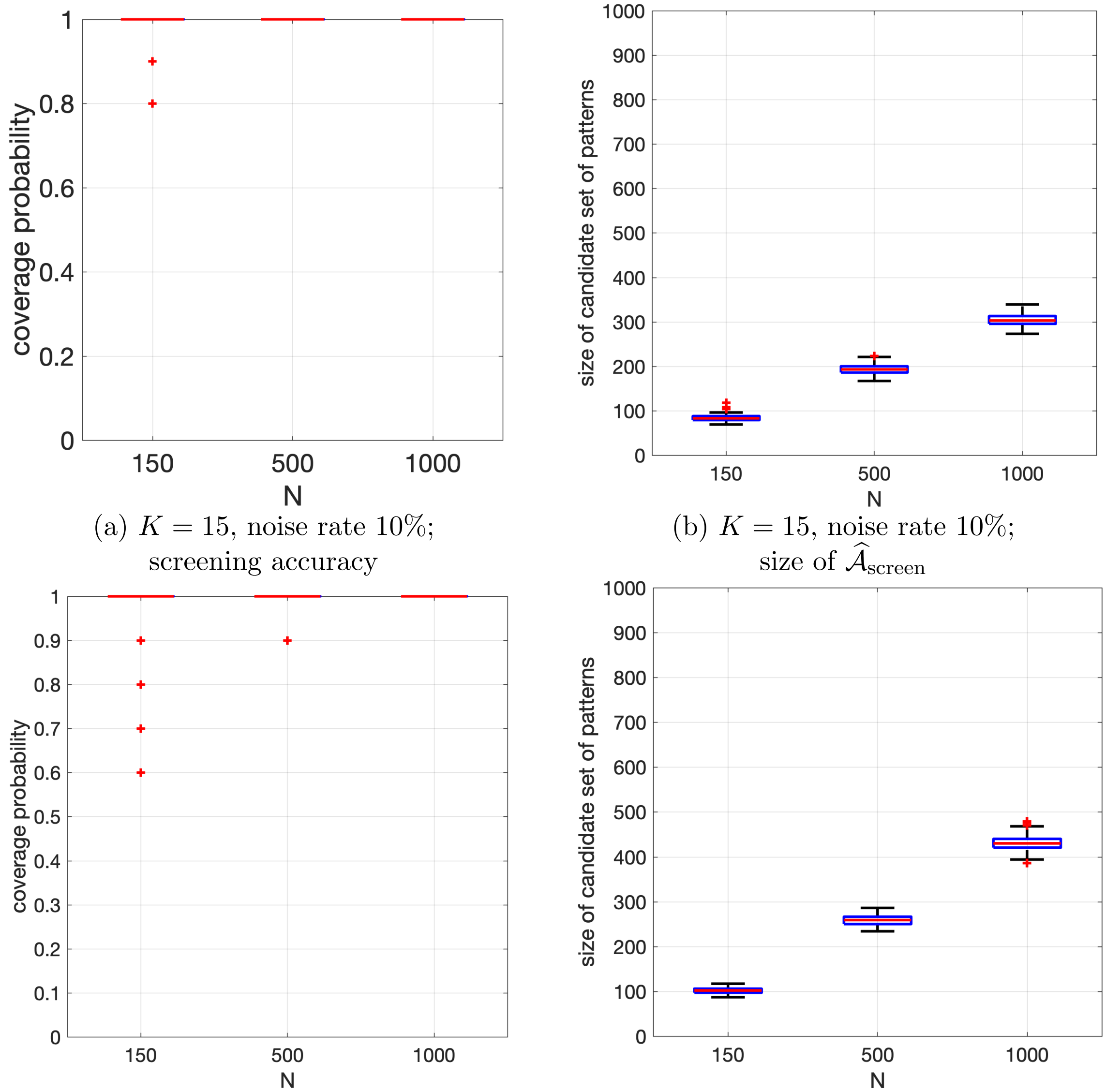


Figure 5.4: Selection accuracies versus thresholds for the two-parameter SLAM with $1-\theta_j^+=\theta_j^-=0.2$ and $N=150$. For each method, "acc1" denotes the True Positive Rate (TPR), the proportion of true patterns that are selected; and "acc2" denotes "1−False Discovery Rate (FDR)", the proportion of selected patterns that are true.

true patterns for each of 200 runs, where in each run $\sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} I(\boldsymbol{\alpha}\in\widehat{\mathcal{A}}_{\text{screen}})/|\mathcal{A}_0|$ is recorded as the coverage probability. The boxplots of coverage probabilities under these scenarios are presented in Figure 5.5(a), (c) and Figure 5.6(a), (c). We also record the size of $\widehat{\mathcal{A}}_{\text{screen}}$, i.e., the number of candidate patterns given by the screening procedure in each run, and present their boxplots in Figure 5.5(b), (d) and Figure 5.6(b), (d). The screening procedure generally has good performance. On the other hand, Figure 5.6(e) and (g) show that for the relatively large noise rate and small sample size, the screening accuracy is not very high.

To improve the performance of screening, we apply the strategy of *screening enhanced by Gibbs exploration* described in Remark V.8 and take $M=3$. That is, along the stochastic EM iterations of the screening algorithm, after every three iterations we add the current set of latent patterns to the candidate set $\widehat{\mathcal{A}}_{\text{screen}}$. The resulting screening accuracies and sizes of $\widehat{\mathcal{A}}_{\text{screen}}$ are presented in Figure 5.7. Compared to Figure 5.6, one can clearly see that the enhancing procedure improves the screening



(a) $K = 15$, noise rate 10%; screening accuracy

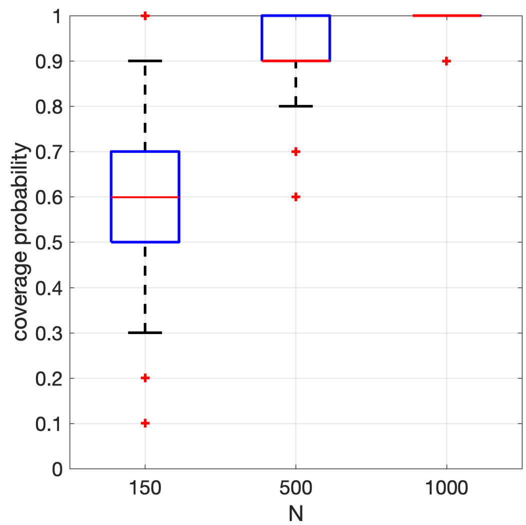
(b) $K = 15$, noise rate 10%; size of $\widehat{\mathcal{A}}_{\text{screen}}$

(c) $K = 20$, noise rate 10%; screening accuracy

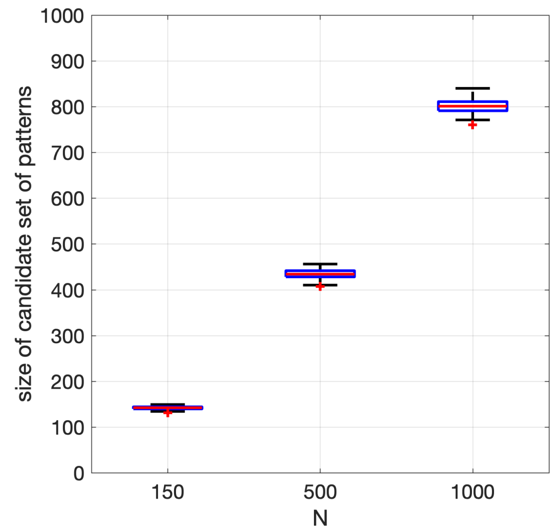
(d) $K = 20$, noise rate 10%; size of $\widehat{\mathcal{A}}_{\text{screen}}$

Figure 5.5: Screening under noise rate 10%: plots (a), (c) are coverage probabilities of the true patterns, from the screening procedure under the two-parameter SLAM; plots (b), (d) are sizes of $\widehat{\mathcal{A}}_{\text{screen}}$. The "noise rate" refers to the value of $1 - \theta_j^+ = \theta_j^-$.

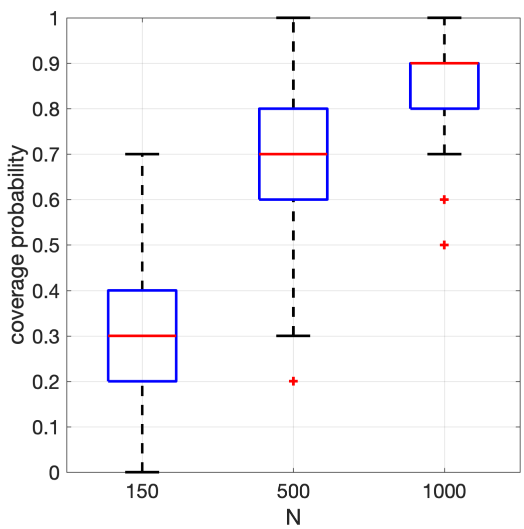
accuracy significantly, while the size of $\mathcal{A}_{\text{screen}}$ also increases but still remains quite manageable. Under the noise rate $1 - \theta_j^+ = \theta_j^- = 20\%$, the size of $\mathcal{A}_{\text{screen}}$ is always below N for screening without enhancing, while for screening with enhancing, the size of $\mathcal{A}_{\text{screen}}$ is around $2N$ for $K = 15$ and around $3N$ for $K = 20$. The enhancing by Gibbs exploration would not sacrifice the efficiency of the screening procedure itself,



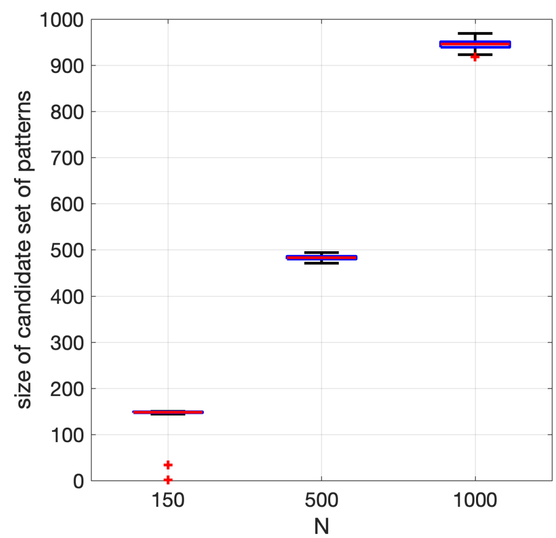
(e) $K = 15$, noise rate 20%; screening accuracy



(f) $K = 15$, noise rate 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$



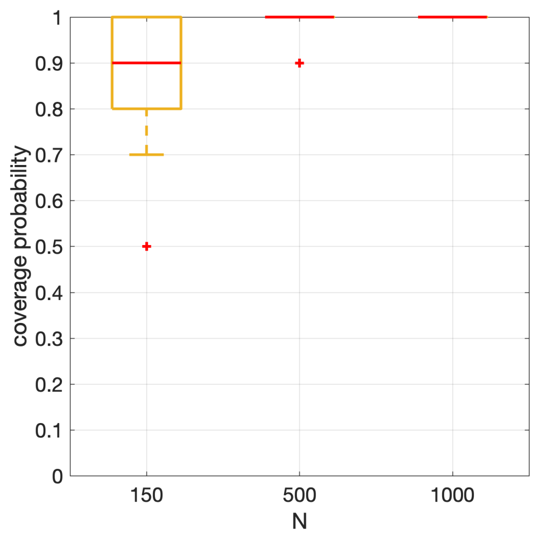
(g) $K = 20$, noise rate 20%; screening accuracy



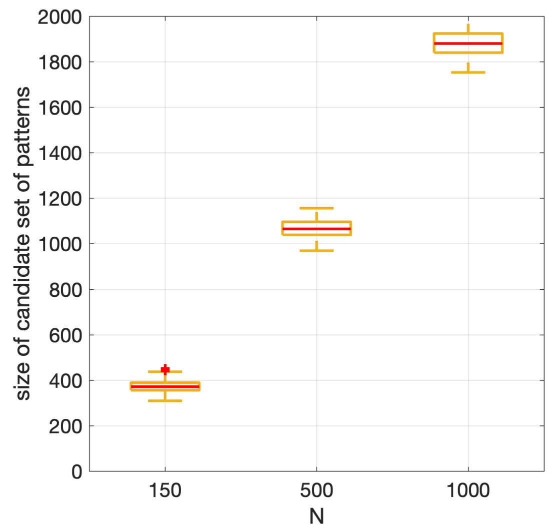
(h) $K = 20$, noise rate 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$

Figure 5.6: Screening under noise rate 20%: plots (a), (c) are coverage probabilities of the true patterns, from the screening procedure under the two-parameter SLAM; plots (b), (d) are sizes of $\widehat{\mathcal{A}}_{\text{screen}}$. The "noise rate" refers to the value of $1 - \theta_j^+ = \theta_j^-$.

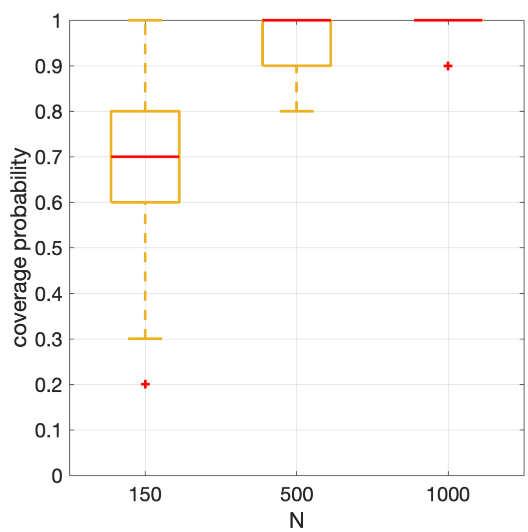
though it results in a larger set of $\widehat{\mathcal{A}}_{\text{screen}}$ which incurs higher computational cost in the shrinkage stage. In practice, one should leverage this tradeoff according to the sample size. Specifically, when sample size N is small, choosing a more conservative screening procedure (with a smaller integer M) is recommended, because this would increase the screening accuracy without causing much computational burden for the



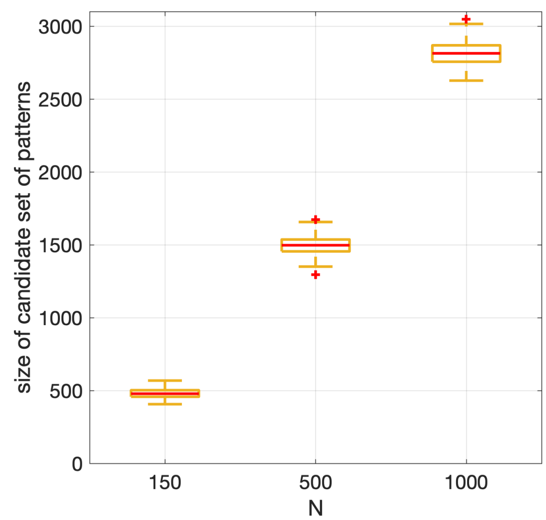
(a) $K = 15$, noise rate 20%; screening accuracy



(b) $K = 15$, noise rate 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$



(c) $K = 20$, noise rate 20%; screening accuracy



(d) $K = 20$, noise rate 20%; size of $\widehat{\mathcal{A}}_{\text{screen}}$

Figure 5.7: Screening enhanced by Gibbs exploration: screening accuracy and size of $\widehat{\mathcal{A}}_{\text{screen}}$. Noise rate is $1 - \theta_j^+ = \theta_j^- = 20\%$.

shrinkage algorithm. With the enhanced screening procedure, in the relatively weak signal case $1 - \theta_j^+ = \theta_j^- = 0.2$ and under $(K, N) = (15, 150)$, the two accuracy measures 1−FDR and TPR for the PEM algorithm, become $(0.850, 0.860)$ (previously it was $(0.523, 0.530)$ in Table 5.2), and those under the FP-VEM algorithm become $(0.839, 0.853)$ (previously $(0.544, 0.543)$ in Table 5.2). Under $(K, N) = (20, 150)$, the two accuracy measures for the PEM become $(0.608, 0.648)$ (previously $(0.213, 0.255)$ in

Table 5.2) and those for the FP-VEM become $(0.620, 0.634)$ (previously $(0.264, 0.271)$ in Table 5.2).

Multi-parameter all-effect SLAM (GDINA Model). We next consider the multi-parameter all-effect SLAM introduced in Example I.3 with an identity link function $f(\cdot)$, that is, the GDINA model proposed in de la Torre (2011). Let the Q-matrix be in the form $Q = (Q_1^\top, Q_2^\top, Q_2^\top)$ with Q_1 and Q_2 specified in (5.14). Similar to the two-parameter simulation study, we consider three dimensions of possible attribute patterns with $2^K = 2^{10}, 2^{15}$, and 2^{20}, and three sample sizes with $N = 150, 500$ and 1000. For each item, we set the baseline probability, the positive response probability of the all-zero attribute pattern $\boldsymbol{\alpha} = \mathbf{0}_K$, to 0.2 (i.e., $\theta_{j,\mathbf{0}_K} = 0.2$), and the positive response probability of $\boldsymbol{\alpha} = \mathbf{1}_K$ to 0.8 (i.e., $\theta_{j,\mathbf{1}_K} = 0.8$). And we set all the main effects and interaction effects parameters of the item to be equal (i.e., $\beta_{j,S_1} = \beta_{j,S_2}$ for any $\varnothing \neq S_1, S_2 \subset \mathcal{K}_j$ for the β-coefficients in (I.3)). We randomly generate the set of true attribute patterns, $\mathcal{A}_0 \subseteq \{0,1\}^K$ with cardinality $|\mathcal{A}_0| = 10$ and set $p_{\boldsymbol{\alpha}} = 0.1$ for all $\boldsymbol{\alpha} \in \mathcal{A}_0$.

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

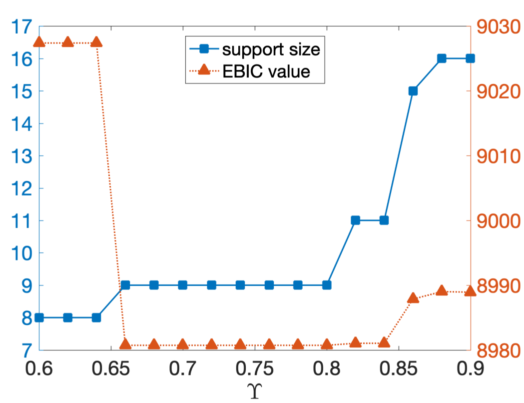
Table 5.3: Pattern selection accuracies for multi-parameter all-effect SLAM. Tuning parameter $\lambda \in \{-0.2, -0.4, \dots, -4.0\}$ in PEM (Algorithm 1) and $\Upsilon \in \{1.0, 0.9, \dots, 0.3\}$ in FP-VEM (Algorithm 2) are selected using EBIC. Signal strengths are $\theta_{j,\mathbf{0}_K} = 0.1,\ \theta_{j,\mathbf{1}_K} = 0.9$.

Similar to the observations in Table 5.2, Table 5.3 shows that the proposed methods also have good pattern selection performance for the more complicated multi-parameter all-effect model. The approximate screening algorithm based on the likelihood of the two-parameter submodel is quite effective here for obtaining candidate patterns under the multi-parameter model. And similarly to the two-parameter case, the EM algorithm tends to severely overselects the attribute patterns. Please see Appendix B for additional results on the performance of the screening procedure.

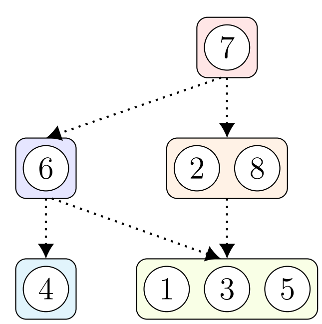
5.5 Data Analysis

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

Analysis of Fraction Subtraction Data. As introduced in Chapter I, the fraction subtraction dataset contains $N = 536$ middle school students' binary (correct or wrong) responses to 20 questions that were designed for the diagnostic assessment of 8 skill attributes related to fraction and subtraction. See Table 1.3 in Chapter I for the Q-matrix specified in de la Torre and Douglas (2004a).

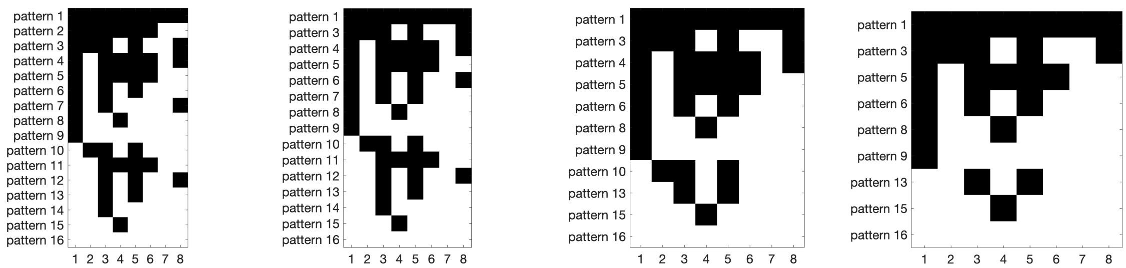


(a) EBIC values and support sizes vs. Υ



(b) attribute structure selected by EBIC

Figure 5.8: Results of Fraction Subtraction Data analyzed using two-parameter SLAM.

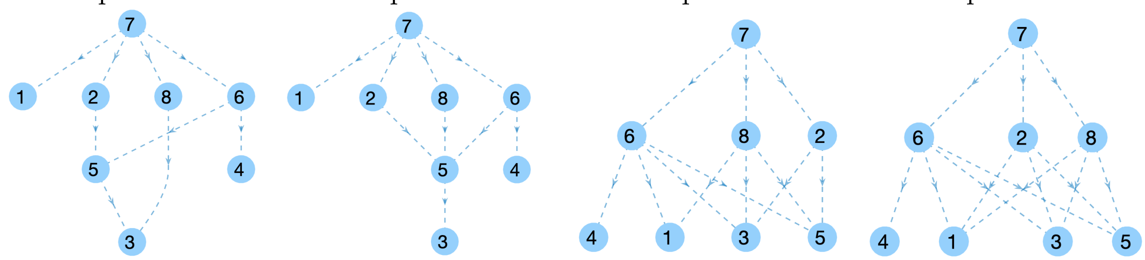


(a) $\Upsilon = 0.90$,
16 patterns

(b) $\Upsilon = 0.86$,
15 patterns

(c) $\Upsilon \in [0.82, 0.84]$,
11 patterns

(d) $\Upsilon \in [0.66, 0.80]$,
9 patterns



(e) $\Upsilon = 0.90$,
8 groups of attributes

(f) $\Upsilon = 0.86$,
7 groups of attributes

(g) $\Upsilon \in [0.82, 0.84]$,
7 groups of attributes

(h) $\Upsilon \in [0.66, 0.80]$,
5 groups of attributes

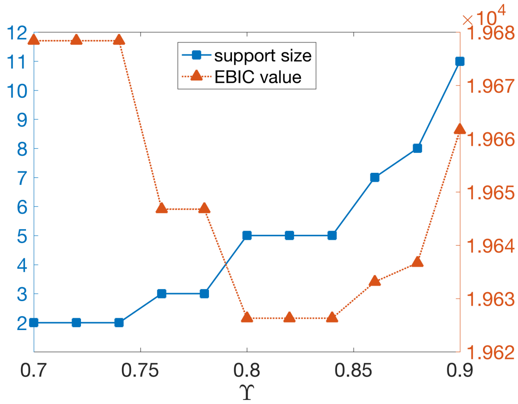
Figure 5.9: Fraction Subtraction Data: different sets of estimated patterns (a)–(d) (black for "0" and white for "1") and the corresponding attribute structures (e)–(h) under various Υ's in Algorithm 2. Plot (h) here is equivalent to Figure 5.8(b).

Many studies in the literature use the two-parameter SLAM to fit the dataset, mostly due to that it is reasonable to assume the required attributes of each item act together to form a "capable" knowledge state and an "incapable" knowledge state. This results in two levels of item parameters for each item. We first use the two-parameter model to analyze the data. Given this 20×8 Q-matrix, the number of equivalence classes induced by the Q-matrix $Q_{20\times 8}$ under the two-parameter model is $|\{\vee_{j\in S}\boldsymbol{q}_j : S \subseteq \{1,\dots,J\}\}| = 58$. We apply Algorithm 2, the FP-VEM algorithm with a sequence of fractional power values $\Upsilon \in \{0.90, 0.89, \cdots, 0.60\}$ and use EBIC to select the tuning parameter Υ while keeping the Dirichlet hyperparameter $\beta = 0.01$. Figure 5.8(a) plots the EBIC values and the support sizes of $\boldsymbol{p}$, both against the Υ values. It can be seen that $\Upsilon = 0.8$ yields the smallest EBIC value 8.98×10^3, and it is the largest Υ value in the flat window of $[0.66, 0.8]$ that gives 9 equivalence classes of attribute patterns. We also use the multi-parameter all-effect model (GDINA model) introduced in Example I.3 to fit the dataset. For a range of values of the tuning parameters Υ, the smallest EBIC value is above 1.02×10^4, which is much higher than the smallest EBIC 8.98×10^3 given by the two-parameter model. This also aligns with the results in the literature that the two-parameter model fits the fraction subtraction dataset better than other models (DeCarlo, 2011; de la Torre and Douglas, 2004a). Therefore next we only present and discuss the results given by the two-parameter model.

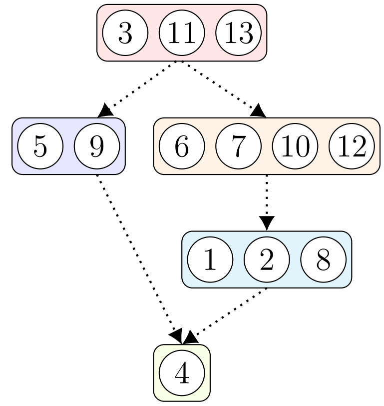
Figure 5.8(b) plots the attribute structure corresponding to the 9 equivalence classes of attribute patterns selected by EBIC. We obtain this attribute structure using the following procedure. First, we obtain the representatives of these 9 equivalence classes and construct a 9×8 matrix of selected attribute patterns. We denote this 9×8 matrix by $\widehat{\mathcal{A}}$, with each row of $\widehat{\mathcal{A}}$ a 8-dimensional binary vector denoting one selected knowledge state (i.e., attribute pattern). We next examine the partial orders among the columns of this matrix to determine the relationships among attributes. In particular, if $\widehat{\mathcal{A}}(\cdot, k_1) \succeq \widehat{\mathcal{A}}(\cdot, k_2)$, then attribute k_1 is considered as a prerequisite for attribute k_2. Examining these 9 selected knowledge states, we find that the total number of 8 attributes are separated into 5 groups $G_1 = \{7\}$, $G_2 = \{2, 8\}$, $G_3 = \{6\}$ and $G_4 = \{4\}$ and $G_5 = \{1, 3, 5\}$, such that the attributes in the same group play the same role in clustering the students population into the 9 knowledge states. In particular, based on the observed data, attributes 2 and 8 are equivalent in distinguishing the students population's knowledge states; and so are attributes 1, 3, 5. The estimated prerequisite relationship among these 5 groups is depicted in Figure 5.8(b). Figure 5.8(b) implies that attribute (α_7) *Subtract numerators*, is a quite basic skill attribute and serves as prerequisite for all the remaining attributes. This suits the common sense that in the problems about fraction and subtraction, the ability of subtracting integers should be the most basic. Figure 5.8 also shows that attributes (α_2), (α_6), (α_8) are middle level skills that only has one prerequisite attribute (α_7), and serve as prerequisites for multiple other skills. Finally, the remaining attributes (α_4), (α_1), (α_3) and (α_5) are high level skills in the hierarchical structure. We would like to point out that the directed edges in the attribute hierarchy in Figure 5.8(b) (and also in the later Figure 5.10 for the TIMSS dataset) do not necessarily correspond to causal relations between the skill attributes. Instead, the attribute hierarchy results from the learned subset of attribute patterns, and it just reflects the estimated cognitive structure of the students being measured.

For the Fraction Subtraction data, in addition to the attribute structure chosen by EBIC shown in Figure 5.8(b), we also present those sets of attribute patterns selected by different Υ's in the solution path. The four sets of patterns and their corresponding attribute structures are presented in Figure 5.9. As shown in Figure 5.9(a)–(d), the latent patterns selected by a smaller Υ always form a subset of those patterns selected by a larger Υ. Also, the attribute structures selected by different Υ's share some commonalities. Among the second row of Figure 5.9, plot (h) is equivalent to the attribute structure in Figure 5.8(b).

Analysis of TIMSS Data. We also apply the proposed method to the TIMSS 2003 8th grade data. The dataset contains $N = 757$ students' responses to $J = 23$ test items, and the Q-matrix is of size 23×13. Under the two-parameter SLAM, the Q-matrix gives $|\{\vee_{j\in S}\boldsymbol{q}_j : S \subseteq \{1,\ldots,J\}\}| = 1625$ equivalence classes. Figure 5.10 shows the results of fitting the two-parameter SLAM with $\beta = 0.01$. The fractional power Υ selected by EBIC is 0.84 and the corresponding number of equivalence classes is 5. The smallest EBIC value in Figure 5.10(a) is 1.96×10^4. We remark that we also fit the general multi-parameter all-effect SLAM to the dataset, while the smallest EBIC given by the multi-parameter model is 7.38×10^4, which is much larger than the best EBIC given by the two-parameter SLAM. So we next focus on the results given by the two-parameter SLAM.



(a) EBIC values and support sizes vs. Υ



(b) attribute structure selected by EBIC

Figure 5.10: Results of TIMSS 2003 8th Grade Data analyzed using two-parameter SLAM.

Figure 5.10(b) plots the attribute structure given by the selected 5 knowledge states. The 13 attributes are separated into five groups $G_1 = \{3, 11, 13\}$, $G_2 = \{5, 9\}$, $G_3 = \{6, 7, 10, 12\}$ and $G_4 = \{1, 2, 8\}$ and $G_5 = \{4\}$, such that the attributes in the same group play the same role in clustering the student population into the five knowledge states. The prerequisite relationships among groups of attributes is also shown in Figure 5.10(b). Attribute (α_3) *compute fluently with multi-digit numbers and find common factors and multiples*, attribute (α_{11}) *compare two fractions with different numerators and different denominators*, attribute (α_{13}) *use equivalent fraction as a strategy to add and subtract fractions*, are the most basic skills in the attribute hierarchy and serve as the prerequisites for all the remaining attributes. Indeed, these three are basic algorithmic operations needed to solve the mathematical problems in the TIMSS test. In addition to the structure selected by EBIC presented in Figure 5.10(b), other attribute structures corresponding to different $\Upsilon \in [0.7, 0.9]$ are presented in Figure D.5 in Appendix B.

Existing works in the literature analyzing the fraction subtraction data and the TIMSS data either make the assumption that all possible configurations of latent attribute patterns exist in the population or pre-specify the attribute structure based on domain experts' judgements (Su et al., 2013). To our knowledge, there has not been a systematic approach to selecting a potentially small set of latent patterns from a high-dimensional space. For the two real datasets, we also find that the EBIC values of the existing EM algorithm are much larger than the proposed method, as indicated in Figures 5.8 and 5.10 when Υ close to 1; thus the proposed method provides a better fit of the two datasets.

5.6 Discussion

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

This work assumes the design matrix Q is prespecified and correct. In practice, if there is reason to suspect that the Q-matrix could be misspecified, then one needs to simultaneously estimate the Q-matrix and learn the attribute patterns from data. Given fixed number of attribute patterns, previous works including Xu and Shang (2018) and Chen et al. (2018a) used the likelihood based methods and the Bayesian methods, respectively, to estimate Q. It is also desirable to develop methods to jointly estimate Q and learn attribute patterns with the existence of large number of attributes. We would like to point out that the identifiability results developed in this work (in Section 5.2) directly apply to this case, and can guarantee both the design matrix Q and the set of significant attribute patterns are learnable from data.

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

CHAPTER VI

Identification and Estimation of Hierarchical Latent Attribute Models

As briefly mentioned in the introduction chapter, Hierarchical Latent Attribute Models (HLAMs) build upon SLAMs by incorporating an additional ingredient: the hierarchical constraints on which configurations of the latent attributes are allowed. HLAMs have connections to other multivariate discrete latent variable models in the machine learning literature, including latent tree graphical models (Choi et al., 2011; Anandkumar et al., 2011; Hsu et al., 2012; Mourad et al., 2013), restricted Boltzmann machines (Hinton, 2002; Hinton and Salakhutdinov, 2006; Salakhutdinov et al., 2007; Larochelle and Bengio, 2008) and restricted Boltzmann forests (RB-Forests) (Larochelle et al., 2010), latent feature models (Ghahramani and Griffiths, 2006; Bernardo et al., 2007; Miller et al., 2009; Yen et al., 2017), sum-product networks (Poon and Domingos, 2011), Probabilistic Sentential Decision Diagrams (PSDD) (Kisa et al., 2014), and cutset networks (Rahman et al., 2014). All these models and HLAMs allow for tractable inference on high-dimensional discrete variables and are closely related. However, HLAMs have two key differences from these models. Other than the structural matrix Q which is unique to the structured latent attribute models, HLAMs additionally incorporate the hierarchical structure among the latent attributes into the model. For instance, in cognitive diagnosis, the possession of certain attributes are often assumed to be the prerequisite for possessing some others (Leighton et al., 2004; Templin and Bradshaw, 2014). Such hierarchical structures differ from the latent tree models in that, the latter use a probabilistic graphical model to model the hierarchical tree structure among latent variables, while in an HLAM the hierarchy is a directed acyclic graph (DAG) encoding hard constraints on allowable configurations of latent attributes. This type of hierarchical constraints in HLAMs have a similar flavor as those of RBForests proposed in Larochelle et al. (2010), though the DAG-structure constraints in an HLAM are more flexible than a forest-structure (i.e., group of trees) one in an RBForest (see Example VI.1).

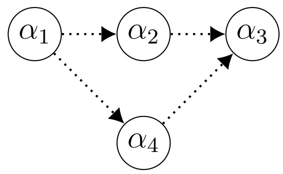
One major issue in the applications of HLAMs is that, the structural matrix and the attribute hierarchy often suffer from potential misspecification by domain experts in confirmatory-type applications, or even entirely unknown in exploratory-type applications. A key question is then how to efficiently learn both the structural Q-matrix and the attribute hierarchy from noisy observations. More fundamentally, it is an important yet open question whether and when the latent structural Q-matrix and the attribute hierarchy are identifiable. Identifiability of HLAMs has a close connection to the uniqueness of tensor decompositions as the probability distribution of an HLAM can be written as a mixture of higher-order tensors. However, related works on identifiability of latent class models and uniqueness of tensor decompositions, such as Allman et al. (2009); Anandkumar et al. (2014, 2015); Bhaskara et al. (2014), cannot be directly applied to HLAMs due to the constraints induced by the structural Q-matrix. To tackle identifiability under such structural constraints, Xu (2017); Xu and Shang (2018); Gu and Xu (2019b, 2020a, 2019a) recently proposed identifiability conditions for latent attribute models. However, Xu (2017); Xu and Shang (2018); Gu and Xu (2019b) considered latent attribute models without any attribute hierarchy; Gu and Xu (2020a) assumed both the structural Q-matrix and true configurations of attribute patterns are known *a priori*; Gu and Xu (2019a) considered the problem of learning the set of truly existing attribute patterns but assumed the structural Q-matrix is correctly specified beforehand. Establishing identifiability without assuming any knowledge of the structural Q-matrix and the attribute hierarchy is a technically much more challenging task and still remains unaddressed in the literature. Moreover, computationally, the existing methods for learning latent attribute models Chen et al. (2015); Xu and Shang (2018); Gu and Xu (2019a) could not simultaneously estimate the structural Q-matrix and the attribute hierarchy.

This chapter has two main contributions. First, we address the challenging identifiability issue of HLAMs. We develop sufficient and almost necessary conditions for identifying the attribute hierarchy, the structural Q-matrix, and the related model parameters in an HLAM. Second, we develop a scalable algorithm to estimate the latent structure and attribute hierarchy of an HLAM. Specifically, we propose a novel approach to simultaneously estimating the structural Q-matrix and performing dimension reduction of attribute patterns. The superior performance of the proposed algorithm is demonstrated in various settings of synthetic data and an application to an educational assessment dataset. The proof of the main theorem and additional numerical results are included in Appendix E.

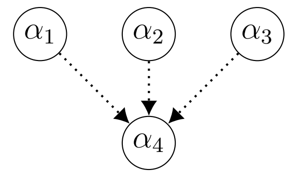
6.1 Hierarchical Latent Attribute Models

This section introduces the model setup of HLAMs in details. An HLAM consists of two types of subject-specific binary variables, the observed responses $\boldsymbol{r} = (r_1, \dots, r_J) \in \{0, 1\}^J$ to J items; and the latent attribute pattern $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K) \in \{0, 1\}^K$. First consider the latent attributes. Attribute α_k is said to be the prerequisite of α_ℓ and denoted by $\alpha_k \to \alpha_\ell$ (or $k \to \ell$), if any $\boldsymbol{\alpha}$ with $\alpha_k = 0$ and $\alpha_\ell = 1$ is "forbidden" to exist. This is a common assumption in applications such as cognitive diagnosis (Leighton et al., 2004; Templin and Bradshaw, 2014). A subject's latent pattern $\boldsymbol{a}$ is assumed to follow a categorical distribution of population proportion parameters ($p_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \{0,1\}^K$), with $p_{\boldsymbol{\alpha}} \geq 0$ and $\sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = 1$. In particular, any pattern $\boldsymbol{\alpha}$ not respecting the hierarchy is deemed impossible to exist with population proportion $p_{\boldsymbol{\alpha}} = 0$. An attribute hierarchy is a set of prerequisite relations between the K attributes, which we denote by $\mathcal{E} = \{k \to \ell :$ attribute k is a prerequisite for $\ell\}$. Any hierarchy $\mathcal{E}$ would induce a set of allowable configurations of attribute patterns, which we denote by $\mathcal{A}$. The set $\mathcal{A}$ is a proper subset of $\{0,1\}^K$ if $\mathcal{E} \neq \varnothing$. So an attribute hierarchy determines the sparsity pattern of the vector of proportion parameters $\boldsymbol{p}$.

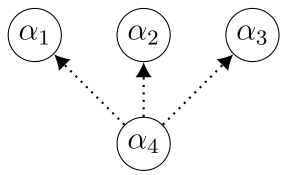
Example VI.1. Figure 6.1 presents several hierarchies with the size of the associated $\mathcal{A}$, where a dotted arrow from α_k to α_ℓ indicates $k \to \ell$. The attribute hierarchy in an HLAM is a DAG generally. In the literature, the RBForests proposed in Larochelle et al. (2010) also introduce hard constraints on allowable configurations of the binary hidden (latent) variables in a restricted Boltzmann machine (RBM). The modeling goal of RBForests is to make computing the probability mass function of observed variables tractable, while not having to limit the number of latent variables. Specifically, in an RBForest, latent variables are grouped in several full and complete binary trees of a certain depth, with variables in a tree respecting the following constraints: if a latent variable takes value zero with $\alpha_i = 0$, then all latent variables in its left subtree must take value d_l; while if $\alpha_i = 1$, all latent variables in its right subtree must take value d_r ($d_l = d_r = 0$ in the paper Larochelle et al. (2010)). The attribute hierarchy model in an HLAM has a similar spirit to RBForests, and actually includes the RBForests as a special case. For instance, the hierarchy in Figure 6.1(d) is equivalent to a tree of depth 3 in an RBForest with $d_l = 1 - d_r = 0$. HLAMs allow for more general attribute hierarchies to encourage better interpretability. Another key difference between HLAMs and RBForests is the different joint model of the observed variables and the latent ones. An RBForest is an extension of an RBM, and they both use the same energy function, while HLAMs model the distribution differently, as to be specified below.



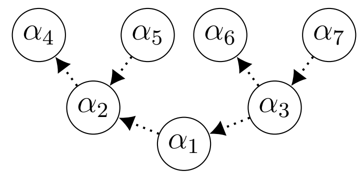
(a) $|\mathcal{A}_1| = 6$



(b) $|\mathcal{A}_2| = 8$



(c) $|\mathcal{A}_3| = 9$

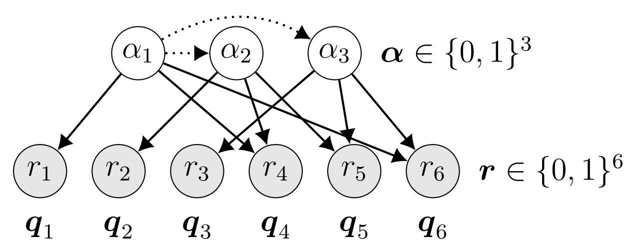


(d) RBForest with $|\mathcal{A}| = 16$

Figure 6.1: Different attribute hierarchies among binary attributes, the first three for $K = 4$ (where $|\{0,1\}^4| = 16$) and the last for $K = 7$ (where $|\{0,1\}^7| = 128$). E.g., the set of allowed attribute patterns under hierarchy (a) is $\mathcal{A}_1 = \{(0000), (1000), (1100), (1001), (1101), (1111)\}$.

$$Q_{6\times 3} := \begin{pmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \boldsymbol{q}_3 \\ \boldsymbol{q}_4 \\ \boldsymbol{q}_5 \\ \boldsymbol{q}_6 \end{pmatrix} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix};$$

(a) binary structural Q-matrix



(b) graphical model along with attribute hierarchy

Figure 6.2: A binary structural Q-matrix and the corresponding graphical model with directed edges from the latent to the observed variables representing dependencies. Below the observed variables in (b) are the row vectors of $Q_{6\times 3}$, i.e., the item loading vectors. There is $\mathcal{E} = \{1 \to 2,\ 1 \to 3\}$.

On top of the model of the latent attributes, an HLAM uses a $J \times K$ binary matrix $Q = (q_{j,k})$ to encode the structural relationship between the J items and the K attributes. In cognitive diagnostic assessments, the matrix Q is often specified by domain experts to summarize which cognitive abilities each test item targets on (Junker and Sijtsma, 2001; von Davier, 2008; George and Robitzsch, 2015). Specifically, $q_{j,k} = 1$ if and only if the response r_j to the jth item has statistical dependence on latent variable α_k. The distribution of r_j, i.e., $\theta_{j,\boldsymbol{\alpha}} := \mathbb{P}(r_j = 1 \mid \boldsymbol{\alpha})$, only depends on its "parent" latent attributes α_k's that are connected to r_j, i.e., $\{\alpha_k : q_{j,k} = 1\}$. The structural matrix Q naturally induces a bipartite graph connecting the latent and the observed variables, with edges corresponding to entries of "1" in $Q = (q_{j,k})$. Figure 6.2 presents an example of a structural matrix Q and its corresponding directed graphical model between the $K = 3$ latent attributes and $J = 6$ observed variables. The solid edges from the latent attributes to the observed variables are specified by $Q_{6\times 3}$. As also can be seen from the graphical model, the observed responses to the J items are conditionally independent given the latent attribute pattern.

In the psychometrics literature, various HLAMs adopting the Q-matrix concept have been proposed with the goal of diagnosing targeted attributes (Junker and Sijtsma, 2001; Templin and Henson, 2006; von Davier, 2008; Henson et al., 2009; de la Torre, 2011). They are often called the cognitive diagnosis models. The general family of latent attribute models are also widely used in other scientific areas including psychiatric evaluation (Templin and Henson, 2006; Jaeger et al., 2006; de la Torre et al., 2018) with the goal of diagnosing patients mental disorders, and epidemiological diagnosis of disease etiology (Wu et al., 2017, 2018; Deloria Knoll et al., 2017; O'Brien et al., 2019). These applications share the common key interest in identifying the multivariate discrete latent attributes.

In this chapter, we focus on two popular and basic types of modeling assumptions under such a framework; as to be revealed soon, these two types of assumptions also have close connections to Boolean matrix decomposition (Ravanbakhsh et al., 2016; Rukat et al., 2017). We would like to point out that this chapter has generalizability beyond these two models. For other more general model assumptions like those considered in Gu and Xu (2019a), our proposed two-stage procedure in Section 6.3 can be easily applied based on their specific likelihood functions (i.e., first reducing dimension and estimating Q by the proposed Alternating Direction Gibbs EM algorithm, and then further shrinking latent patterns; see Section 6.3 for details). Specifically, the HLAMs considered in this paper assume a logical *ideal response* $\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}}$ given an attribute pattern $\boldsymbol{\alpha}$ and an item loading vector $\boldsymbol{q}_j$ in the noiseless case. Then item-level noise parameters are further introduced to account for uncertainty of observations. The following are two popular ways to define the ideal response.

The first is the AND-Model (DINA model in Example I.1) that assumes a conjunctive "and" relationship among the binary attributes. The ideal response of attribute pattern $\boldsymbol{\alpha}$ to item j is

$$\text{(AND-model)} \qquad \Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = \prod_{k=1}^{K} \alpha_k^{q_{j,k}}. \tag{6.1}$$

To interpret, $\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}}$ in (6.1) indicates whether a pattern $\boldsymbol{\alpha}$ possesses all the attributes specified by the item loading vector $\boldsymbol{q}_j$. This conjunctive relationship is often assumed for diagnosis of students' mastery or deficiency of skill attributes in educational assessments, and $\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}}$ naturally indicates whether a student with $\boldsymbol{\alpha}$ has mastered all the attributes required by the test item j. With $\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}}$, the uncertainty of the responses is further modeled by the item-specific Bernoulli parameters

$$\begin{aligned} \theta_j^+ &= 1 - \mathbb{P}(r_j = 0 \mid \Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = 1), \\ \theta_j^- &= \mathbb{P}(r_j = 1 \mid \Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = 0), \end{aligned} \tag{6.2}$$

where $\theta_j^+ > \theta_j^-$ is assumed for identifiability. For each item j, the ideal response $\Gamma_{\boldsymbol{q}_j,\cdot}$, if viewed as a function of attribute patterns, divides the patterns into two latent classes $\{\boldsymbol{\alpha} : \Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = 1\}$ and $\{\boldsymbol{\alpha} : \Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = 0\}$; and for these two latent classes, respectively, the item parameters quantify the noise levels of the response to item j that deviates from the ideal response. Note that the $\theta_{j,\boldsymbol{\alpha}}$ equals either θ_j^+ or θ_j^-. Denote the item parameter vectors by $\boldsymbol{\theta}^+ = (\theta_1^+, \ldots, \theta_J^+)^\top$ and $\boldsymbol{\theta}^- = (\theta_1^-, \ldots, \theta_J^-)^\top$.

The second model is the OR-model (DINO model in Example I.1) that assumes the following ideal response

$$
\begin{aligned}
&\text{(OR-model)} && (6.3)\\
&\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = I(q_{j,k} = \alpha_k = 1 \text{ for at least one } k \in [K]),
\end{aligned}
$$

Such a disjunctive relationship is often assumed in psychiatric measurement. In the Boolean matrix factorization literature, a similar model was proposed by Ravanbakhsh et al. (2016); Rukat et al. (2017). Adapted to the terminology here, Rukat et al. (2017) assumes the ideal response takes the form

$$\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = 1 - \prod_{k=1}^{K}(1 - \alpha_k q_{j,k}),$$

which is equivalent to (6.3), while Rukat et al. (2017) constrains all the item-level noise parameters to be the same.

The last equivalent formulation of the OR-model reveals that its ideal response is symmetric about the two vectors $\boldsymbol{\alpha}$ and $\boldsymbol{q}_j$; while for the AND-model this is not the case. There is an interesting duality (Chen et al., 2015) between the AND-model and the OR-model with $\Gamma^{\text{OR}}_{\boldsymbol{q}_j,\boldsymbol{\alpha}} = 1 - \Gamma^{\text{AND}}_{\boldsymbol{q}_j,\mathbf{1}-\boldsymbol{\alpha}}$. Due to this duality, we next will focus on the asymmetric AND-model without loss of generality.

Due to the duality between the AND-model and the OR-model, we next will focus on the asymmetric AND-model without loss of generality.

6.2 Joint Identifiability of Q-matrix and Attribute Hierarchy

This section presents the main theoretical result on model identifiability. Denote the $J \times |\mathcal{A}|$ ideal response matrix by $\Gamma(Q, \mathcal{A})$. The $\Gamma(Q, \mathcal{A})$ has rows indexed by the J items and columns by attribute patterns in $\mathcal{A}$, and its $(j, \boldsymbol{\alpha})$th entry is defined to be the ideal response $\Gamma_{j,\boldsymbol{\alpha}}$ in (6.1). Given an attribute hierarchy $\mathcal{E}$ and the resulting $\mathcal{A}$, two matrices Q_1 and Q_2 are equivalent if $\Gamma(Q_1, \mathcal{A}) = \Gamma(Q_2, \mathcal{A})$. We also equivalently write it as $Q_1 \overset{\mathcal{E}}{\sim} Q_2$ (or $Q_1 \overset{\mathcal{A}}{\sim} Q_2$). The following example illustrates how an attribute hierarchy determines a set of equivalent Q-matrices.

Example VI.2. Consider the attribute hierarchy $\mathcal{E} = \{1 \to 2,\, 1 \to 3\}$ in Figure 6.2, which results in $\mathcal{A} = \{(000), (100), (110), (101), (111)\}$. The identity matrix I_3 is equivalent to the following matrices under $\mathcal{E}$,

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \overset{\mathcal{E}}{\sim} \begin{pmatrix} 1 & 0 & 0 \\ \mathbf{1} & 1 & 0 \\ \mathbf{1} & 0 & 1 \end{pmatrix} \overset{\mathcal{E}}{\sim} \begin{pmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ * & 0 & 1 \end{pmatrix}, \tag{6.4}$$

where the "$*$"'s in the third matrix above indicate unspecified values, any of which can be either 0 or 1. This equivalence is due to that attribute α_1 serves as the prerequisite for both α_2 and α_3, and any item loading vector measuring α_2 or α_3 is equivalent to a modified one that also measures α_1, in terms of classifying the patterns in $\mathcal{A}$ into two categories $\{\boldsymbol{\alpha} : \Gamma_{\boldsymbol{q}_j, \boldsymbol{\alpha}} = 1\}$ and $\{\boldsymbol{\alpha} : \Gamma_{\boldsymbol{q}_j, \boldsymbol{\alpha}} = 0\}$.

The following main theorem establishes identifiability for an HLAM. See Supplement A for its proof.

Theorem VI.1. *Consider an HLAM under the AND-model assumption with a Q and a hierarchy $\mathcal{E}$.*
(i) $(\Gamma(Q, \mathcal{A})$, $\boldsymbol{\theta}^+$, $\boldsymbol{\theta}^-$, $\boldsymbol{p})$ *are jointly identifiable if the true Q satisfies the following conditions.*

A. *The Q contains a $K \times K$ submatrix Q^0; and setting $Q^0_{j,k}$ to "0" for any $k \to h$ and $Q^0_{j,h} = 1$ results a matrix equal to I_K up to column permutation.*

   (Assume first K rows of Q form Q^0, and denote the remaining submatrix of Q by $Q^\star$.)

B. For any item $j > K$, $q_{j,h} = 1$ and any $k \to h$, we set $q_{j,k}$ to "1" and obtain a modified $Q^{\star,B}$. The $Q^{\star,B}$ contains K distinct column vectors.

C. For any item $j > K$, $q_{j,k} = 1$ and any $k \to h$, we set $q_{j,h}$ to "0" and obtain a modified $Q^{\star,C}$, with entries $q^c_{j,k}$. The $Q^{\star,C}$ satisfies that $\sum_{j=1}^{J-K} q^c_{j,k} \geq 2$ for all $k \in [K]$.

To identify $(\Gamma(Q, \mathcal{A}), \boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$, Condition A is necessary; moreover, Conditions A, B and C are necessary and sufficient when there exists no hierarchy with $p_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$.

(ii) In addition to Conditions A–C, if Q is constrained to contain an I_K, then $(\mathcal{A}, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ are identifiable, and Q can be identified up to the equivalence class under the true $\mathcal{A}$. On the other hand, it is indeed necessary for Q to contain an I_K to ensure an arbitrary $\mathcal{A}$ identifiable.

When estimating an HLAM with the goal of recovering the ideal response structure $\Gamma(Q, \mathcal{A})$ and the model parameters, Theorem VI.1(i) guarantees that Conditions A, B and C suffice and are close to being necessary. While the goal is to uniquely determine the attribute hierarchy from the identified $\Gamma(Q, \mathcal{A})$, the additional condition that Q contains an I_K becomes necessary. This phenomenon can be better understood if one relates it to the identification criteria for the factor loading matrix in factor analysis (Anderson, 2009; Bai and Li, 2012); the loading matrix there is often required to include an identity submatrix or satisfy certain rank constraints, since otherwise the loading matrix can be identifiable only up to a matrix transformation. We point out that developing identifiability theory for HLAMs that can have arbitrarily complex hierarchies is more difficult than the case without hierarchy, and hence Theorem VI.1 is a significant technical advancement over previous works (Gu and Xu, 2019a). We next present an example as an illustration of Theorem VI.1.

Example VI.3. Consider the attribute hierarchy $\{\alpha_1 \to \alpha_2,\, \alpha_1 \to \alpha_3\}$ among $K = 3$ attributes as in Figure 6.2. The following 7×3 structural matrix Q satisfies Conditions A, B and C in Theorem VI.1. In particular, the first 3 rows of Q serves as Q^0 in Condition A. We call the two types of modifications of matrix Q described in Conditions B and C by the name "Operation" B and C, respectively. In the following equation, the matrix entries modified by Operations B and C are highlighted, and the resulting Q^B and Q^C indeed satisfies the requirements in Conditions B and C. So the HLAM associated with Q is identifiable.

$$Q = \begin{pmatrix} & I_3 & \\ \hline 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad \stackrel{\text{Operation } B}{\Longrightarrow} \quad Q^B = \begin{pmatrix} & I_3 & \\ \hline 1 & 1 & 0 \\ \mathbf{1} & 1 & 1 \\ \mathbf{1} & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix};$$

$$Q \quad \stackrel{\text{Operation } C}{\Longrightarrow} \quad Q^C = \begin{pmatrix} & I_3 & \\ \hline 1 & \mathbf{0} & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & \mathbf{0} \end{pmatrix}.$$

6.3 A Scalable Algorithm for Estimating HLAMs

This section presents an efficient two-step algorithm for structure learning of HLAMs. The EM algorithm is popular for estimating latent variable models; however for HLAMs, it needs to evaluate subjects' and items' probabilities of all configurations of K-dimensional patterns in each E step, so it is computationally intractable for moderate to large K with complexity $O((N + J)2^K)$. In this chapter, we propose a scalable two-step algorithm which is able to simultaneously learn the structural matrix Q and latent patterns. Our new first step jointly estimates Q and performs dimension reduction of the latent patterns in a scalable way, with computational complexity $O((N+J)K)$. Then based on the estimated Q and candidate patterns, our second step imposes further regularization on proportions of patterns to extract the set of truly existing patterns and the corresponding attribute hierarchy.

For a sample of size N, denote by $\boldsymbol{R} = (r_{i,j})$ the $N \times J$ matrix containing the N subjects' response vectors as rows, and denote by $\mathbf{A} = (a_{i,k})$ the $N \times K$ matrix containing subjects' latent attribute patterns as rows. Our first step treats $(Q, \mathbf{A})$ as random effect variables with noninformative marginal distributions and $\boldsymbol{\Theta} = (\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ as fixed effect parameters. The log-likelihood of the complete data, $\boldsymbol{R} = (r_{i,j})_{N\times J}$ and $(Q, \mathbf{A})$, is as follows under the AND-model,

$$\ell^{1\text{st}}(\boldsymbol{\Theta};\, Q,\, \mathbf{A},\, \boldsymbol{R}) = \sum_{i=1}^{N}\sum_{j=1}^{J}\Big[r_{i,j}\Big(\prod_{k} a_{i,k}^{q_{j,k}} \log\theta_j^+ + \tag{6.5}$$

$$(1-\prod_k a_{i,k}^{q_{j,k}})\log\theta_j^-\Big)\Big] + (1-r_{i,j})\Big(\prod_k a_{i,k}^{q_{j,k}}\log(1-\theta_j^+) + (1-\prod_k a_{i,k}^{q_{j,k}})\log(1-\theta_j^-)\Big)\Big].$$

We develop a stochastic EM algorithm for structure learning. In particular, in the E step, we use M Gibbs samples of $(Q, \mathbf{A})$ to stochastically approximate their target posterior expectation; then in the M step we update the estimates of the item parameters $\boldsymbol{\Theta}$. We call the algorithm EM with Alternating Direction Gibbs (ADG-EM) as each E step iteratively draws Gibbs samples of $\mathbf{A}$ (along the direction of updating attribute patterns) and Q (along the direction of updating item loadings). The details of ADG-EM are presented in Algorithm 4. In practice we draw $2M$ samples of $(Q, \mathbf{A})$ with the first M as burn-in in each E step; we find usually a small number M suffices for good performance and $M = 3$ is taken in the experiments. Algorithm 4 has a desirable property of performing dimension reduction to obtain a set of candidate patterns, as can be seen from its last step of including all the unique row vectors of the matrix $I(\mathbf{A}^{\text{ave}} > 1/2)$ in the $\mathcal{A}_{\text{candi}}$. This is because the matrix $\mathbf{A}^{\text{ave}}$ has size

$N \times K$, which means the number of selected candidate patterns can be at most N, no matter how large 2^K might be. Indeed, in the experimental setting with $K = 15$ in Section 6.4, the $2^K = 32768 \gg N = 1200$, while the proposed algorithm successfully reduces $|\mathcal{A}_{\text{candi}}|$ to several hundreds (see Table 6.1), and then estimates true latent structure with good accuracy and scalability.

After using Algorithm 4 to obtain the estimated structural matrix $\widehat{Q}$ and a set of candidate latent patterns $\mathcal{A}_{\text{candi}}$, we further impose penalty on the proportion parameters of these candidate patterns $\boldsymbol{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}_{\text{candi}})$ for sparse estimation. Denote $\boldsymbol{\Theta} = (\theta_{j,\boldsymbol{\alpha}} : j \in [J], \boldsymbol{\alpha} \in \mathcal{A}_{\text{candi}})$. Motivated by Gu and Xu (2019a), the second stage maximizes the following objective,

$$\ell_{\lambda}^{\text{2nd}}(\boldsymbol{p}, \boldsymbol{\Theta}; \boldsymbol{R}, \widehat{Q}) = \lambda \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{candi}}} \log_{\rho}(p_{\boldsymbol{\alpha}}) + \sum_{i=1}^{N} \log \Big\{ \sum_{\boldsymbol{\alpha} \in \mathcal{A}_{\text{candi}}} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{r_{i,j}} (1 - \theta_{j,\boldsymbol{\alpha}})^{1 - r_{i,j}} \Big\} \tag{6.6}$$

where $\lambda \in (-\infty, 0)$ is a tuning parameter encouraging sparsity of $(p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}_{\text{candi}})$, and $\rho \asymp N^{-d}$ for some $d \geq 1$ is used to avoid the singularity issue of the log function at zero. Note that the $\widehat{Q}$ estimated by Algorithm 4 implicitly appears in the above (6.6), because it determines the ideal response of patterns to items and further determines whether a $\theta_{j,\boldsymbol{\alpha}}$ should equal θ_j^+ or θ_j^-. To maximize (6.6), we apply the Penalized EM (PEM) algorithm proposed in Gu and Xu (2019a) to obtain the set of selected latent patterns $\mathcal{A}_{\text{final}}$. The PEM algorithm has complexity $O(N|\mathcal{A}_{\text{candi}}|)$ in each E step, thanks to the dimension reduction of ADG-EM algorithm in the first stage. We also use the Extended Bayesian Information Criterion (EBIC) (Chen and Chen, 2008) to select the tuning parameter λ and obtain the best set of attribute patterns $\mathcal{A}_{\text{final}}$. Then finally, the attribute hierarchy $\widehat{\mathcal{E}}$ can be determined by examining the order between columns of the $|\mathcal{A}_{\text{final}}| \times K$ binary matrix $\mathbf{D}$ containing the selected patterns. Denote the columns of the matrix $\mathbf{D}$ by $\mathbf{D}_{\cdot,k}$'s where $k \in [K]$. Specifically, if $\mathbf{D}_{\cdot,k} \succeq \mathbf{D}_{\cdot,\ell}$, then $\{k \to \ell\}$ should be included in $\widehat{\mathcal{E}}$. Combined with the proposed Algorithm 4, the final output is $(\widehat{Q}, \mathcal{A}_{\text{final}}, \widehat{\mathcal{E}})$, including the structural matrix Q, latent patterns, and attribute hierarchy.

We make several remarks about the proposed algorithm. First, in terms of computational complexity, Algorithm 4 has $O((N+J)K)$ complexity in each iterative step, in contrast to the $O((N+J)2^K)$ complexity of the regularized EM algorithm that evaluates the probabilities of all the 2^K configurations of the binary attribute patterns (Chen et al., 2015; Xu and Shang, 2018). This reduction to linear complexity in K is remarkable in the literature of estimating latent attribute models. Second, our algorithm is indeed the first in the literature to simultaneously estimate both $\mathbf{Q}$ and the attribute hierarchy, which itself is a methodological advancement since both quantities are of interest to practitioners; while that in Chapter V (Gu and Xu, 2019a) estimates the latent attribute patterns assuming $\mathbf{Q}$ is known as input.

We also remark that it is straightforward to handle missing data in an HLAM and still perform structure learning. Indeed, it suffices to replace the objective functions (6.5) and (6.6) that are over the $\{r_{i,j} : (i,j) \in [N] \times [J]\}$ by functions over $\{r_{i,j} : (i,j) \in \Omega\}$, where $\Omega \subseteq [N] \times [J]$ is the set of indices in $\boldsymbol{R}$ corresponding to those observed entries. Supplement B contains more details on computation.

6.4 Simulations and Real Data Analysis

Simulations. We perform simulations in two different settings, the first having relatively small J with $(N,J) = (1200, 120)$ and the second having relatively large J with $(N,J) = (1200, 1200)$. Two different numbers of attributes $K = 8$ and $K = 15$ are considered. We next specify the structural matrices $Q_{120\times K}$ and $Q_{1200\times K}$ used to generate the synthetic data. Let $Q^1 = (q^1_{k,\ell})$ be a $K \times K$ matrix with $q^1_{k,\ell} = 1$ if $\ell = k$ or $\ell = k+1$ and zero otherwise; then let $Q_{\text{block}} = (I_K, Q_1^\top, Q_1^\top)^\top$ be a $3K \times K$ matrix that consists of one submatrix I_K and two copies of Q^1. Under $J = 120$ or 1200, the $Q_{J\times K}$

Algorithm 4: ADG-EM: Alternating Direction Gibbs EM for Q estimation and dimension reduction

Data: Binary response matrix $\boldsymbol{R} = (r_{i,j})_{N\times J}$.
Initialize $\mathbf{A} = (a_{i,k})_{N\times K} \in \{0,1\}^{N\times K}$ and $Q = (q_{j,k})_{J\times K} \in \{0,1\}^{J\times K}$.
Initialize parameters $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$. Set $t = 1$, $\mathbf{A}^{\text{ave}} = \mathbf{0}$.
while *not converged* **do**
  for $(i,j) \in [N]\times[J]$ **do**
  $\psi_{i,j} \leftarrow r_{i,j}\log[\theta_j^+/\theta_j^-] + (1-r_{i,j})\log[(1-\theta_j^+)/(1-\theta_j^-)]$;
  $\mathbf{A}^{\text{s}} \leftarrow \mathbf{0}, \quad Q^{\text{s}} \leftarrow \mathbf{0}$.
  for $r \in [2M]$ **do**
    for $(i,k) \in [N]\times[K]$ **do** Draw
    $a_{i,k} \sim \text{Bernoulli}\Big(\sigma\Big(\sum_j q_{j,k}\prod_{m\neq k} a_{i,m}^{q_{j,m}}\psi_{i,j}\Big)\Big)$;
    if $r > M$ **then** $\mathbf{A}^{\text{s}} \leftarrow \mathbf{A}^{\text{s}} + \mathbf{A}$;
  $\mathbf{A}^{\text{ave}} \leftarrow \frac{1}{t}\mathbf{A}^{\text{s}}/M + \Big(1-\frac{1}{t}\Big)\mathbf{A}^{\text{ave}}; \quad t \leftarrow t+1.$
  for $r \in [2M]$ **do**
    for $(j,k) \in [J]\times[K]$ **do** Draw
    $q_{j,k} \sim \text{Bernoulli}\Big(-\sigma\Big(\sum_i (1-a_{i,k})\prod_{m\neq k} a_{i,m}^{q_{j,m}}\psi_{i,j}\Big)\Big)$;
    if $r > M$ **then** $Q^{\text{s}} \leftarrow Q^{\text{s}} + Q$;
  $Q = I(Q^{\text{s}}/M > \frac{1}{2})$ element-wisely; $\mathbf{I}^{\text{ave}} = \Big(\prod_k \{a_{i,k}^{\text{ave}}\}^{q_{j,k}}\Big)_{N\times J}$;
  for $j \in [J]$ **do** $\theta_j^+ \leftarrow (\sum_i r_{i,j} I_{i,j}^{\text{ave}})/(\sum_i I_{i,j}^{\text{ave}})$,
  $\theta_j^- \leftarrow (\sum_i r_{i,j}(1-I_{i,j}^{\text{ave}}))/(\sum_i (1-I_{i,j}^{\text{ave}}))$;
$\widehat{\mathbf{A}} = I(\mathbf{A}^{\text{ave}} > \frac{1}{2})$ element-wisely.
Output: $\mathcal{A}_{\text{candi}}$ containing the unique row vectors of $\widehat{\mathbf{A}}$, and binary structural matrix $\widehat{Q}$.
Then $(\widehat{Q}, \mathcal{A}_{\text{candi}})$ are fed to the Penalized EM algorithm in Gu and Xu (2019a) to maximize (6.6) and obtain $\mathcal{A}_{\text{final}}$.

vertically stacks an appropriate number of Q_{block}. The algorithm is implemented in Matlab. For all the scenarios, 200 independent runs are carried out. The second step PEM algorithm is always run over a range of $\lambda \in \{-0.2\times i:\ i = 1,\ldots,20\}$, from which EBIC selects the best. Figure 6.3 presents two particular hierarchies among $K = 8$ attributes, the diamond and the tree, together with the hierarchy estimation results. More extensive simulation results are presented in Table 6.1. In the table, the column "acc$[Q]^{\mathcal{A}}$" records the average accuracy of estimating the structural Q-matrix up to the equivalence class under $\mathcal{A}$, as illustrated in Example VI.2; the "TPR" denotes

True Positive Rate, the average proportion of true patterns that are selected in $\mathcal{A}_{\text{final}}$; and "1−FDR" denotes "1−False Discovery Rate (FDR)", the average proportion of selected patterns in $\mathcal{A}_{\text{final}}$ that are true. In terms of running time, in scenarios of Table 1, with $2^K = 2^8$ or 2^{15}, $J = 120$ and noise rate 20%, the proposed algorithm takes < 30 seconds on average; even for challenging cases with $(2^K, J) = (2^{15}, 1200)$, the running time is around 1 minute. In contrast, algorithms in previous works (Chen et al., 2015; Xu and Shang, 2018) with exponential complexity in K usually take > 10 minutes even for $(2^K, J) = (2^5, 30)$.

Results in Table 6.1 not only demonstrate the algorithm's excellent performance, but also provide interesting insight into the differences between the two settings, (I) $(N, J) = (1200, 120)$ and (II) $(N, J) = (1200, 1200)$. In setting (I), the first stage ADG-EM algorithm tends to produce a relatively large number of candidate patterns $|\mathcal{A}_{\text{candi}}|$ (though definitely below sample size N, even for $2^K = 2^{15}$), and the second stage PEM algorithm significantly reduces the number of patterns, usually yielding $|\mathcal{A}_{\text{final}}| = |\mathcal{A}_0|$. In contrast, in setting (II), Algorithm 4 itself usually can successfully reduce the number of candidate patterns, giving $|\mathcal{A}_{\text{candi}}|$ close to $|\mathcal{A}_0|$, and the PEM algorithm does not seem to improve the selection results very much in such scenarios. One explanation for this phenomenon is that in the small J case, there are not enough items "measuring" subjects' latent attributes, so the ADG-EM algorithm is not very sure about which false attribute patterns to exclude (very nicely, ADG-EM does not tend to exclude truly existing patterns), and further regularization of patterns in the PEM algorithm is very necessary and helpful; while in the large J case, there exists enough information about the subjects extracted by the large number of items, and hence the ADG-EM can be more confident about discarding those non-existing patterns in the data. Inspired by this observation, we also apply the ADG-EM algorithm to the task of factorization and reconstruction of large and noisy binary matrices. Supplement C contains some interesting simulations.

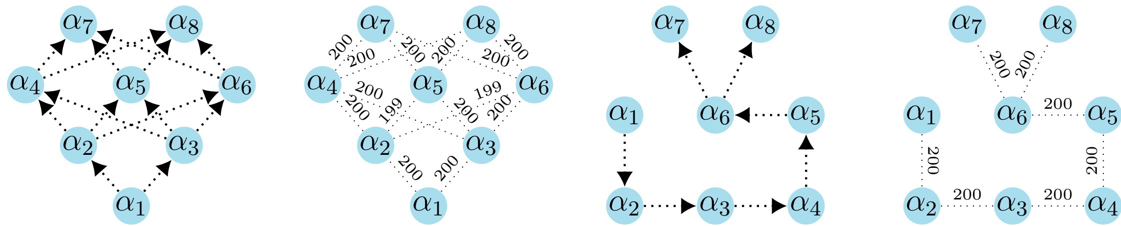


Figure 6.3: Among $K = 8$ attributes, on the upper-left is a diamond shape hierarchy, resulting in 15 patterns; and on the bottom-left is a tree shape hierarchy resulting in 10 patterns. The upper-right and bottom-right plots show how many times out of the 200 runs each true prerequisite relation is successfully recovered. The setting is $(N, J) = (1200, 1200)$ and $1 - \theta_j^+ = \theta_j^- = 20\%$.

Table 6.1: Accuracy of learning the structural Q-matrix and the attribute hierarchy. The "noise" in the table refers to the value of $1 - \theta_j^+ = \theta_j^-$. Numbers in the column "size" record the *median values* of the cardinality of $|\mathcal{A}_{\text{final}}|$ (and $|\mathcal{A}_{\text{candi}}|$ in the parenthesis), based on 200 runs in each scenario.

2^K	$\lvert\mathcal{A}_0\rvert$	noise	$(N, J) = (1200, 120)$			
			acc$[Q]^{\mathcal{A}}$	TPR	1-FDR	size
2^8	10	20%	1.00	1.00	0.96	10 (113)
		30%	1.00	1.00	0.96	10 (166)
	15	20%	1.00	1.00	0.95	15 (120)
		30%	1.00	0.99	0.94	16 (179)
2^{15}	10	20%	0.98	0.91	0.90	10 (272)
		30%	0.99	1.00	0.88	10 (851)
	15	20%	0.99	0.96	0.95	15 (309)
		30%	0.99	0.99	0.89	15 (894)

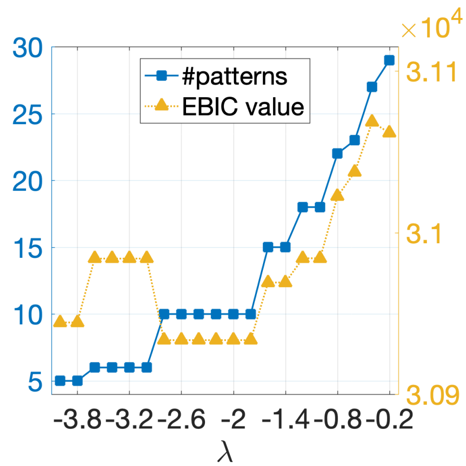
2^K	$\lvert\mathcal{A}_0\rvert$	noise	$(N, J) = (1200, 1200)$			
			acc$[Q]^{\mathcal{A}}$	TPR	1-FDR	size
2^8	10	20%	1.00	1.00	1.00	10 (10)
		30%	1.00	1.00	0.68	15 (16)
	15	20%	1.00	1.00	1.00	15 (15)
		30%	1.00	1.00	0.80	19 (20)
2^{15}	10	20%	0.99	0.99	0.97	10 (11)
		30%	0.97	0.94	0.62	15 (28)
	15	20%	1.00	1.00	0.99	15 (16)
		30%	0.99	0.98	0.71	21 (41)

Real data analysis. We use the proposed method to analyze real data from a large-scale educational assessment, the Trends in International Mathematics and Science Study (TIMSS). This dataset is part of the TIMSS 2011 Austrian data, which was also used in (George and Robitzsch, 2015) to analyze students' abilities in mathematical sub-competences and can be found in the `R` package `CDM`. For this real dataset, there seems no widely-accepted domain knowledge regarding the attribute hierarchy, and our study here provides an exploratory analysis. It includes responses of $N = 1010$ Austrian fourth grade students and $J = 47$ items. A number of $K = 9$ attributes is pre-specified in George and Robitzsch (2015), together with a tentative Q-matrix. One structure specific to such large scale assessments is that only a subset of all items in the entire study is presented to each of students (George and Robitzsch, 2015). This results in many missing values in the $N \times J$ data matrix, and the considered dataset has a missing rate 51.73%. After running the ADG-EM algorithm with missing data firstly, there is $|\mathcal{A}_{\text{candi}}| = 384$, out of the $2^K = 512$ possible patterns. Figure 6.4(a) presents the results of the second stage PEM algorithm. The smallest EBIC value is achieved when $\lambda \in [-2.8, -1.8]$, with 10 estimated latent patterns in $\mathcal{A}_{\text{final}}$ presented in Figure 6.4(b). The hierarchy corresponding to $\mathcal{A}_{\text{final}}$ in Figure 6.4(c) reveals that attribute α_{GR} "Geometry & Reasoning" has the largest number of prerequisites. And in general, attributes related to either "reasoning" or "geometry" seem to be higher level skills in the hierarchy.

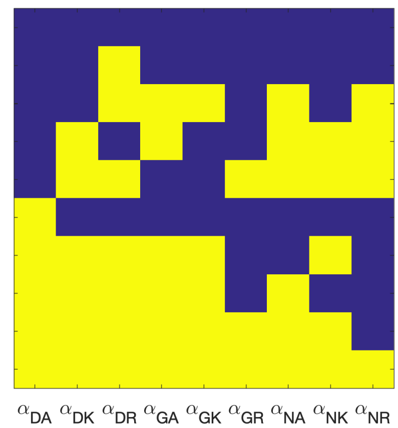
6.5 Discussion

In this chapter, we have proposed transparent conditions on the structural matrix Q for identifying an HLAM and developed a scalable algorithm for estimating an HLAM. The algorithm has great empirical performance on both small- and large-scale structure learning tasks. We next make a remark about the comparison between the new ADG-EM algorithm and the screening algorithm in Gu and Xu (2019a), which takes a known Q as input. Some additional experiments reveal that, (a) when $\mathbf{Q}$ is correctly specified as input to the algorithm in Gu and Xu (2019a), the new ADG-EM has as high accuracy of estimating latent patterns *as* that in Gu and Xu (2019a), in addition to also giving an accurately estimated $\widehat{\mathbf{Q}}$; in this case the ADG-EM takes a little longer due to the additional estimation of $\mathbf{Q}$; (b) while if $\mathbf{Q}$ is misspecified, the algorithm in Gu and Xu (2019a) often has convergence issues due to the misspecification of the latent structural matrix $\mathbf{Q}$; in contrast, the new ADG-EM algorithm is able to take a misspecified $\mathbf{Q}$ as an initial value and then iterates towards convergence to the correct $\mathbf{Q}$ and attribute hierarchy with high accuracy. As about algorithmic robustness, our proposed algorithm is not limited to an identifiable model and can generally be applied to any HLAM where both $\mathbf{Q}$ and attribute patterns are unknown. If, however, the model does not satisfy the proposed identifiability conditions, then the strongest possible identification argument for any estimation method would be partial identifiability (Gu and Xu, 2020a). In this case, the proposed algorithm can still be applied to estimate those parameters up to partial identifiability.

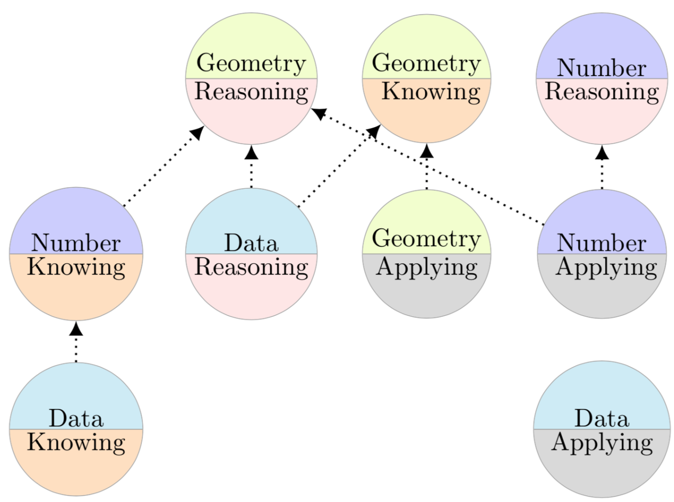
This chapter focuses on basic types of HLAMs that have two item-specific parameters per item, i.e., two-parameter models. It would be interesting to generalize the theory and algorithm to other latent attribute models. More broadly, this chapter makes an attempt to bridge the two fields of psychometrics and machine learning. In psychometrics, various latent attribute models have been recently proposed, which carry good scientific interpretability in the underlying latent structure; while in machine learning, relevant latent variable models including RBM and its extensions are popular, which enjoy computational efficiency. This chapter sheds light on further research that can combine strengths from both fields to analyze large and complex datasets from educational and psychological assessments.



(a) #patterns and EBIC



(b) yellow for 1, blue for 0



(c) attribute hierarchy of $\mathcal{A}_{\text{final}}$

Figure 6.4: Results of Austrian TIMSS 2011 data.

APPENDICES

APPENDIX A

Appendix of Chapter II

This is the appendix to Chapter II and it is organized as follows. Section A.1 presents the proof of the main result Theorem II.1. Appendix A.2 gives the derivation of Equation (2.6) in Example II.2. Appendix A.3 presents the proof of Corollary V.1. Appendix A.4 presents the proof of Proof of Proposition A.2. Appendix A.5 presents the proof of Lemma A.1.

A.1 Proof of Theorem II.1

To study model identifiability, directly working with (5.9) is technically challenging. To facilitate the proof of the theorem, we introduce a key technical quantity following that of Xu (2017), the marginal probability matrix called the T-matrix. We first introduce two new notations, $\boldsymbol{\theta}^+ = \mathbf{1} - \boldsymbol{s}$ and $\boldsymbol{\theta}^- = \boldsymbol{g}$. The T-matrix $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$, is a defined as a $2^J \times 2^K$ matrix, where the entries are indexed by row index $\boldsymbol{r} \in \{0,1\}^J$ and column index $\boldsymbol{\alpha}$. Suppose that the columns of $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ indexed by $(\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^{2^K})$ are arranged in the following order of $\{0,1\}^K$

$$\boldsymbol{\alpha}^1 = \mathbf{0},\ \boldsymbol{\alpha}^2 = \boldsymbol{e}_1,\ \ldots,\ \boldsymbol{\alpha}^{K+1} = \boldsymbol{e}_K,\ \boldsymbol{\alpha}^{K+2} = \boldsymbol{e}_1 + \boldsymbol{e}_2,\ \boldsymbol{\alpha}^{K+3} = \boldsymbol{e}_1 + \boldsymbol{e}_3,\ \ldots,$$
$$\boldsymbol{\alpha}^{2^K} = \sum_{k=1}^{K} \boldsymbol{e}_k = \mathbf{1},$$

where $\mathbf{0}$ denotes the column vector of zeros, $\mathbf{1}$ denotes the column vector of ones, and $\boldsymbol{e}_k$ denotes a standard basis vector, whose kth element is one and the rest are zero; to simplify notation, we omit the dimension indices of $\mathbf{0}, \mathbf{1}$ and $\boldsymbol{e}_k$'s. Similarly, suppose that the rows of $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ indexed by $(\boldsymbol{r}^1, \ldots, \boldsymbol{r}^{2^J})$ are arranged in the following order

$$\boldsymbol{r}^1 = \mathbf{0},\ \boldsymbol{r}^2 = \boldsymbol{e}_1,\ \ldots,\ \boldsymbol{r}^{J+1} = \boldsymbol{e}_J,\ \boldsymbol{r}^{J+2} = \boldsymbol{e}_1 + \boldsymbol{e}_2,\ \boldsymbol{r}^{J+3} = \boldsymbol{e}_1 + \boldsymbol{e}_3,\ \ldots,$$
$$\boldsymbol{r}^{2^J} = \sum_{j=1}^{J} \boldsymbol{e}_j = \mathbf{1}.$$

The $\boldsymbol{r} = (r_1, \ldots, r_J)$th row and $\boldsymbol{\alpha}$th column element of $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$, denoted by $t_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$, is the probability that a subject with attribute profile $\boldsymbol{\alpha}$ answers all items in the subset $\{j : r_j = 1\}$ positively, that is, $t_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-) = P(\boldsymbol{R} \succeq \boldsymbol{r} \mid Q, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\alpha})$. When $\boldsymbol{r} = \mathbf{0}$, $t_{\mathbf{0},\boldsymbol{\alpha}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-) = P(\boldsymbol{r} \succeq \mathbf{0}) = 1$ for any $\boldsymbol{\alpha}$. When $\boldsymbol{r} = \boldsymbol{e}_j$, for $1 \leq j \leq J$, $t_{\boldsymbol{e}_j,\boldsymbol{\alpha}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-) = P(R_j = 1 \mid Q, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\alpha})$. Let $T_{\boldsymbol{r},\bullet}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ be the row vector in the T-matrix corresponding to $\boldsymbol{r}$. Then for any $\boldsymbol{r} \neq \mathbf{0}$, we can write $T_{\boldsymbol{r},\bullet}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-) = \bigodot_{j:r_j=1} T_{\boldsymbol{e}_j,\bullet}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$, where $\odot$ is the element-wise product of the row vectors.

By definition, multiplying the T-matrix by the distribution of attribute profiles $\boldsymbol{p}$ results in a vector, $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p}$, containing the marginal probabilities of successfully responding to each subset of items positively. The $\boldsymbol{r}$th entry of this vector is

$$
\begin{aligned}
&T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)\boldsymbol{p} = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} P(\boldsymbol{R}\succeq\boldsymbol{r}\mid Q,\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{\alpha})p_{\boldsymbol{\alpha}}\\
=&P(\boldsymbol{R}\succeq\boldsymbol{r}\mid Q,\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{p}).
\end{aligned}
$$

We can see that there is a one-to-one mapping between the two 2^J-dimensional vectors $T(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)\boldsymbol{p}$ and $\big(P(\boldsymbol{R}=\boldsymbol{r}\mid Q,\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{p}):\ \boldsymbol{r}\in\{0,1\}^J\big)$. Therefore, Definition 1 directly implies the following proposition.

Proposition A.1. The parameters $(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{p})$ are identifiable if and only if for any $(\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{\theta}}^-,\bar{\boldsymbol{p}})\neq(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{p})$, there exists $\boldsymbol{r}\in\{0,1\}^J$ such that

$$
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)\boldsymbol{p}\neq T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}}. \tag{A.1}
$$

Proposition 1 shows that to establish the identifiability of $(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{p})$, we only need to focus on the T-matrix structure.

The following proposition characterizes the equivalence between the identifiability of the DINA model associated with a Q-matrix with some zero $\boldsymbol{q}$-vectors and that associated with the submatrix of Q containing all of those nonzero $\boldsymbol{q}$-vectors. The proof of Proposition A.2 is given in the Supplementary Material.

Proposition A.2. Suppose the Q-matrix of size $J\times K$ takes the form

$$
Q=\begin{pmatrix} Q' \\ \mathbf{0}\end{pmatrix},
$$

where Q' denotes a $J'\times K$ submatrix containing the J' nonzero $\boldsymbol{q}$-vectors of Q, and $\mathbf{0}$ denotes a $(J-J')\times K$ submatrix containing those zero $\boldsymbol{q}$-vectors of Q. Then the DINA model associated with Q is identifiable if and only if the DINA model associated with Q' is identifiable.

By Proposition A.2, without loss of generality, in the following we assume the Q-matrix does not contain any zero $\boldsymbol{q}$-vectors and prove the necessity and sufficiency of the proposed Conditions 1 and 2.

Proof of Necessity The necessity of Condition 1 comes from Theorem 3 in Xu and Zhang (2016). Now suppose Condition 1 holds but Condition 2 is not satisfied. Without loss of generality, suppose the first two columns in Q^* are the same and the Q takes the following form

$$Q = \begin{pmatrix} & & I_K & & \\ \hdashline \boldsymbol{v} & \boldsymbol{v} & \vdots & \vdots & \vdots \end{pmatrix}_{J\times K}, \tag{A.2}$$

where $\boldsymbol{v}$ is any binary vector of length $J-K$. To show the necessity of Condition 2, from Proposition 1, we only need to find two different sets of parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p}) \neq (\bar{\boldsymbol{s}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}})$ such that for any $\boldsymbol{r} \in \{0,1\}^J$, the following equation holds

$$T_{\boldsymbol{r},\cdot}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p} = T_{\boldsymbol{r},\cdot}(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}}. \tag{A.3}$$

We next construct such $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ and $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{p}})$. We assume in the following that $\bar{\boldsymbol{\theta}}^+ = \boldsymbol{\theta}^+$ and $\bar{\theta}_j^- = \theta_j^-$ for any $j > 2$, and focus on the construction of $(\bar{\theta}_1^-, \bar{\theta}_2^-, \bar{\boldsymbol{p}}) \neq (\theta_1^-, \theta_2^-, \boldsymbol{p})$ satisfying (C.22) for any $\boldsymbol{r} \in \{0,1\}^J$. For notational convenience, we write the positive response probability for item j and attribute profile $\boldsymbol{\alpha}$ in the following general form $\theta_{j,\boldsymbol{\alpha}} := (\theta_j^+)^{\xi_{j,\boldsymbol{\alpha}}}(\theta_j^-)^{1-\xi_{j,\boldsymbol{\alpha}}}$. So based on our construction, for any $j > 2$, $\theta_{j,\boldsymbol{\alpha}} = \bar{\theta}_{j,\boldsymbol{\alpha}}$.

We define two subsets of items S_0 and S_1 to be

$$S_0 = \{j : q_{j,1} = q_{j,2} = 0\} \text{ and } S_1 = \{j : q_{j,1} = q_{j,2} = 1\},$$

where S_0 includes those items not requiring any of the first two attributes, and S_1

includes those items requiring both of the first two attributes. Then since Condition 2 is not satisfied, we must have $S_0 \cup S_1 = \{3, 4, \dots, J\}$, i.e., all but the first two items either fall in S_0 or S_1. Now consider any $\boldsymbol{\alpha}^* \in \{0,1\}^{K-2}$, for any item $j \in S_0$, the four attribute profiles $(0,0,\boldsymbol{\alpha}^*)$, $(0,1,\boldsymbol{\alpha}^*)$, $(1,0,\boldsymbol{\alpha}^*)$ and $(1,1,\boldsymbol{\alpha}^*)$ always have the same positive response probabilities to j, and for any $j \in S_1$, the three attribute profiles $(0,0,\boldsymbol{\alpha}^*)$, $(1,0,\boldsymbol{\alpha}^*)$, $(0,1,\boldsymbol{\alpha}^*)$ always have the same positive response probabilities to j. In summary,

$$\begin{cases} \theta_{j,(0,0,\boldsymbol{\alpha}^*)} = \theta_{j,(0,1,\boldsymbol{\alpha}^*)} = \theta_{j,(1,0,\boldsymbol{\alpha}^*)} = \theta_{j,(1,1,\boldsymbol{\alpha}^*)} & \text{for } j \in S_0; \\ \theta_{j,(0,0,\boldsymbol{\alpha}^*)} = \theta_{j,(0,1,\boldsymbol{\alpha}^*)} = \theta_{j,(1,0,\boldsymbol{\alpha}^*)} \leq \theta_{j,(1,1,\boldsymbol{\alpha}^*)} & \text{for } j \in S_1. \end{cases} \tag{A.4}$$

For any response vector $\boldsymbol{r} \in \{0,1\}^J$ such that $\boldsymbol{r}_{S_1} := (r_j : j \in S_1) \neq \mathbf{0}$, namely $r_j = 1$ for some item j requiring both of the first two attributes, we discuss the following four cases.

(a) For any $\boldsymbol{r}$ such that $(r_1, r_2) = (0,0)$ and $\boldsymbol{r}_{S_1} \neq \mathbf{0}$, from (A.4) and the definition of the T-matrix, (C.22) is equivalent to

$$\begin{aligned} &\sum_{\boldsymbol{\alpha}^*} \Big\{ \Big[\prod_{j>2:\, r_j=1} \theta_{j,(0,0,\boldsymbol{\alpha}^*)} \Big] \big[p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(0,1,\boldsymbol{\alpha}^*)} + p_{(1,0,\boldsymbol{\alpha}^*)} \big] + \\ &\qquad \Big[\prod_{j>2:\, r_j=1} \theta_{j,(1,1,\boldsymbol{\alpha}^*)} \Big] p_{(1,1,\boldsymbol{\alpha}^*)} \Big\} \\ = &\sum_{\boldsymbol{\alpha}^*} \Big\{ \Big[\prod_{j>2:\, r_j=1} \bar{\theta}_{j,(0,0,\boldsymbol{\alpha}^*)} \Big] \big[\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(0,1,\boldsymbol{\alpha}^*)} + \bar{p}_{(1,0,\boldsymbol{\alpha}^*)} \big] \\ &\qquad + \Big[\prod_{j>2:\, r_j=1} \bar{\theta}_{j,(1,1,\boldsymbol{\alpha}^*)} \Big] \bar{p}_{(1,1,\boldsymbol{\alpha}^*)} \Big\} \\ = &\sum_{\boldsymbol{\alpha}^*} \Big\{ \Big[\prod_{j>2:\, r_j=1} \theta_{j,(0,0,\boldsymbol{\alpha}^*)} \Big] \big[\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(0,1,\boldsymbol{\alpha}^*)} + \bar{p}_{(1,0,\boldsymbol{\alpha}^*)} \big] \\ &\qquad + \Big[\prod_{j>2:\, r_j=1} \theta_{j,(1,1,\boldsymbol{\alpha}^*)} \Big] \bar{p}_{(1,1,\boldsymbol{\alpha}^*)} \Big\}, \end{aligned}$$

where the last equality above follows from $\theta_{j,\boldsymbol{\alpha}} = \bar{\theta}_{j,\boldsymbol{\alpha}}$ for any $j > 2$. To ensure the above equations hold, it suffices to have the following equations satisfied for any $\boldsymbol{\alpha}^* \in \{0,1\}^{K-2}$

$$\begin{cases} p_{(1,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(1,1,\boldsymbol{\alpha}^*)}; \\ p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(1,0,\boldsymbol{\alpha}^*)} + p_{(0,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(1,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(0,1,\boldsymbol{\alpha}^*)}. \end{cases} \tag{A.5}$$

(b) For any $\boldsymbol{r}$ such that $(r_1, r_2) = (1,0)$ and $\boldsymbol{r}_{S_1} \neq \mathbf{0}$, from (A.4) and the definition of the T-matrix, (C.22) can be equivalently written as

$$\begin{aligned} &\sum_{\boldsymbol{\alpha}^*} \Big\{ \Big[\prod_{j>2:\, r_j=1} \theta_{j,\,(0,0,\boldsymbol{\alpha}^*)} \Big] \big[g_1(p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(0,1,\boldsymbol{\alpha}^*)}) + (1-s_1) p_{(1,0,\boldsymbol{\alpha}^*)} \big] \\ &\qquad + \Big[\prod_{j>2:\, r_j=1} \theta_{j,\,(1,1,\boldsymbol{\alpha}^*)} \Big] (1-s_1) p_{(1,1,\boldsymbol{\alpha}^*)} \Big\} \\ = &\sum_{\boldsymbol{\alpha}^*} \Big\{ \Big[\prod_{j>2:\, r_j=1} \theta_{j,\,(0,0,\boldsymbol{\alpha}^*)} \Big] \big[\bar{g}_1(\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(0,1,\boldsymbol{\alpha}^*)}) + (1-s_1) \bar{p}_{(1,0,\boldsymbol{\alpha}^*)} \big] \\ &\qquad + \Big[\prod_{j>2:\, r_j=1} \theta_{j,\,(1,1,\boldsymbol{\alpha}^*)} \Big] (1-s_1) \bar{p}_{(1,1,\boldsymbol{\alpha}^*)} \Big\}. \end{aligned}$$

To ensure the above equation holds, it suffices to have the following equations satisfied for any $\boldsymbol{\alpha}^* \in \{0,1\}^{K-2}$

$$\begin{cases} p_{(1,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(1,1,\boldsymbol{\alpha}^*)}; \\ g_1[p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(0,1,\boldsymbol{\alpha}^*)}] + (1-s_1) p_{(1,0,\boldsymbol{\alpha}^*)} = \\ \qquad \bar{g}_1[\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(0,1,\boldsymbol{\alpha}^*)}] + (1-s_1) \bar{p}_{(1,0,\boldsymbol{\alpha}^*)}. \end{cases} \tag{A.6}$$

(c) For any $\boldsymbol{r}$ such that $(r_1, r_2) = (0,1)$ and $\boldsymbol{r}_{S_1} \neq \mathbf{0}$, by symmetry to the previous case of $(r_1, r_2) = (1,0)$, when the following equations hold for any $\boldsymbol{\alpha}^* \in$

$\{0,1\}^{K-2}$, equation (C.22) is guaranteed to hold

$$\begin{cases} p_{(1,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(1,1,\boldsymbol{\alpha}^*)}; \\ g_2[p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(1,0,\boldsymbol{\alpha}^*)}] + (1-s_2)p_{(0,1,\boldsymbol{\alpha}^*)} = \\ \qquad \bar{g}_2[\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(1,0,\boldsymbol{\alpha}^*)}] + (1-s_2)\bar{p}_{(0,1,\boldsymbol{\alpha}^*)}. \end{cases} \tag{A.7}$$

(d) For any $\boldsymbol{r}$ such that $(r_1, r_2) = (1,1)$ and $\boldsymbol{r}_{S_1} \neq \mathbf{0}$, similarly to the previous cases, equation (C.22) can be equivalently written as

$$\begin{aligned} &\sum_{\boldsymbol{\alpha}^*}\Big\{\Big[\prod_{j>2:\, r_j=1}\theta_{j,\,(0,0,\boldsymbol{\alpha}^*)}\Big]\big[g_1g_2p_{(0,0,\boldsymbol{\alpha}^*)} + (1-s_1)g_2p_{(1,0,\boldsymbol{\alpha}^*)} + g_1(1-s_2)p_{(0,1,\boldsymbol{\alpha}^*)}\big] \\ &\qquad + \Big[\prod_{j>2:\, r_j=1}\theta_{j,\,(1,1,\boldsymbol{\alpha}^*)}\Big](1-s_1)(1-s_2)p_{(1,1,\boldsymbol{\alpha}^*)}\Big\} \\ =&\sum_{\boldsymbol{\alpha}^*}\Big\{\Big[\prod_{j>2:\, r_j=1}\theta_{j,\,(0,0,\boldsymbol{\alpha}^*)}\Big]\big[\bar{g}_1\bar{g}_2\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + (1-s_1)\bar{g}_2\bar{p}_{(1,0,\boldsymbol{\alpha}^*)} + \bar{g}_1(1-s_2)\bar{p}_{(0,1,\boldsymbol{\alpha}^*)}\big] \\ &\qquad + \Big[\prod_{j>2:\, r_j=1}\theta_{j,\,(1,1,\boldsymbol{\alpha}^*)}\Big](1-s_1)(1-s_2)\bar{p}_{(1,1,\boldsymbol{\alpha}^*)}\Big\}. \end{aligned}$$

To ensure the above equation hold, it suffices to have the following equations hold for any $\boldsymbol{\alpha}^* \in \{0,1\}^{K-2}$

$$\begin{cases} p_{(1,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(1,1,\boldsymbol{\alpha}^*)}; \\ g_1g_2p_{(0,0,\boldsymbol{\alpha}^*)} + (1-s_1)g_2p_{(1,0,\boldsymbol{\alpha}^*)} + g_1(1-s_2)p_{(0,1,\boldsymbol{\alpha}^*)} \\ \qquad = \bar{g}_1\bar{g}_2\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + (1-s_1)\bar{g}_2\bar{p}_{(1,0,\boldsymbol{\alpha}^*)} + \bar{g}_1(1-s_2)\bar{p}_{(0,1,\boldsymbol{\alpha}^*)}. \end{cases} \tag{A.8}$$

We further consider those response vectors with $\boldsymbol{r}_{S_1} = \mathbf{0}$. A similar argument gives that, to ensure (C.22) holds for any $\boldsymbol{r}$ with $\boldsymbol{r}_{S_1} = \mathbf{0}$, it suffices to have equations (A.5)–(A.8) hold. Together with the results in cases (a)–(d) discussed above, we know that equations (A.5)–(A.8) are a set of sufficient conditions for (C.22) to hold for any $\boldsymbol{r} \in \{0,1\}^J$. Therefore, to show the necessity of Condition 2, we only need to construct $(\bar{g}_1, \bar{g}_2, \bar{\boldsymbol{p}}) \neq (g_1, g_2, \boldsymbol{p})$ satisfying (A.5)–(A.8), which can be equivalently written as, for any $\boldsymbol{\alpha}^* \in \{0,1\}^{K-2}$, $p_{(1,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(1,1,\boldsymbol{\alpha}^*)}$ and

$$\begin{cases} p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(1,0,\boldsymbol{\alpha}^*)} + p_{(0,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(1,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(0,1,\boldsymbol{\alpha}^*)}; \\ g_1[p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(0,1,\boldsymbol{\alpha}^*)}] + (1-s_1)p_{(1,0,\boldsymbol{\alpha}^*)} \\ \qquad = \bar{g}_1[\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(0,1,\boldsymbol{\alpha}^*)}] + (1-s_1)\bar{p}_{(1,0,\boldsymbol{\alpha}^*)}; \\ g_2[p_{(0,0,\boldsymbol{\alpha}^*)} + p_{(1,0,\boldsymbol{\alpha}^*)}] + (1-s_2)p_{(0,1,\boldsymbol{\alpha}^*)} \\ \qquad = \bar{g}_2[\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + \bar{p}_{(1,0,\boldsymbol{\alpha}^*)}] + (1-s_2)\bar{p}_{(0,1,\boldsymbol{\alpha}^*)}; \\ g_1 g_2 p_{(0,0,\boldsymbol{\alpha}^*)} + (1-s_1)g_2 p_{(1,0,\boldsymbol{\alpha}^*)} + g_1(1-s_2)p_{(0,1,\boldsymbol{\alpha}^*)} \\ \qquad = \bar{g}_1\bar{g}_2\bar{p}_{(0,0,\boldsymbol{\alpha}^*)} + (1-s_1)\bar{g}_2\bar{p}_{(1,0,\boldsymbol{\alpha}^*)} + \bar{g}_1(1-s_2)\bar{p}_{(0,1,\boldsymbol{\alpha}^*)}. \end{cases} \tag{A.9}$$

To construct $(\bar{g}_1, \bar{g}_2, \bar{\boldsymbol{p}}) \neq (g_1, g_2, \boldsymbol{p})$, we focus on the family of parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ such that for any $\boldsymbol{\alpha}^* \in \{0,1\}^{K-2}$,

$$\frac{p_{(0,1,\boldsymbol{\alpha}^*)}}{p_{(0,0,\boldsymbol{\alpha}^*)}} = u \text{ and } \frac{p_{(1,0,\boldsymbol{\alpha}^*)}}{p_{(0,0,\boldsymbol{\alpha}^*)}} = v,$$

where u and v are some positive constants. Next we choose $\bar{\boldsymbol{p}}$ such that for any $\boldsymbol{\alpha}^* \in \{0,1\}^{K-2}$

$$p_{(1,1,\boldsymbol{\alpha}^*)} = \bar{p}_{(1,1,\boldsymbol{\alpha}^*)}, \quad \bar{p}_{(0,0,\boldsymbol{\alpha}^*)} = \bar{\rho} \cdot p_{(0,0,\boldsymbol{\alpha}^*)}, \quad \frac{\bar{p}_{(0,1,\boldsymbol{\alpha}^*)}}{\bar{p}_{(0,0,\boldsymbol{\alpha}^*)}} = \bar{u}, \text{ and } \frac{\bar{p}_{(1,0,\boldsymbol{\alpha}^*)}}{\bar{p}_{(0,0,\boldsymbol{\alpha}^*)}} = \bar{v},$$

for some positive constants $\bar{\rho}$, $\bar{u}$ and $\bar{v}$ to be determined. In particular, we choose $\bar{\rho}$

close enough to 1 and then (A.9) is equivalent to

$$\begin{cases} (1+u+v) = \bar\rho(1+\bar u+\bar v); \\ g_1(1+u)+(1-s_1)v = \bar\rho\ [\ \bar g_1(1+\bar u)+(1-s_1)\bar v\ ]; \\ g_2(1+v)+(1-s_2)u = \bar\rho\ [\ \bar g_2(1+\bar v)+(1-s_2)\bar u\ ]; \\ g_1g_2+g_1(1-s_2)u+(1-s_1)g_2v = \bar\rho\ [\ \bar g_1\bar g_2+\bar g_1(1-s_2)\bar u+(1-s_1)\bar g_2\bar v\ ]. \end{cases} \tag{A.10}$$

For any g_1, g_2, s_1, s_2, u and v, the above system of equations contain 5 free parameters $\bar\rho$, $\bar u$, $\bar v$, $\bar g_1$ and $\bar g_2$, while only have 4 constraints, so there are infinitely many sets of solutions of $(\bar\rho, \bar u, \bar v, \bar g_1, \bar g_2)$ to (A.10). This gives the non-identifiability of $(g_1, g_2, \boldsymbol{p})$ and hence justifies the necessity of Condition 2. □

Proof of Sufficiency It suffices to show that if $T(\boldsymbol\theta^+, \boldsymbol\theta^-)\boldsymbol p = T(\bar{\boldsymbol\theta}^+, \bar{\boldsymbol\theta}^-)\bar{\boldsymbol p}$, then $(\boldsymbol\theta^+, \boldsymbol\theta^-, \boldsymbol p) = (\bar{\boldsymbol\theta}^+, \bar{\boldsymbol\theta}^-, \bar{\boldsymbol p})$. Under Condition 1, Theorem 4 in Xu and Zhang (2016) gives that $\boldsymbol s = \bar{\boldsymbol s}$ and $g_j = \bar g_j$ for $j \in \{K+1, \ldots, J\}$. It remains to show $g_j = \bar g_j$ for $j \in \{1, \ldots, K\}$. To facilitate the proof, we introduce the following lemma, whose proof is given in the Supplementary Material.

Lemma A.1. *Suppose Condition 1 is satisfied. For an item set S, define $\vee_{h\in S}\,\boldsymbol q_h$ to be the vector of the element-wise maximum of the $\boldsymbol q$-vectors in the set S. For any $k \in \{1, \ldots, K\}$, if there exist two item sets, denoted by S_k^- and S_k^+, which are not necessarily nonempty or disjoint, such that*

$$g_h = \bar g_h \text{ for any } h \in S_k^- \cup S_k^+, \text{ and } \vee_{h\in S_k^+}\boldsymbol q_h - \vee_{h\in S_k^-}\boldsymbol q_h = \boldsymbol e_k^\top = (\boldsymbol 0, \underbrace{1}_{\text{column } k}, \boldsymbol 0), \tag{A.11}$$

then $g_k = \bar g_k$.

Suppose the Q-matrix takes the form of (2.1), then under Condition 2, any two different columns of the $(J-K)\times K$ sub-matrix Q^* as specified in (2.1) are distinct.

Before proceeding with the proof, we first introduce the concept of the "lexicographic order". We denote the lexicographic order on $\{0,1\}^{J-K}$, the space of all $(J-K)$-dimensional binary vectors, by "$\prec_{\text{lex}}$". Specifically, for any $\mathbf{A} = (a_1, \ldots, a_{J-K})^\top$, $\boldsymbol{b} = (b_1, \ldots, b_{J-K})^\top \in \{0,1\}^{J-K}$, we write $\mathbf{A} \prec_{\text{lex}} \boldsymbol{b}$ if either $a_1 < b_1$; or there exists some $i \in \{2, \ldots, J-K\}$ such that $a_i < b_i$ and $a_j = b_j$ for all $j < i$. For instance, the following four vectors $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4$ in $\{0,1\}^2$ are sorted in an increasing lexicographic order:

$$\mathbf{A}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \prec_{\text{lex}} \mathbf{A}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \prec_{\text{lex}} \mathbf{A}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \prec_{\text{lex}} \mathbf{A}_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

It is not hard to see that if the K column vectors of the submatrix Q^* are mutually distinct, then there exists a unique way to sort them in an increasing lexicographic order. Thus under Condition 2, there exists a unique permutation $(k_1, k_2, \ldots, k_K)$ of $(1, 2, \ldots, K)$ such that column k_1 has the smallest lexicographic order among the K columns of Q^*, column k_2 has the second smallest lexicographic order, and so on, i.e., $Q^*_{\bullet,k_1} \prec_{\text{lex}} Q^*_{\bullet,k_2} \prec_{\text{lex}} \cdots \prec_{\text{lex}} Q^*_{\bullet,k_K}$. As an illustration, consider the leftmost Q-matrix presented in Example 1, Equation (2.4):

$$Q = \begin{pmatrix} & I_3 & \\ \hdashline 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

then the permutation is $(k_1, k_2, k_3) = (3, 2, 1)$, since the third column of Q^* has the smallest lexicographic order while the first column has the largest. Recall that we denote $\mathbf{A} \succeq \boldsymbol{b}$ if $a_i > b_i$ for all i, and denote $\mathbf{A} \nsucceq \boldsymbol{b}$ otherwise. Then by definition, if $\mathbf{A} \prec_{\text{lex}} \boldsymbol{b}$, then $\mathbf{A} \nsucceq \boldsymbol{b}$ must hold. Therefore for any $1 \leq i < j \leq K$, since $Q_{\bullet,k_i} \prec_{\text{lex}} Q_{\bullet,k_j}$, we must have $Q_{\bullet,k_i} \nsucceq Q_{\bullet,k_j}$. This fact will be useful in the following proof.

Equipped with the permutation $(k_1, \ldots, k_K)$, we first prove $g_{k_1} = \bar{g}_{k_1}$. Define a subset of items

$$S_{k_1}^- = \{j > K : q_{j,k_1} = 0\},$$

which includes those items from $\{K+1, \ldots, J\}$ that do not require attribute k_1. Since $Q^*_{\bullet,k_1}$ is of the smallest lexicographic order among column vectors of Q^*, for any $k \in \{1, \ldots, K\} \backslash \{k_1\}$, we must have $Q^*_{\bullet,k} \npreceq Q^*_{\bullet,k_1}$. Thus, for any $k \in \{1, \ldots, K\} \backslash \{k_1\}$ there must exist some item $j_k \in \{K+1, \ldots, J\}$ such that $q_{j_k,k} = 1 > 0 = q_{j_k,k_1}$, which indicates that the union of the attributes required by items in $S_{k_1}^-$ include all the attributes other than k_1, i.e

$$\vee_{h \in S_{k_1}^-} \boldsymbol{q}_h = (\mathbf{1}, \underbrace{0}_{\text{column } k_1}, \mathbf{1}).$$

We further define $S_{k_1}^+ = \{K+1, \ldots, J\}$. Since $S_{k_1}^-$ and $S_{k_1}^+$ satisfy conditions (A.11) in Lemma A.1 for attribute k_1, we have $g_{k_1} = \bar{g}_{k_1}$.

Next we use the induction method to prove that for $l = 2, \ldots, K$, we also have $g_{k_l} = \bar{g}_{k_l}$. In particular, suppose for any $1 \leq m \leq l-1$, we already have $g_{k_m} = \bar{g}_{k_m}$. Note that each k_l is an integer in $\{1, \ldots, K\}$ that can be viewed as *either* the index of the k_lth attribute *or* the index of the k_lth item. Define a set of items

$$S_{k_l}^- = \{j > K : q_{j,k_l} = 0\} \cup \{k_m : 1 \leq m \leq l-1\}, \tag{A.12}$$

where the set $\{j > K : q_{j,k_l} = 0\}$ contains those items, among the last $J - K$ items, which do not require attribute k_l; while the set $\{k_m : 1 \leq m \leq l-1\}$ contains those items for which we have already established the identifiability of the guessing parameter in steps $m = 1, 2, \dots, l-1$ of the induction method, i.e., $g_{k_m} = \bar{g}_{k_m}$ for $m = 1, \dots, l-1$. Thus for any item $j \in S^-_{k_l}$, we have $g_j = \bar{g}_j$. Namely, $S^-_{k_l}$ includes the items whose guessing parameters have already been identified prior to step l of the induction method. Moreover, we claim

$$\vee_{h\in S^-_{k_l}} \boldsymbol{q}_h = (\mathbf{1}, \underbrace{0}_{\text{column } k_l}, \mathbf{1}). \tag{A.13}$$

This is because for any $1 \le m \le l-1$, the item k_m, whose $\boldsymbol{q}$-vector is $\boldsymbol{e}^\top_{k_m}$, is included in the set $S^-_{k_l}$ and hence attribute k_m is required by the set $S^-_{k_l}$; on the other hand, for any $h \in \{l+1, \dots, K\}$, the column vector $Q^*_{\bullet,k_h}$ is of greater lexicographic order than $Q^*_{\bullet,k_l}$ and hence there must exist some item in $S^-_{k_l}$ that does not require attribute k_l but requires attribute k_h. We further define $S^+_{k_l} = \{K+1, \dots, J\}$. The chosen $S^-_{k_l}$ and $S^+_{k_l}$ satisfy the conditions (A.11) in Lemma A.1 and therefore $g_{k_l} = \bar{g}_{k_l}$.

Now that all the slipping and guessing parameters have been identified, $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p} = T(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}} = T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\bar{\boldsymbol{p}}$. Then the fact that $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ has full column rank, which is shown in the Proof of Theorem 1 in Xu and Zhang (2016), implies $\boldsymbol{p} = \bar{\boldsymbol{p}}$. This completes the proof. □

A.2 Derivation of Equation (2.6) in Example II.2

In Example 2, we claimed that, given the Q-matrix in the following form where there are J_0 items with $\boldsymbol{q}$-vectors being $(0, 0)$ and $J - 2 - J_0$ items with $\boldsymbol{q}$-vectors being $(1, 1)$,

$$Q = \begin{pmatrix} \multicolumn{2}{c}{I_2} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix}_{J\times 2},$$

to construct $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{p}}) \neq (\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ satisfying Equation (5.9) where $\bar{\boldsymbol{s}} = \boldsymbol{s}$, $\bar{g}_j = g_j$ for all $j = 3, \ldots, J$, and $\bar{p}_{(1,1)} = p_{(1,1)}$, it suffices to ensure the Equations (2.6) hold. Now we prove this argument. Following the proof of the necessity of Conditions C1 and C2 in the Appendix, we can obtain the following equations in (A.14) from Equations (A.9) in the main text by replacing $(\alpha_1, \alpha_2, \boldsymbol{\alpha}^*)$ in (A.9) with (α_1, α_2) here, since in this case there are only two attributes. And similarly we have the conclusion that Equation (5.9) holds as long as Equations (A.14) hold,

$$\begin{cases} p_{(0,0)} + p_{(1,0)} + p_{(0,1)} = \bar{p}_{(0,0)} + \bar{p}_{(1,0)} + \bar{p}_{(0,1)}; \\ g_1[p_{(0,0)} + p_{(0,1)}] + (1 - s_1)p_{(1,0)} = \bar{g}_1[\bar{p}_{(0,0)} + \bar{p}_{(0,1)}] + (1 - s_1)\bar{p}_{(1,0)}; \\ g_2[p_{(0,0)} + p_{(1,0)}] + (1 - s_2)p_{(0,1)} = \bar{g}_2[\bar{p}_{(0,0)} + \bar{p}_{(1,0)}] + (1 - s_2)\bar{p}_{(0,1)}; \\ g_1 g_2 p_{(0,0)} + (1 - s_1)g_2 p_{(1,0)} + g_1(1 - s_2)p_{(0,1)} \\ \qquad = \bar{g}_1\bar{g}_2\bar{p}_{(0,0)} + (1 - s_1)\bar{g}_2\bar{p}_{(1,0)} + \bar{g}_1(1 - s_2)\bar{p}_{(0,1)}. \end{cases} \tag{A.14}$$

Adding $p_{(1,1)}$ to both hand sides of the first equation in (A.14), adding $(1 - s_1)p_{(1,1)}$ to the second equation, adding $(1 - s_2)p_{(1,1)}$ to the third equation and adding $(1 - s_1)(1 - s_2)p_{(1,1)}$ to the last equation, we exactly obtain (2.6) in Example 2.

A.3 Proof of Corollary V.1

When the identifiability conditions are satisfied, the maximum likelihood estimators of $\bar{\boldsymbol{\theta}}^{+}, \bar{\boldsymbol{\theta}}^{-}$, and $\widehat{\boldsymbol{p}}$ are consistent as the sample size $N \to \infty$. Specifically, we introduce a 2^J-dimensional empirical response vector

$$\begin{aligned}\boldsymbol{\gamma} \;=\; & \Big\{1, N^{-1}\sum_{i=1}^{N} I(\boldsymbol{r}_i \succeq \boldsymbol{e}_1), \cdots, N^{-1}\sum_{i=1}^{N} I(\boldsymbol{r}_i \succeq \boldsymbol{e}_J),\\ & N^{-1}\sum_{i=1}^{N} I(\boldsymbol{r}_i \succeq \boldsymbol{e}_1 + \boldsymbol{e}_2), \cdots, N^{-1}\sum_{i=1}^{N} I(\boldsymbol{r}_i \succeq \mathbf{1})\Big\}^{\top},\end{aligned}$$

where elements of $\boldsymbol{\gamma}$ are indexed by response vectors arranged in the same order as the rows of the T-matrix. From the definition of the T-matrix and the law of large numbers, we know $\boldsymbol{\gamma} \to T(\boldsymbol{\theta}^{+}, \boldsymbol{\theta}^{-})\boldsymbol{p}$ almost surely as $N \to \infty$. On the other hand, the maximum likelihood estimators $\bar{\boldsymbol{\theta}}^{+}, \bar{\boldsymbol{\theta}}^{-}$, and $\widehat{\boldsymbol{p}}$ satisfy $\|\boldsymbol{\gamma} - T(\bar{\boldsymbol{\theta}}^{+}, \bar{\boldsymbol{\theta}}^{-})\widehat{\boldsymbol{p}}\| \to 0$, where $\|\cdot\|$ is the L_2 norm. Therefore,

$$\|T(\boldsymbol{\theta}^{+}, \boldsymbol{\theta}^{-})\boldsymbol{p} - T(\bar{\boldsymbol{\theta}}^{+}, \bar{\boldsymbol{\theta}}^{-})\widehat{\boldsymbol{p}}\| \to 0$$

almost surely. Then from the proof of Theorem II.1, we can obtain the consistency result that $(\widehat{\boldsymbol{\theta}}^{+}, \widehat{\boldsymbol{\theta}}^{-}, \widehat{\boldsymbol{p}}) \to (\boldsymbol{\theta}^{+}, \boldsymbol{\theta}^{-}, \boldsymbol{p})$ almost surely as $N \to \infty$. □

A.4 Proof of Proposition A.2

Consider a Q-matrix of size $J \times K$ in the form

$$Q = \begin{pmatrix} Q' \\ \mathbf{0} \end{pmatrix}, \tag{A.15}$$

where Q' is of size $J' \times K$ and contains those nonzero $\boldsymbol{q}$-vectors of Q. Recall from the model setup in Section 2 of the main text, for any item $j \in \{J'+1, \ldots, J\}$ which has

$\boldsymbol{q}_j = \mathbf{0}$, the guessing parameter is not needed by the DINA model and for notational convenience, we set $g_j \equiv \bar{g}_j \equiv 0$, so the slipping parameter s_j is the only unknown item parameter associated with such j. Taking the response pattern $\boldsymbol{r} = \boldsymbol{e}_j$ for any item $j \in \{J'+1, \dots, J\}$ in Equation (C.22) gives

$$T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p} = (1 - s_j) \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} = (1 - \bar{s}_j) \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \bar{p}_{\boldsymbol{\alpha}} = T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}},$$

then since $\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \bar{p}_{\boldsymbol{\alpha}} = 1$, we have $s_j = \bar{s}_j$ for any $j \in \{J' + 1, \dots, J\}$.

Now denote $\boldsymbol{s}' = (s_1, \dots, s_{J'})$, $\boldsymbol{g}' = (g_1, \dots, g_{J'})$ and similarly denote $\bar{\boldsymbol{s}}'$, $\bar{\boldsymbol{g}}'$. Denote the $2^{J'} \times 2^K$ T-matrix associated with matrix Q' by $T'(\boldsymbol{s}', \boldsymbol{g}')$. For any response pattern $\boldsymbol{r} = (r_1, \dots, r_{J'}, r_{J'+1}, \dots, r_J) \in \{0,1\}^J$, denote $\boldsymbol{r}' = (r_1, \dots, r_{J'})$ and $(\boldsymbol{r}', \mathbf{0}) = (r_1, \dots, r_{J'}, 0, \dots, 0)$ of length J; then we have

$$\begin{aligned}
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p} &= \left\{T_{(\boldsymbol{r}',\mathbf{0}),\boldsymbol{\cdot}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p}\right\} \prod_{j>J'} (1 - s_j)^{r_j} = \left\{T'_{\boldsymbol{r}',\boldsymbol{\cdot}}(\boldsymbol{\theta}^{+\prime}, \boldsymbol{g}')\boldsymbol{p}\right\} \prod_{j>J'} (1 - s_j)^{r_j}, \\
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{g}})\bar{\boldsymbol{p}} &= \left\{T_{(\boldsymbol{r}',\mathbf{0}),\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}}\right\} \prod_{j>J'} (1 - s_j)^{r_j} = \left\{T'_{\boldsymbol{r}',\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^{+\prime}, \bar{\boldsymbol{g}}')\boldsymbol{p}\right\} \prod_{j>J'} (1 - s_j)^{r_j}.
\end{aligned}$$

Using the above equalities, by Proposition A.1, we have the following equivalent arguments,

$$\begin{aligned}
& (\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p}) \text{ associated with } Q \text{ are identifiable,} \\
\iff\quad & \forall(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{p}}) \neq (\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p}),\ \exists \boldsymbol{r} \in \{0,1\}^J \text{ such that } T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p} \neq T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}}, \\
\iff\quad & \forall(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{p}}) \neq (\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p}),\ \exists \boldsymbol{r}' \in \{0,1\}^{J'} \text{ such that } T'_{\boldsymbol{r}',\boldsymbol{\cdot}}(\boldsymbol{\theta}^{+\prime}, \boldsymbol{g}')\boldsymbol{p} \neq T'_{\boldsymbol{r}',\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^{+\prime}, \bar{\boldsymbol{g}}')\bar{\boldsymbol{p}}, \\
\iff\quad & (\boldsymbol{\theta}^{+\prime}, \boldsymbol{g}', \boldsymbol{p}) \text{ associated with } Q' \text{ are identifiable.}
\end{aligned}$$

Therefore we have shown identifiability of DINA associated with Q in the form of (A.15) is equivalent to that of DINA associated with submatrix Q' in (A.15) and the proof of the proposition is complete.

A.5 Proof of Lemma A.1

To facilitate the proof of the lemma, we introduce the following proposition, which is from Proposition 3 in Xu and Zhang (2016). We first generalize the definition of the T-matrix. For any $\boldsymbol{x} = (x_1, \ldots, x_J)^\top \in \mathbb{R}^J$ and $\boldsymbol{y} = (y_1, \ldots, y_J)^\top \in \mathbb{R}^J$, we still define the T-matrix $T(\boldsymbol{x}, \boldsymbol{y})$ to be a $2^J \times 2^K$ matrix, where the entries are indexed by row index $\boldsymbol{r} \in \{0,1\}^J$ and column index $\boldsymbol{\alpha}$. For any row indexed by $\boldsymbol{e}_j$ with $j = 1, \ldots, J$, we let $t_{\boldsymbol{e}_j,\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) = (1 - x_j)^{\xi_{j,\boldsymbol{\alpha}}} y_j^{1-\xi_{j,\boldsymbol{\alpha}}}$; for any $\boldsymbol{r} \neq \boldsymbol{0}$, let the $\boldsymbol{r}$th row vector of $T(\boldsymbol{x}, \boldsymbol{y})$ be $T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{x}, \boldsymbol{y}) = \bigodot_{j:r_j=1} T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}(\boldsymbol{x}, \boldsymbol{y})$.

Proposition A.3. If $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{p} = T(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}}$, then for any $\boldsymbol{\theta} \in \mathbb{R}^J$, $T(\boldsymbol{\theta}^+ - \boldsymbol{\theta}, \boldsymbol{\theta}^- - \boldsymbol{\theta})\boldsymbol{p} = T(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta})\bar{\boldsymbol{p}}$.

Let G be the set of items whose guessing parameters have been identified in the sense that $g_j = \bar{g}_j$, for any $j \in G$. Let $G^c := \{1, \ldots, J\} \backslash G$ be the complement of G. Note that $\{K+1, \ldots, J\} \cup S_k^- \cup S_k^+ \subseteq G$. Define

$$\boldsymbol{\theta} = \sum_{j \in G^c} (1 - s_j)\boldsymbol{e}_j + \sum_{j \in G} g_j \boldsymbol{e}_j. \tag{A.16}$$

Denote $T := T(\boldsymbol{\theta}^+ = \boldsymbol{1}, \boldsymbol{g} = \boldsymbol{0})$ and denote the $(\boldsymbol{r}, \boldsymbol{\alpha})$-entry of T by $t_{\boldsymbol{r},\boldsymbol{\alpha}}$, then by definition,

$$t_{\boldsymbol{r},\boldsymbol{\alpha}} = \prod_{j:\, r_j = 1} 1^{I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j)} 0^{1 - I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j)} = I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j \ \forall j \text{ s.t. } r_j = 1), \tag{A.17}$$

where $I(\cdot)$ denotes the indicator function. Proposition B.2 implies that $T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ + \boldsymbol{\theta}, \boldsymbol{g} - \boldsymbol{\theta}) = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ + \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta})\bar{\boldsymbol{p}}$ for $\boldsymbol{\theta}$ defined in (A.16). We use $\theta_{j,\boldsymbol{\alpha}}$ to denote the positive response probability of attribute profile $\boldsymbol{\alpha}$ to item j, i.e., $\theta_{j,\boldsymbol{\alpha}} = 1 - s_j$ for $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$, and $\theta_{j,\boldsymbol{\alpha}} = g_j$ for $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_j$. For any response pattern $\boldsymbol{r}$

such that $r_j = 0$ for all $j \in G^c$,

$$
\begin{aligned}
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^{+}-\boldsymbol{\theta},\boldsymbol{\theta}^{-}-\boldsymbol{\theta})\boldsymbol{p} &= \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j\in G}[\theta_{j,\boldsymbol{\alpha}} - g_j]^{r_j} \prod_{j\in G^c}[\theta_{j,\boldsymbol{\alpha}} - (1-s_j)]^{r_j} \\
&= \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j\in G}(\theta_{j,\boldsymbol{\alpha}} - g_j)^{r_j},
\end{aligned}
\tag{A.18}
$$

where in the above summation over $\boldsymbol{\alpha} \in \{0,1\}^K$, one can see that the product term $\prod_{j\in G}(\theta_{j,\boldsymbol{\alpha}} - g_j)^{r_j}$ is nonzero only for those $\boldsymbol{\alpha}$ such that $\theta_{j,\boldsymbol{\alpha}} = 1 - s_j > g_j$ for all j where $r_j = 1$; and when the product term is nonzero, it equals $\prod_{j\in G}(1 - s_j - g_j)^{r_j}$. Further examining those $\boldsymbol{\alpha}$ that make the product term nonzero in (A.18), one can find it is exactly those $\boldsymbol{\alpha}$ such that $t_{\boldsymbol{r},\boldsymbol{\alpha}} = 1$ according to (E.3). Noting that $t_{\boldsymbol{r},\boldsymbol{\alpha}}$ can either be 1 or 0, (A.18) can be further written as

$$
\begin{aligned}
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^{+}-\boldsymbol{\theta},\boldsymbol{\theta}^{-}-\boldsymbol{\theta})\boldsymbol{p} &= \sum_{\boldsymbol{\alpha}:\, t_{\boldsymbol{r},\boldsymbol{\alpha}}=1} p_{\boldsymbol{\alpha}} \prod_{j\in G}(1 - s_j - g_j)^{r_j} \\
&= \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r},\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} \prod_{j\in G}(1 - s_j - g_j)^{r_j}.
\end{aligned}
\tag{A.19}
$$

Following the same argument, we also have

$$
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{s}+\boldsymbol{\theta},\bar{\boldsymbol{g}}-\boldsymbol{\theta})\bar{\boldsymbol{p}} = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r},\boldsymbol{\alpha}} \bar{p}_{\boldsymbol{\alpha}} \prod_{j\in G}(1 - s_j - g_j)^{r_j},
$$

then Proposition B.2 implies

$$
\sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r},\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r},\boldsymbol{\alpha}} \bar{p}_{\boldsymbol{\alpha}}, \text{ for any } \boldsymbol{r} \text{ such that } r_j = 0 \text{ for all } j \in G^c. \tag{A.20}
$$

We then define a response vector $\boldsymbol{r}^* = (r_1^*, \dots, r_J^*)^\top$ to be $\boldsymbol{r}^* = \sum_{j\in G}(1 - q_{j,k})\boldsymbol{e}_j$, that is, $\boldsymbol{r}^*$ has correct responses to and only to those items among the set G that do not require the kth attribute. Let $S_{\boldsymbol{r}^*}$ denote the set of items that $\boldsymbol{r}^*$ has correct responses to, i.e., $S_{\boldsymbol{r}^*} = \{j : r_j^* = 1\}$. Since $S_k^- \subseteq G$ and $q_{j,k} = 0$ for any $j \in S_k^-$, we know $S_{\boldsymbol{r}^*}$ is nonempty. Now consider the row vector in the transformed T-matrix $T(\boldsymbol{s}+\boldsymbol{\theta}, \boldsymbol{g}-\boldsymbol{\theta})$ corresponding to response vector $\boldsymbol{r}^* + \boldsymbol{e}_k$, then we have that $T_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}(\boldsymbol{s}+\boldsymbol{\theta}, \boldsymbol{g}-\boldsymbol{\theta}) \neq 0$ if and only if

$$\boldsymbol{\alpha} \succeq \boldsymbol{q}_j \text{ for any item } j \in S_{\boldsymbol{r}^*}, \text{ and } \alpha_k = 0.$$

In other words, $T_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}(\boldsymbol{s}+\boldsymbol{\theta}, \boldsymbol{g}-\boldsymbol{\theta}) \neq 0$ if and only if $\boldsymbol{\alpha}$ satisfies $t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 1$ and $t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}} = 0$. This implies that

$$\begin{aligned} & T_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}, \boldsymbol{\theta}^- - \boldsymbol{\theta})\boldsymbol{p} \\ & = (g_k + s_k - 1) \prod_{j \in S_{\boldsymbol{r}^*}} (1 - s_j - g_j) \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} (t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} - t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}) p_{\boldsymbol{\alpha}} \end{aligned} \tag{A.21}$$

and

$$\begin{aligned} & T_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}) \cdot \bar{\boldsymbol{p}} \\ & = (\bar{g}_k + s_k - 1) \prod_{j \in S_{\boldsymbol{r}^*}} (1 - s_j - g_j) \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} (t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} - t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}) \bar{p}_{\boldsymbol{\alpha}}. \end{aligned} \tag{A.22}$$

Note that (A.21) = (A.22) by Proposition 2.

We next show that the summation terms in (A.21) and (A.22) satisfy

$$\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} (t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} - t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}) p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} (t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} - t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}) \bar{p}_{\boldsymbol{\alpha}} \neq 0. \tag{A.23}$$

Note $\boldsymbol{r}^*$ satisfies the condition in (D.1) that $r_j^* = 0$ for all $j \in G^c$. Therefore,

$$\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \bar{p}_{\boldsymbol{\alpha}}. \tag{A.24}$$

We further consider the response vector $\boldsymbol{r}^* + \boldsymbol{e}_k$. Under the conditions of Lemma 1, there exists some item $h \in G$ such that

$$q_{h,k} = 1 \text{ and } \{l : q_{h,l} = 1,\ l \neq k\} \subseteq \bigcup_{j \in S_{\boldsymbol{r}^*}} \{l : q_{j,l} = 1\}.$$

That is, the item h requires the kth attribute and h's any other required attribute is also required by some item in the set $S_{\boldsymbol{r}^*}$. Therefore we have $T_{\boldsymbol{r}^*+\boldsymbol{e}_k,\bullet} = T_{\boldsymbol{r}^*\vee\boldsymbol{r}^\#,\bullet}$, where $\boldsymbol{r}^\# := \sum_{h\in S_j^+\setminus S_j^-}\boldsymbol{e}_h$; in addition, since the response vector $\boldsymbol{r}^*\vee\boldsymbol{r}^\#$ satisfies the condition in (D.1) that its jth element $(\boldsymbol{r}^*\vee\boldsymbol{r}^\#)_j = 0$ for any $j\in G^c$, we have

$$\begin{aligned}\sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}\cdot p_{\boldsymbol{\alpha}} &= \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r}^*\vee\boldsymbol{r}^\#,\boldsymbol{\alpha}}\cdot p_{\boldsymbol{\alpha}} \\ &= \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r}^*\vee\boldsymbol{r}^\#,\boldsymbol{\alpha}}\cdot \bar{p}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}\cdot \bar{p}_{\boldsymbol{\alpha}}.\end{aligned} \tag{A.25}$$

The first equation in (A.23) then follows from (A.24) and (A.25). The inequality in (A.23) also holds since $t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \geq t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}$ for any $\boldsymbol{\alpha}$ and $t_{\boldsymbol{r}^*,\boldsymbol{\alpha}} > t_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\alpha}}$ for those $\boldsymbol{\alpha}$ with $\alpha_k = 0$ and $\boldsymbol{\alpha}\succeq\boldsymbol{q}_j$ for any item $j\in S_{\boldsymbol{r}^*}$.

With the results in (A.23), we have $g_k = \bar{g}_k$ from the equality of (A.21) and (A.22). This completes the proof. $\square$

APPENDIX B

Appendix of Chapter III

This is the appendix to Chapter III and it is organized as follows. Appendix B.1 presents the details of establishing model identifiability under Q-matrices associated with real data. Appendices B.2 and B.3 provide the proofs of the main theoretical results for the two-parameter and multi-parameter restricted latent class models in Sections 3.2 and 3.3 of the main text, respectively. Appendix B.4 gives the proofs of the results in Section 3.4 in the main text. Appendix B.5 gives the proofs of some technical lemmas.

B.1 Identifiability under Q-matrices associated with real data

B.1.1 TIMSS Data Q-matrix and its identifiability.

Table 1.2 presents the full 43×12 Q-matrix $Q_{43\times 12}$ for the TIMSS data, which is introduced in Example I.6 of the main text. The Q-matrix was constructed by mathematics educators and researchers and its form was specified in Choi et al. (2015). Please refer to Choi et al. (2015) for more details about the test items and fine-grained attributes. We next show how our theoretical results guarantee $\boldsymbol{p}$-partial identifiability of two-parameter models and generic identifiability of multi-parameter models under this $Q_{43\times 12}$.

$\boldsymbol{p}$-partial identifiability. We show that the $Q_{43\times 12}$ satisfies conditions (C1*) and (C2*). The $Q_{43\times 12}$ in Table 1.2 contains 9 basis items $S_{basis} = \{4, 8, 15, 16, 19, 24, 30, 34, 38\}$ and the remaining 34 non-basis items. We can check that each basis item is S_{non}-differentiable and conditions (C1*) and (C2*) hold. Thus Corollary III.1 implies $\boldsymbol{p}$-partial identifiability of the two-parameter restricted latent class models, and also guarantees estimability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$.

Generic identifiability. We show that the $Q_{43\times 12}$ satisfies conditions (C5) and (C6). In particular, let $S_1 = \{1, 3, 4, 5, 7, 8, 12, 13, 15, 17, 19, 38\}$ and $S_2 = \{2, 11, 16, 20, 22, 23, 24, 26, 30, 31, 33, 34\}$, then items in each of S_1 and S_2 can be arranged in a way such that the sub-Q-matrices Q_1 and Q_2 take the form of (5.5), which implies condition (C5). In addition, each attribute is required by at least one item in $(S_1 \cup S_2)^c$ and thus condition (C6) is also satisfied. Theorem III.7 then gives the generic identifiability of any multi-parameter model associated with this Q-matrix.

B.1.2 Identifiability with the Fraction Subtraction Data Q-matrix.

We next show how our theoretical results guarantee $\boldsymbol{p}$-partial identifiability of two-parameter models and generic identifiability of multi-parameter models under the $Q_{20\times 8}$ associated with the Fraction Subtraction Data.

$\boldsymbol{p}$-partial identifiability. We apply Theorem III.2 since attribute $k = 6$ is only required by two items $\{1, 18\}$ and condition (C1*) is violated. Specifically, we transform the original Q-matrix to the form of (3.15) with $\boldsymbol{v}_1 = (0, 0, 0, 1, 0, 1, 0)$, $\boldsymbol{v}_2 = (0, 1, 0, 0, 1, 1, 0)$ and submatrix Q' as specified in (1.3), by first exchanging the second and the eighteenth rows and then exchanging the first and the sixth columns. The transformed Q-matrix falls into the case (a) of (B.1) in Theorem III.2, and it suffices to show that the Q'-matrix in (1.3) satisfy condition (C1*) and (C2*). We can check that (C1*) holds for Q' that attributes required by each $\boldsymbol{q}$-vector in Q' are repeatedly measured by at least two disjoint sets of other items. In addition, (C2*) is satisfied because Q' only has one basis item $S_{basis}(Q') = \{9\}$ and the item $j = 9$ is S_{non}-differentiable.

$$
Q' = \begin{array}{c} \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \\ 15 \\ 16 \\ 17 \\ 19 \\ 20 \end{array}
\begin{array}{c}
\begin{array}{ccccccc} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_7 & \alpha_8 \end{array} \\
\begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 & 0
\end{pmatrix}
\end{array}
\begin{array}{l} \\
\preceq \boldsymbol{q}_3,\ \boldsymbol{q}_5 \\
\preceq \boldsymbol{q}_2,\ \boldsymbol{q}_5 \\
\preceq \boldsymbol{q}_9,\ \boldsymbol{q}_{20} \\
\preceq \boldsymbol{q}_3 \vee \boldsymbol{q}_4 \vee \boldsymbol{q}_{10},\ \boldsymbol{q}_{12} \vee \boldsymbol{q}_{13} \\
\preceq \boldsymbol{q}_2,\ \boldsymbol{q}_8; \\
\preceq \boldsymbol{q}_{14} \vee \boldsymbol{q}_{15},\ \boldsymbol{q}_{19} \\
\preceq \boldsymbol{q}_2,\ \boldsymbol{q}_6 \\
\preceq \boldsymbol{q}_4,\ \boldsymbol{q}_5; \\
\preceq \boldsymbol{q}_4 \vee \boldsymbol{q}_5,\ \boldsymbol{q}_7 \vee \boldsymbol{q}_{12} \vee \boldsymbol{q}_{13} \\
\preceq \boldsymbol{q}_4 \vee \boldsymbol{q}_5,\ \boldsymbol{q}_7 \vee \boldsymbol{q}_{12} \vee \boldsymbol{q}_{13} \\
\preceq \boldsymbol{q}_4 \vee \boldsymbol{q}_5,\ \boldsymbol{q}_7 \vee \boldsymbol{q}_{12} \vee \boldsymbol{q}_{13} \\
\preceq \boldsymbol{q}_4 \vee \boldsymbol{q}_5,\ \boldsymbol{q}_{13} \\
\preceq \boldsymbol{q}_4,\ \boldsymbol{q}_5 \\
\preceq \boldsymbol{q}_{14} \vee \boldsymbol{q}_{15},\ \boldsymbol{q}_{19} \\
\preceq \boldsymbol{q}_4,\ \boldsymbol{q}_5 \\
\preceq \boldsymbol{q}_{13},\ \boldsymbol{q}_{18} \\
\preceq \boldsymbol{q}_4 \vee \boldsymbol{q}_7,\ \boldsymbol{q}_{15} \vee \boldsymbol{q}_{19} \\
\preceq \boldsymbol{q}_4 \vee \boldsymbol{q}_5,\ \boldsymbol{q}_{13}
\end{array}
\tag{B.1}
$$

Theorem III.2 therefore gives the $\boldsymbol{p}$-partial identifiability of the two-parameter models.

Generic identifiability. We apply Theorem III.8 since attribute 6 is required by only two items, item 6 and item 18. Rearranging the columns and rows of this Q-matrix to the form of (3.17) with $\boldsymbol{v}_1 = (0, 0, 0, 1, 0, 1, 0)$ and $\boldsymbol{v}_2 = (0, 1, 0, 0, 1, 1, 0)$, we have $\boldsymbol{v}_1 \vee \boldsymbol{v}_2 \neq \mathbf{1}_{K-1}$ and the sub-matrix Q' part satisfies conditions (C5) and (C6), so Theorem III.8 gives the generic identifiability of multi-parameter Q-restricted latent class models.

B.2 Proof of Main Results in Section 3.2

In this section we first introduce some technical quantities and their properties which will be useful in later proofs, then present the proofs of the main results in Section 3.2 for two-parameter restricted latent class models.

To facilitate the study of parameter identifiability of restricted latent class models, we consider a marginal probability matrix $T(\boldsymbol{\Theta})$ of size $2^J \times m$ as follows, where $J = |\mathcal{S}|$ denotes the number of items and $m = |\mathcal{A}|$ denotes the number of classes. Rows of $T(\boldsymbol{\Theta})$ are indexed by the 2^J possible response patterns $\boldsymbol{r} = (r_1, \ldots, r_J)^\top \in \{0,1\}^J$ and columns of $T(\boldsymbol{\Theta})$ are indexed by latent classes $\boldsymbol{\alpha} \in \mathcal{A}$, while the $(\boldsymbol{r}, \boldsymbol{\alpha})$th entry of $T(\boldsymbol{\Theta})$, denoted by $T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\Theta})$, represents the marginal probability that subjects in latent class $\boldsymbol{\alpha}$ provide positive responses to the set of items $\{j : r_j = 1\}$, namely

$$T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\Theta}) = P(\boldsymbol{R} \succeq \boldsymbol{r} \mid \boldsymbol{\Theta}, \boldsymbol{\alpha}) = \prod_{j=1}^{J} \theta_{j,\boldsymbol{\alpha}}^{r_j}.$$

Denote the $\boldsymbol{\alpha}$th column vector and the $\boldsymbol{r}$th row vector of the T-matrix by $T_{\boldsymbol{\cdot},\boldsymbol{\alpha}}(\boldsymbol{\Theta})$ and $T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\Theta})$ respectively. Let $\boldsymbol{e}_j$ denote the J-dimensional unit vector with the jth element being one and all the other elements being zero, then any response pattern $\boldsymbol{r}$ can be written as a sum of some $\boldsymbol{e}$-vectors, namely $\boldsymbol{r} = \sum_{j:r_j=1} \boldsymbol{e}_j$. The $\boldsymbol{r}$th element of the 2^J-dimensional vector $T(\boldsymbol{\Theta})\boldsymbol{p}$ is

$$\{T(\boldsymbol{\Theta})\boldsymbol{p}\}_{\boldsymbol{r}} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\Theta})\boldsymbol{p} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\Theta})p_{\boldsymbol{\alpha}} = P(\boldsymbol{R}\succeq\boldsymbol{r}\mid\boldsymbol{\Theta}).$$

Based on the T-matrix, we have the following definition of identifiability for model parameters $(\boldsymbol{\Theta},\boldsymbol{p})$, equivalent to definition (5.9) in Section 2.3 of the main text. The equivalence of the two definitions comes from that two sets of model parameters lead to the same marginal distribution of responses $\{P(\boldsymbol{R}\succeq\boldsymbol{r}\mid\boldsymbol{\Theta}),\forall\boldsymbol{r}\in\{0,1\}^J\}$ if and only if they lead to the same distribution of the responses $\{P(\boldsymbol{R}=\boldsymbol{r}\mid\boldsymbol{\Theta}),\forall\boldsymbol{r}\in\{0,1\}^J\}$.

Proposition B.1. Under a restricted latent class model, the model parameters are identifiable if and only if for any $(\boldsymbol{\Theta},\,\boldsymbol{p})$ and $(\bar{\boldsymbol{\Theta}},\,\bar{\boldsymbol{p}})$,

$$T(\boldsymbol{\Theta})\boldsymbol{p} = T(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}} \tag{B.2}$$

implies $(\boldsymbol{\Theta},\boldsymbol{p}) = (\bar{\boldsymbol{\Theta}},\bar{\boldsymbol{p}})$.

Together with this equivalent definition, the following proposition, which was introduced in Xu (2017), describes an important algebraic property of the T-matrix and will be used in our proofs.

Proposition B.2. For any $\boldsymbol{\theta}^* = (\theta_1,\ldots,\theta_J)^\top \in \mathbb{R}^J$, there exists an invertible lower triangular matrix $D(\boldsymbol{\theta}^*)$ depending solely on $\boldsymbol{\theta}^*$, such that the diagonal elements of $D(\boldsymbol{\theta}^*)$ are all 1, and

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

Another useful property of the T-matrix is given by the following lemma, whose proof is given in Section B.4.

Lemma B.1. *Denote the T-matrix corresponding to a subset of items S by $T(\boldsymbol{\Theta}_S)$, where $\boldsymbol{\Theta}_S = (\theta_{j,\boldsymbol{\alpha}},\ j\in S,\ \boldsymbol{\alpha}\in\mathcal{A})$. If for an item set S, the Γ-matrix Γ^S of size*

$|S| \times m$ is separable, then the corresponding T-matrix $T(\mathbf{\Theta}_S)$ of size $2^{|S|} \times m$ has full column rank m.

Equipped with the above developments, now we are ready to prove the main results.

Proof of Proposition III.1 and Proposition III.2. When $\mathbf{\Theta}$ is known, by Proposition (B.1), we only need to show that if $T(\mathbf{\Theta})\boldsymbol{p} = T(\mathbf{\Theta})\bar{\boldsymbol{p}}$, then $\boldsymbol{p} = \bar{\boldsymbol{p}}$. This directly follows from the result in Lemma B.1 that when Γ is separable, the T-matrix $T(\mathbf{\Theta})$ has full column rank m.

We next prove the necessity part of Proposition III.1 that the separability of the Γ-matrix is necessary for identifiability of $\boldsymbol{p}$. Suppose Γ is inseparable and consider the representatives $\boldsymbol{\alpha}_{\mathcal{A}_1}, \ldots, \boldsymbol{\alpha}_{\mathcal{A}_C}$ from the C equivalence classes, respectively. It suffices to show that for any $\boldsymbol{p} \neq \bar{\boldsymbol{p}}$, if $\boldsymbol{\nu} = \bar{\boldsymbol{\nu}}$, where $\bar{\boldsymbol{\nu}} = (\bar{\nu}_{[\boldsymbol{\alpha}_{\mathcal{A}_i}]}, i = 1, \ldots, C)$ and $\bar{\nu}_{[\boldsymbol{\alpha}_{\mathcal{A}_i}]} = \sum_{\boldsymbol{\alpha}:\, \boldsymbol{\alpha} \in \mathcal{A}_i} \bar{p}_{\boldsymbol{\alpha}}$, then $T(\mathbf{\Theta})\boldsymbol{p} = T(\mathbf{\Theta})\bar{\boldsymbol{p}}$. Note that under the two-parameter restricted latent class models, any two equivalence latent classes $\boldsymbol{\alpha} \overset{\Gamma}{\sim} \boldsymbol{\alpha}'$ have identical item parameter vectors, i.e. $\boldsymbol{\theta}_{\bullet,\boldsymbol{\alpha}} = \boldsymbol{\theta}_{\bullet,\boldsymbol{\alpha}'}$. This further implies $T_{\bullet,\boldsymbol{\alpha}}(\mathbf{\Theta}) = T_{\bullet,\boldsymbol{\alpha}'}(\mathbf{\Theta})$ by the definition of the T-matrix. Let Γ^{eq} be the $J \times C$ submatrix of Γ that consists of the column vectors indexed by $\boldsymbol{\alpha}_{\mathcal{A}_i}$, $i = 1, \ldots, C$, and $T^{eq}(\mathbf{\Theta})$ be the corresponding $2^J \times C$ submatrix of $T(\mathbf{\Theta})$. Then if $\boldsymbol{\nu} = \bar{\boldsymbol{\nu}}$,

$$
\begin{aligned}
T(\mathbf{\Theta})\bar{\boldsymbol{p}} &= \sum_{i=1}^{C} \sum_{\boldsymbol{\alpha} \in \mathcal{A}_i} T_{\bullet,\boldsymbol{\alpha}}(\mathbf{\Theta})\bar{p}_{\boldsymbol{\alpha}} = \sum_{i=1}^{C} T_{\bullet,\boldsymbol{\alpha}_{\mathcal{A}_i}}(\mathbf{\Theta})\bar{\nu}_{\mathcal{A}_i} \\
&= T^{eq}(\mathbf{\Theta})\bar{\boldsymbol{\nu}} = T^{eq}(\mathbf{\Theta})\boldsymbol{\nu} \\
&= \sum_{i=1}^{C} T_{\bullet,\boldsymbol{\alpha}_{\mathcal{A}_i}}(\mathbf{\Theta})\nu_{\mathcal{A}_i} = \sum_{i=1}^{C} \sum_{\boldsymbol{\alpha} \in \mathcal{A}_i} T_{\bullet,\boldsymbol{\alpha}}(\mathbf{\Theta})p_{\boldsymbol{\alpha}} = T(\mathbf{\Theta})\boldsymbol{p},
\end{aligned}
$$

This proves that given an inseparable Γ-matrix, $\boldsymbol{p}$ is not identifiable.

Lastly, we prove Proposition III.2 that when the Γ-matrix is inseparable, the grouped proportion parameters $\boldsymbol{\nu}$ is identifiable. By Proposition B.1, we only need to show that if $T(\boldsymbol{\Theta})\boldsymbol{p} = T(\boldsymbol{\Theta})\bar{\boldsymbol{p}}$, then $\boldsymbol{\nu} = \bar{\boldsymbol{\nu}}$. From the calculation in the previous paragraph, we know $T(\boldsymbol{\Theta})\boldsymbol{p} = T^{eq}(\boldsymbol{\Theta})\boldsymbol{\nu}$ and $T(\boldsymbol{\Theta})\bar{\boldsymbol{p}} = T^{eq}(\boldsymbol{\Theta})\bar{\boldsymbol{\nu}}$. Since Γ^{eq} is separable by its construction, Lemma B.1 gives that $T^{eq}(\boldsymbol{\Theta})$ has full column rank C. Therefore, $T(\boldsymbol{\Theta})\boldsymbol{p} = T(\boldsymbol{\Theta})\bar{\boldsymbol{p}}$ implies $\boldsymbol{\nu} = \bar{\boldsymbol{\nu}}$ and $\boldsymbol{\nu}$ is identifiable. This completes the proof. $\square$

Proof of Equation (3.9) *in Remark III.1.* We introduce a notation first. For any set of items $S \subset \{1, \ldots, J\}$, we denote $\boldsymbol{q}_S = \vee_{h\in S}\, \boldsymbol{q}_h$. To prove the first part of (3.9), it suffices to show that under the conjunctive DINA model, (i) for any $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{R}^{Q,conj}$ and $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$, we have $\Gamma_{\bullet,\boldsymbol{\alpha}_1} \neq \Gamma_{\bullet,\boldsymbol{\alpha}_2}$; and (ii) for any $\boldsymbol{\alpha} \in \{0,1\}^K$, there exists $\boldsymbol{\alpha}' \in \mathcal{R}^{Q,conj}$ such that $\Gamma_{\bullet,\boldsymbol{\alpha}} = \Gamma_{\bullet,\boldsymbol{\alpha}'}$.

For any $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{R}^{Q,conj}$, without loss of generality, we can denote $\boldsymbol{\alpha}_1 = \vee_{h\in S_1}\boldsymbol{q}_h$ and $\boldsymbol{\alpha}_2 = \vee_{h\in S_2}\boldsymbol{q}_h$ where $S_1, S_2 \subset \{1, \ldots, J\}$ are two different sets of items. Then by definition, in the vector $\Gamma_{\bullet,\boldsymbol{\alpha}_1}$, the entry $\Gamma_{j,\boldsymbol{\alpha}_1} = 1$ if and only if $j \in S_1$; and similarly in the vector $\Gamma_{\bullet,\boldsymbol{\alpha}_2}$, the entry $\Gamma_{j,\boldsymbol{\alpha}_2} = 1$ if and only if $j \in S_2$. Since $S_1 \neq S_2$, we must have the two vectors different, i.e., $\Gamma_{\bullet,\boldsymbol{\alpha}_1} \neq \Gamma_{\bullet,\boldsymbol{\alpha}_2}$. This proves (i). Next, for any $\boldsymbol{\alpha} \in \{0,1\}^K$, we collect the items that $\boldsymbol{\alpha}$ is capable of in the set $S_{\boldsymbol{\alpha}} = \{j \in \mathcal{S} : \boldsymbol{\alpha} \succeq \boldsymbol{q}_j\}$, and just define $\boldsymbol{\alpha}' = \boldsymbol{q}_{S_{\boldsymbol{\alpha}}}$. then clearly $\boldsymbol{\alpha}' \in \mathcal{R}^{Q,conj}$. Additionally, the set of items that $\boldsymbol{\alpha}'$ is capable of is also $S_{\boldsymbol{\alpha}}$, so $\Gamma_{\bullet,\boldsymbol{\alpha}} = \Gamma_{\bullet,\boldsymbol{\alpha}'}$. This proves (ii). So the first part of (3.9) holds.

To prove the second part of (9), it suffices to show that under the disjunctive DINO model, (iii) for any $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{R}^{Q,comp}$ and $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$, we have $\Gamma^{disj}_{\bullet,\boldsymbol{\alpha}_1} \neq \Gamma^{disj}_{\bullet,\boldsymbol{\alpha}_2}$; and (iv) for any $\boldsymbol{\alpha} \in \{0,1\}^K$, there exists $\boldsymbol{\alpha}' \in \mathcal{R}^{Q,disj}$ such that $\Gamma^{disj}_{\bullet,\boldsymbol{\alpha}} = \Gamma^{disj}_{\bullet,\boldsymbol{\alpha}'}$. First, for $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{R}^{Q,disj}$ and $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$, they can be written as $\boldsymbol{\alpha}_1 = \mathbf{1}_K^\top - \boldsymbol{q}_{S_1}$ and $\boldsymbol{\alpha}_2 = \mathbf{1}_K^\top - \boldsymbol{q}_{S_2}$ where S_1, S_2 are two different item sets. Then

$$\begin{aligned}
\Gamma^{disj}_{j,\boldsymbol{\alpha}_1} &= I(\boldsymbol{\alpha}_1 \not\succeq \boldsymbol{q}_j) = I(\mathbf{1}_K^\top - \boldsymbol{q}_{S_1} \not\succeq \boldsymbol{q}_j) \\
&= I(\exists k \ s.t.\ q_{j,k} = 1,\ q_{S_1,k} = 0) = I(j \notin S_1),
\end{aligned} \tag{B.3}$$

and similarly $\Gamma^{disj}_{j,\boldsymbol{\alpha}_2} = I(j \notin S_2)$. Since $S_1 \neq S_2$, we have the inequalities of the two column vectors $\Gamma^{disj}_{\bullet,\boldsymbol{\alpha}_1} \neq \Gamma^{disj}_{\bullet,\boldsymbol{\alpha}_2}$. This proves (iii). Next, for any $\boldsymbol{\alpha} \in \{0,1\}^K$, we define $S^{\star}_{\boldsymbol{\alpha}} = \{j \in \mathcal{S} : \boldsymbol{\alpha} \prec \boldsymbol{q}_j\}$, which is the set of items $\boldsymbol{\alpha}$ is *not* capable of under the disjunctive model. Define

$$\boldsymbol{\alpha}' = \mathbf{1}_K^{\top} - \boldsymbol{q}_{S^{\star}_{\boldsymbol{\alpha}}},$$

then clearly $\boldsymbol{\alpha}' \in \mathcal{R}^{Q,disj}$. Further, similar to the derivation in (B.3), we have

$$\Gamma^{disj}_{j,\boldsymbol{\alpha}'} = I(j \notin S^{\star}_{\boldsymbol{\alpha}}),$$

which implies the set of items $\boldsymbol{\alpha}'$ is *not* capable of is also $S^{\star}_{\boldsymbol{\alpha}}$. This means $\Gamma^{disj}_{\bullet,\boldsymbol{\alpha}} = \Gamma^{disj}_{\bullet,\boldsymbol{\alpha}'}$ and proves (iv). $\square$

In the following proofs of the results for two-parameter restricted latent class models, for any latent class $\boldsymbol{\alpha}$, we use $[\boldsymbol{\alpha}]$ to denote the Γ-induced equivalence class containing $\boldsymbol{\alpha}$. Then by definition, with j ranging in the set of all items and $[\boldsymbol{\alpha}]$ ranging in the set of all equivalence classes, $(\theta_{j,[\boldsymbol{\alpha}]})$ give all the item parameters of interest while $(\nu_{[\boldsymbol{\alpha}]})$ give all the grouped proportion parameters of interest, under the framework of $\boldsymbol{p}$-partial identifiability. In the following, when there is no ambiguity, we write the item parameters as $\boldsymbol{\Theta} = (\theta_{j,[\boldsymbol{\alpha}]})$; and write $T^{eq}(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{\nu}} = T^{eq}(\boldsymbol{\Theta})\boldsymbol{\nu}$ as $T(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{\nu}} = T(\boldsymbol{\Theta})\boldsymbol{\nu}$, for $\boldsymbol{\Theta} = (\theta_{j,[\boldsymbol{\alpha}]})$, $\boldsymbol{\nu} = (\nu_{[\boldsymbol{\alpha}]})$, and $\bar{\boldsymbol{\Theta}} = (\bar{\theta}_{j,[\boldsymbol{\alpha}]})$, $\bar{\boldsymbol{\nu}} = (\bar{\nu}_{[\boldsymbol{\alpha}]})$.

Proof of Theorem III.1. To show the $\boldsymbol{p}$-partial identifiability, Proposition B.1 implies that we only need to show for any $(\boldsymbol{\Theta},\ \boldsymbol{\nu})$ and $(\bar{\boldsymbol{\Theta}},\ \bar{\boldsymbol{\nu}})$, $T(\boldsymbol{\Theta})\boldsymbol{\nu} = T(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{\nu}}$ implies $(\boldsymbol{\Theta},\boldsymbol{\nu}) = (\bar{\boldsymbol{\Theta}},\bar{\boldsymbol{\nu}})$. We prove this in two steps: in Step 1, we show the Repeated Measurement Condition (C1) ensures identifiability of $(\boldsymbol{\theta}^{+},\boldsymbol{\theta}^{-}_{non})$; in Step 2, we show the Sequentially Differentiable Condition (C2) additionally ensures identifiability of the remaining parameters $(\boldsymbol{\theta}^{-}_{basis},\boldsymbol{\nu})$, where $\boldsymbol{\theta}^{-}_{basis} = (\theta^{-}_{j},\ j \in S_{basis})$. In both steps, we frequently use the following lemma, whose proof is postponed to Section B.4.

Lemma B.2. *Under the two-parameter restricted latent class models, Equation* (C.1) *implies that* $\theta_j^+ \neq \bar{\theta}_j^-$ *and* $\theta_j^- \neq \bar{\theta}_j^+$ *for any item* j.

Step 1. To show identifiability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}_{non}^-)$ under (C1), we start with two identifiability results in the following two cases (a) and (b).

(a) If for item j, there exist two disjoint sets of items S_1 and S_2, both not containing j, such that

$$\mathcal{C}_j \supseteq \mathcal{C}_{S_1}, \quad \mathcal{C}_j \supseteq \mathcal{C}_{S_2}, \tag{B.4}$$

then we have the identifiability of θ_j^+, as proved in the following.

Define

$$\boldsymbol{\theta}^* = \sum_{h \in S_1} \theta_h^- \boldsymbol{e}_h + \sum_{m \in S_2} \bar{\theta}_m^- \boldsymbol{e}_m;$$

then consider the two row vectors corresponding to response pattern $\boldsymbol{r}^* = \sum_{h \in S_1 \cup S_2} \boldsymbol{e}_h$ in the transformed T-matrices $T(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)$ and $T(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)$, respectively, and we have the following expressions:

$$\begin{aligned}
&T_{\boldsymbol{r}^*,[\boldsymbol{\alpha}]}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top) = T_{\boldsymbol{r}^*,[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*) \\
&= \begin{cases} \prod_{h \in S_1} (\theta_h^+ - \theta_h^-) \prod_{m \in S_2} (\theta_m^+ - \bar{\theta}_m^-), & \text{for } [\boldsymbol{\alpha}] \in \mathcal{C}_{S_1} \cap \mathcal{C}_{S_2}; \\ 0, & \text{otherwise.} \end{cases} \\
&T_{\boldsymbol{r}^*,[\boldsymbol{\alpha}]}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top) = T_{\boldsymbol{r}^*,[\boldsymbol{\alpha}]}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^*) \\
&= \begin{cases} \prod_{h \in S_1} (\bar{\theta}_h^+ - \theta_h^-) \prod_{m \in S_2} (\bar{\theta}_m^+ - \bar{\theta}_m^-), & \text{for } [\boldsymbol{\alpha}] \in \mathcal{C}_{S_1} \cap \mathcal{C}_{S_2}; \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{B.5}$$

Since $\theta_h^+ \neq \bar{\theta}_h^-$ and $\bar{\theta}_h^+ \neq \theta_h^-$ for all h by Lemma B.2, we have

$$
\begin{aligned}
& T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}, \boldsymbol{\theta}^- - \boldsymbol{\theta})\boldsymbol{\nu} \\
= & \Big(\sum_{[\boldsymbol{\alpha}]\in\mathcal{C}_{S_1}\cap\mathcal{C}_{S_2}} \nu_{[\boldsymbol{\alpha}]}\Big) \prod_{h\in S_1}(\bar{\theta}_h^+ - \theta_h^-) \prod_{m\in S_2}(\bar{\theta}_m^+ - \bar{\theta}_m^-) \neq 0,
\end{aligned}
$$

and similarly $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta})\bar{\boldsymbol{\nu}} \neq 0$. Therefore, $\mathcal{C}_{S_1} \subseteq \mathcal{C}_j$, $\mathcal{C}_{S_2} \subseteq \mathcal{C}_j$ together with Equation (C.1) indicates

$$
\begin{aligned}
\theta_j^+ &= \frac{T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}, \boldsymbol{\theta}^- - \boldsymbol{\theta})\boldsymbol{\nu}}{T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}, \boldsymbol{\theta}^- - \boldsymbol{\theta})\boldsymbol{\nu}} \\
&= \frac{T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta})\bar{\boldsymbol{\nu}}}{T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta})\bar{\boldsymbol{\nu}}} = \bar{\theta}_j^+.
\end{aligned} \tag{B.6}
$$

(b) If for item j, there exist another item h and an item set S_2 not containing j or h, such that

$$
\mathcal{C}_h \supseteq \mathcal{C}_j \supseteq \mathcal{C}_{S_2}, \tag{B.7}
$$

then we have the identifiability of θ_j^-, as proved in the following.

From the proof of (a) we can obtain the identifiability of θ_h^+, i.e., $\theta_h^+ = \bar{\theta}_h^+$. Define $\boldsymbol{\theta}^* = \theta_h^+\boldsymbol{e}_h$. Consider the two row vectors corresponding to response pattern $\boldsymbol{r}^* = \boldsymbol{e}_h$ in the transformed T-matrices $T(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ and $T(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top)$, respectively, and we have

$$
\begin{aligned}
T_{\boldsymbol{e}_h,[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*) &= \begin{cases} \theta_h^- - \theta_h^+, & \text{for } [\boldsymbol{\alpha}] \in \mathcal{C}_h^c; \\ 0, & \text{otherwise.} \end{cases} \\
T_{\boldsymbol{e}_h,[\boldsymbol{\alpha}]}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^*) &= \begin{cases} \bar{\theta}_h^- - \theta_h^+, & \text{for } [\boldsymbol{\alpha}] \in \mathcal{C}_h^c; \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{B.8}
$$

Moreover, we have

$$T_{\boldsymbol{e}_h,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*)\boldsymbol{\nu} = T_{\boldsymbol{e}_h,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^*)\bar{\boldsymbol{\nu}} \neq 0.$$

Since $\mathcal{C}_h^c \subseteq \mathcal{C}_j^c$, Equation (C.1) indicates

$$\begin{aligned} \theta_j^- &= \frac{T_{\boldsymbol{e}_h+\boldsymbol{e}_j,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}, \boldsymbol{\theta}^- - \boldsymbol{\theta})\boldsymbol{\nu}}{T_{\boldsymbol{e}_h,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}, \boldsymbol{\theta}^- - \boldsymbol{\theta})\boldsymbol{\nu}} \\ &= \frac{T_{\boldsymbol{e}_h+\boldsymbol{e}_j,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta})\bar{\boldsymbol{\nu}}}{T_{\boldsymbol{e}_h,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta})\bar{\boldsymbol{\nu}}} = \bar{\theta}_j^-. \end{aligned} \tag{B.9}$$

With the above results in cases (a) and (b), we show that (C1) ensures the identifiability of $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-_{non})$. Specifically, if condition (C1) is satisfied, then for each item j, there exist two item sets S_1 and S_2 satisfying (B.4). Thus the result for case (a) implies that the items parameters $\boldsymbol{\theta}^+$ are identifiable. Moreover, for any non-basis item j, by definition there must exist an item h such that $\mathcal{C}_h \supseteq \mathcal{C}_j$; condition (C1) further guarantees that there exists another set S_2 not containing j such that $\{h\} \cap S_2 = \varnothing$ and $\mathcal{C}_j \supseteq \mathcal{C}_{S_2}$. Therefore, (B.7) is satisfied and the result for case (b) implies that θ_j^- is identifiable for all $j \in S_{non}$.

Step 2. This step proves that when (C2) additionally holds, the parameter θ_j^- of each basis item j is identifiable. Following the definition of the sequentially expanding procedure in (C2), we first prove that in each expanding step, $\theta_j^- = \bar{\theta}_j^-$ for all $j \in S_{sep}$, namely, every item j included into the separator set through the expanding procedure has its lower level parameter θ_j^- identifiable. To show this, it suffices to prove the result that if an item j is set S-differentiable and $\theta_h^- = \bar{\theta}_h^-$, $\theta_h^+ = \bar{\theta}_h^+$ for any $h \in S$, then $\theta_j^- = \bar{\theta}_j^-$.

If j is S-differentiable, by definition there exist two item sets S_j^+, $S_j^- \subseteq S$ that are not necessarily disjoint such that $\mathcal{C}_{S_j^-} \backslash \mathcal{C}_{S_j^+} \subseteq \mathcal{C}_j^c$. Define

$$\boldsymbol{\theta}^* = \sum_{h \in S_j^+ \cup S_j^-} \theta_h^- \boldsymbol{e}_h,$$

and define response patterns

$$\boldsymbol{r}^+ = \sum_{h \in S_j^+} \boldsymbol{e}_h, \quad \boldsymbol{r}^- = \sum_{h \in S_j^-} \boldsymbol{e}_h.$$

Note that the nonzero entries of the row vectors $T_{\boldsymbol{r}^+, \boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*)$ and $T_{\boldsymbol{r}^-, \boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*)$ correspond to the capable classes of S_j^+ and S_j^-, respectively. Specifically,

$$T_{\boldsymbol{r}^+, [\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*) = \begin{cases} \prod_{j \in S_j^+} (\theta_j^+ - \theta_j^-), & [\boldsymbol{\alpha}] \in \mathcal{C}_{S_j^+}; \\ 0, & [\boldsymbol{\alpha}] \notin \mathcal{C}_{S_j^+}. \end{cases}$$

$$T_{\boldsymbol{r}^-, [\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*) = \begin{cases} \prod_{j \in S_j^-} (\theta_j^+ - \theta_j^-), & [\boldsymbol{\alpha}] \in \mathcal{C}_{S_j^-}; \\ 0, & [\boldsymbol{\alpha}] \notin \mathcal{C}_{S_j^-}. \end{cases}$$

We define a linear transformation of the above vectors $T_{\boldsymbol{r}^+, \boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*)$ and $T_{\boldsymbol{r}^-, \boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*)$ as

$$T_{(\boldsymbol{r}^- + k \cdot \boldsymbol{r}^+), \boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*) := T_{\boldsymbol{r}^-, \boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*) + k \cdot T_{\boldsymbol{r}^+, \boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{\theta}^- - \boldsymbol{\theta}^*)$$

where

$$k = -\frac{\prod_{j \in S_j^-} (\theta_j^+ - \theta_j^-)}{\prod_{j \in S_j^+} (\theta_j^+ - \theta_j^-)} \neq 0.$$

Since the capable classes of S_j^+ must also be capable classes of S_j^-, we have

$$T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{\theta}^--\boldsymbol{\theta}^*)=\begin{cases}\prod_{j\in S_j^-}(\theta_j^+-\theta_j^-), & [\boldsymbol{\alpha}]\in\mathcal{C}_{S_j^-}\setminus\mathcal{C}_{S_j^+};\\ 0, & \text{otherwise.}\end{cases}$$

Under the assumption that $\theta_h^-=\bar{\theta}_h^-$, $\theta_h^+=\bar{\theta}_h^+$ for any $h\in S$, we also have

$$T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),[\boldsymbol{\alpha}]}(\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{\theta}}^--\boldsymbol{\theta}^*)=\begin{cases}\prod_{j\in S_j^-}(\theta_j^+-\theta_j^-), & [\boldsymbol{\alpha}]\in\mathcal{C}_{S_j^-}\setminus\mathcal{C}_{S_j^+};\\ 0, & \text{otherwise.}\end{cases}$$

Note that the condition $\mathcal{C}_{S_j^-}\setminus\mathcal{C}_{S_j^+}\subseteq\mathcal{C}_j^c$ implies for any $[\boldsymbol{\alpha}]\in\mathcal{C}_{S_j^-}\setminus\mathcal{C}_{S_j^+}$, one must have $[\boldsymbol{\alpha}]\in\mathcal{C}_j^c$. Since

$$T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),\boldsymbol{\cdot}}(\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{\theta}^--\boldsymbol{\theta}^*)\boldsymbol{\nu}=T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{\theta}}^--\boldsymbol{\theta}^*)\bar{\boldsymbol{\nu}}\neq 0,$$

Since $j\notin(S_j^-\cup S_j^+)$, Equation (C.1) implies

$$\begin{aligned}\theta_j^- &= \frac{\{T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}(\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{\theta}^--\boldsymbol{\theta}^*)\odot T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),\boldsymbol{\cdot}}(\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{\theta}^--\boldsymbol{\theta}^*)\}\boldsymbol{\nu}}{T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),\boldsymbol{\cdot}}(\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{\theta}^--\boldsymbol{\theta}^*)\boldsymbol{\nu}}\\ &= \frac{\{T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{\theta}}^--\boldsymbol{\theta}^*)\odot T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{\theta}}^--\boldsymbol{\theta}^*)\}\bar{\boldsymbol{\nu}}}{T_{(\boldsymbol{r}^-+k\cdot\boldsymbol{r}^+),\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{\theta}}^--\boldsymbol{\theta}^*)\bar{\boldsymbol{\nu}}}\\ &= \bar{\theta}_j^-,\end{aligned}$$

where $\odot$ denotes the element-wise product of two vectors. This proves the claim that if item j is set S-differentiable and $\theta_h^-=\bar{\theta}_h^-$, $\theta_h^+=\bar{\theta}_h^+$ for any $h\in S$, then $\theta_j^-=\bar{\theta}_j^-$. Together with the result in Step 1 and the definition of the sequentially expanding procedure in (C2), we therefore have the identifiability of $(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)$.

With $(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)=(\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{\theta}}^-)$, Equation (C.1) simplifies to $T(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)\boldsymbol{p}=T(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)\bar{\boldsymbol{p}}=0$. The last part of the proof of Propositions 1 and 4 then gives the identifiability of $\boldsymbol{\nu}$. This completes the proof of the theorem. □

Proof of Proposition III.3. For ease of discussion, in this proof we use $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^- \mid \Gamma(S))$ to denote the T-matrix associated with any S-adjusted design matrix $\Gamma(S)$ and item parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$.

For any S-adjusted $\Gamma(S)$-matrix, we define another set of item parameters $\tilde{\boldsymbol{\theta}}^+ = (\tilde{\theta}_1^+, \ldots, \tilde{\theta}_J^+)$ and $\tilde{\boldsymbol{\theta}}^- = (\tilde{\theta}_1^-, \ldots, \tilde{\theta}_J^-)$, where $\tilde{\theta}_j^- = \theta_j^+$, $\tilde{\theta}_j^+ = \theta_j^-$ for all $j \in S$, and $\tilde{\theta}_j^- = \theta_j^-$, $\tilde{\theta}_j^+ = \theta_j^+$ for all $j \notin S$. We first show that the T-matrix $T(\tilde{\boldsymbol{\theta}}^+, \tilde{\boldsymbol{\theta}}^- \mid \Gamma(S))$ can be viewed as the T-matrix associated with the original Γ-matrix with item parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$, i.e.,

$$T_{\boldsymbol{r},\boldsymbol{\alpha}}(\tilde{\boldsymbol{\theta}}^+, \tilde{\boldsymbol{\theta}}^- \mid \Gamma(S)) = T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^- \mid \Gamma). \tag{B.10}$$

To show this, note that for any response pattern $\boldsymbol{r} \in \{0,1\}^J$, we have

$$T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^- \mid \Gamma) = \prod_{j:r_j=1} \Big[\Gamma_{j,\boldsymbol{\alpha}}\theta_j^+ + (1-\Gamma_{j,\boldsymbol{\alpha}})\theta_j^-\Big]$$

and

$$\begin{aligned}
&T_{\boldsymbol{r},\boldsymbol{\alpha}}(\tilde{\boldsymbol{\theta}}^+, \tilde{\boldsymbol{\theta}}^- \mid \Gamma(S)) \\
&= \prod_{j:r_j=1} \Big[\{\Gamma(S)\}_{j,\boldsymbol{\alpha}}\tilde{\theta}_j^+ + (1-\{\Gamma(S)\}_{j,\boldsymbol{\alpha}})\tilde{\theta}_j^-\Big] \\
&= \prod_{j\in S:r_j=1} \Big[(1-\Gamma_{j,\boldsymbol{\alpha}})\tilde{\theta}_j^+ + \Gamma_{j,\boldsymbol{\alpha}}\tilde{\theta}_j^-\Big] \times \prod_{j\notin S:r_j=1} \Big[\Gamma_{j,\boldsymbol{\alpha}}\tilde{\theta}_j^+ + (1-\Gamma_{j,\boldsymbol{\alpha}})\tilde{\theta}_j^-\Big] \\
&= \prod_{j\in S:r_j=1} \Big[(1-\Gamma_{j,\boldsymbol{\alpha}})\theta_j^- + \Gamma_{j,\boldsymbol{\alpha}}\theta_j^+\Big] \times \prod_{j\notin S:r_j=1} \Big[\Gamma_{j,\boldsymbol{\alpha}}\theta_j^+ + (1-\Gamma_{j,\boldsymbol{\alpha}})\theta_j^-\Big] \\
&= \prod_{j:r_j=1} \Big[\Gamma_{j,\boldsymbol{\alpha}}\theta_j^+ + (1-\Gamma_{j,\boldsymbol{\alpha}})\theta_j^-\Big] \\
&= T_{\boldsymbol{r},\boldsymbol{\alpha}}(\boldsymbol{\theta}^+, \boldsymbol{\theta}^- \mid \Gamma).
\end{aligned}$$

With the result in (B.10), to prove Proposition III.3, it suffices to show that the identifiability argument in Theorem III.1 still holds if the Γ-induced restrictions of the item parameters, $\theta_j^+ > \theta_j^-$ for all $j = 1, \ldots, J$, are replaced by the constraints that $\theta_j^+ < \theta_j^-$ for any $j \in S$ and $\theta_j^+ > \theta_j^-$ for any $j \notin S$.

We next prove this claim. If $\theta_j^+ < \theta_j^-$ for some items j, the conclusion in Lemma B.2 still holds. In particular, following the proof of Lemma B.2, if $\theta_j^+ < \theta_j^-$, then

$$\theta_j^+ = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \theta_j^+ p_{\boldsymbol{\alpha}} \leq \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \theta_{j,\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{\theta}_{j,\boldsymbol{\alpha}} \bar{p}_{\boldsymbol{\alpha}} \leq \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{\theta}_j^- \bar{p}_{\boldsymbol{\alpha}} = \bar{\theta}_j^-,$$

where among the two "$\leq$" there is at least a strict "$<$". This implies $\theta_j^+ \neq \bar{\theta}_j^-$ for all $j = 1, \ldots, J$, and a similar argument gives $\theta_j^- \neq \bar{\theta}_j^+$ for all $j = 1, \ldots, J$. With these results, we can check that all the needed inequalities in the proof of Theorem III.1 still hold and all the proof steps proceed with no changes. This proves the conclusion of the proposition. $\square$

Next we prove the identifiability results for the two-parameter Q-restricted models. We say a Q-matrix of size $J \times K$ is *complete* for the two-parameter model, if after some row permutation it contains an identity submatrix $\mathcal{I}_K$. Under the conjunctive model assumption, let

$$\mathcal{R}^Q = \mathcal{R}^{Q,conj} = \{\mathbf{0}_K^\top\} \cup \{\boldsymbol{\alpha} = \vee_{h\in S}\, \boldsymbol{q}_h : \forall S \subset \mathcal{S}\} \tag{B.11}$$

be defined as in Remark 1 of the main text. Since elements of $\mathcal{R}^Q$ are K-dimensional binary vectors, they can be viewed as attribute profiles and $\mathcal{R}^Q \subseteq \{0,1\}^K$. When Q is complete, clearly $\mathcal{R}^Q = \{0,1\}^K$. The row-union space $\mathcal{R}^Q$ has the following two properties. First, every two attribute profiles in $\mathcal{R}^Q$ have different ideal response vectors, i.e.

$$\forall \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{R}^Q,\ \boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2, \quad \Gamma_{\bullet,\boldsymbol{\alpha}_1} = \Gamma_{\bullet,\boldsymbol{\alpha}_2}. \tag{B.12}$$

Second, when Q is incomplete, for any attribute profile $\boldsymbol{\alpha} \in \{0,1\}^K$, there must exist some $\boldsymbol{\alpha}' \in \mathcal{R}^Q$ that has the same ideal response vector as $\boldsymbol{\alpha}$, i.e.

$$\forall \boldsymbol{\alpha} \in \{0,1\}^K, \quad \exists \boldsymbol{\alpha}' \in \mathcal{R}^Q \text{ such that } \boldsymbol{\alpha} \succeq \boldsymbol{\alpha}' \text{ and } \Gamma_{\bullet,\boldsymbol{\alpha}} = \Gamma_{\bullet,\boldsymbol{\alpha}'}. \tag{B.13}$$

Based on the above two properties, when $\mathcal{A}$ is saturated, $\mathcal{R}^Q$ is a complete set of representatives of the conjunctive equivalence classes. Similarly, we can show $\mathcal{R}^{Q,comp} = \{\mathbf{1}_K^\top - \boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathcal{R}^Q\}$ gives a complete set of representatives of the compensatory equivalence classes. Therefore, this proves the claims in Remark 1 of the main text. In the following proofs of Corollary III.1, Theorem III.2, Theorem III.3 and Theorem III.4 for the two-parameter Q-restricted models, when there is no ambiguity, we will exchangeably say an equivalence class $[\boldsymbol{\alpha}]$ is induced by the Γ-matrix or is induced by the corresponding Q-matrix.

Proof of Corollary III.1. With definitions of non-basis and basis items introduced in (3.12) and definition of S-differentiable item introduced in (3.13), conditions (C1) and (C2) exactly reduce to the new conditions (C1*) and (C2*) regarding the Q-matrix for the two-parameter conjunctive model, therefore by Theorem III.1, (C1*) and (C2*) are sufficient for the $\boldsymbol{p}$-partial identifiability of the conjunctive models.

On the other hand, for the two-parameter compensatory model, if the Q-matrix satisfies the new conditions (C1*) and (C2*), then we have that Γ^{conj} satisfies the original conditions (C1) and (C2). Given an arbitrary Q-matrix, by the definition of the conjunctive Γ^{conj} and compensatory Γ^{comp}, for any item j and any attribute profile $\boldsymbol{\alpha} \in \{0,1\}^K$, we can obtain

$$\Gamma^{comp}_{j,\boldsymbol{\alpha}} = 1 - \Gamma^{conj}_{j,\mathbf{1}-\boldsymbol{\alpha}} = I(\alpha_k = 1 \text{ for some } k \text{ s.t. } q_{j,k} = 1), \tag{B.14}$$

where $\mathbf{1}-\boldsymbol{\alpha} = (1-\alpha_1, \ldots, 1-\alpha_K)$. This means the two matrices Γ^{conj} and $\mathbf{1}_{J\times C}-\Gamma^{comp}$ only differ by a column permutation. Noting that conditions (C1) and (C2) do not depend on the order of the column vectors, so if Γ^{conj} satisfies (C1) and (C2), then $\mathbf{1}_{J\times C} - \Gamma^{comp}$ also satisfies (C1) and (C2). Then Proposition III.3 implies the two parameter compensatory model with design matrix Γ^{comp} is $\boldsymbol{p}$-partially identifiable. $\square$

Proof of Theorem III.2. Without loss of generality, we focus on the proof of the conclusion for the two-parameter conjunctive model, and all the arguments also hold for the two-parameter disjunctive model, following the similar argument in the proof of Proposition III.3. In the following, we first present the proof of part (a), then that of part (b.2), and finally that of part (b.1).

Proof of part (a). Without loss of generality, assume the Q-matrix takes the following form

$$Q = \begin{pmatrix} 1 & \boldsymbol{v}_1 \\ \mathbf{0} & Q' \end{pmatrix},$$

where Q' is a submatrix of size $(J-1) \times (K-1)$ and $\boldsymbol{v}_1$ is a $(K-1)$-dimensional vector. For any attribute profile $\boldsymbol{\alpha} = (0, \boldsymbol{\alpha}_{2:K})$, denote $\boldsymbol{\alpha} + \boldsymbol{e}_1 = (1, \boldsymbol{\alpha}_{2:K})$; and for any $\boldsymbol{\alpha} = (1, \boldsymbol{\alpha}_{2:K})$, denote $\boldsymbol{\alpha} - \boldsymbol{e}_1 = (0, \boldsymbol{\alpha}_{2:K})$. Consider any valid set of parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$. To prove the conclusion in (A), we next construct another set of parameters $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{\nu}}) \neq (\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$ but $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{\nu} = T(\bar{\boldsymbol{\theta}}^+, \boldsymbol{\theta}^-)\bar{\boldsymbol{\nu}}$. In particular, we set $\bar{\boldsymbol{\theta}}^- = \boldsymbol{\theta}^-$, $\bar{\theta}_j^+ = \theta_j^+$ for $j = 2, \ldots, J$, and choose $\bar{\theta}_1^+$ close enough but not equal to θ_1^+. Define

$$\mathcal{R}_0 = \{\boldsymbol{\alpha} \in \mathcal{R}^Q : \alpha_1 = 0, \boldsymbol{\alpha} \succeq (0, \boldsymbol{v}_1)\},$$
$$\mathcal{R}_1 = \{\boldsymbol{\alpha} \in \mathcal{R}^Q : \alpha_1 = 1, \boldsymbol{\alpha} \succeq (0, \boldsymbol{v}_1)\},$$

then we can see that the two sets $\mathcal{R}_0$ and $\mathcal{R}_1$ are disjoint and their elements are paired in the sense that for any $\boldsymbol{\alpha} \in \mathcal{R}_0$, one has $\boldsymbol{\alpha} + \boldsymbol{e}_1 \in \mathcal{R}_1$ and for any $\boldsymbol{\alpha} \in \mathcal{R}_1$, one has $\boldsymbol{\alpha} - \boldsymbol{e}_1 \in \mathcal{R}_0$. To construct the proportion parameters $\bar{\boldsymbol{\nu}}$, we set

$$\begin{cases} \bar{\nu}_{[\boldsymbol{\alpha}]} = \nu_{[\boldsymbol{\alpha}]} + \left(1 - \frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-}\right)\nu_{[\boldsymbol{\alpha}+\boldsymbol{e}_1]}, & \forall \boldsymbol{\alpha} \in \mathcal{R}_0; \\ \bar{\nu}_{[\boldsymbol{\alpha}]} = \frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-}\nu_{[\boldsymbol{\alpha}]}, & \forall \boldsymbol{\alpha} \in \mathcal{R}_1; \\ \bar{\nu}_{[\boldsymbol{\alpha}]} = \nu_{[\boldsymbol{\alpha}]}, & \forall \boldsymbol{\alpha} \in \mathcal{R}^Q \setminus (\mathcal{R}_0 \cup \mathcal{R}_1). \end{cases} \tag{B.15}$$

For notational simplicity, denote $\mathcal{R}_c = \mathcal{R}^Q \setminus (\mathcal{R}_0 \cup \mathcal{R}_1)$. Next we show that under the two different sets of parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$ and $(\bar{\boldsymbol{\theta}}^+, \boldsymbol{\theta}^-, \bar{\boldsymbol{\nu}})$, for any response pattern $\boldsymbol{r} \in \{0,1\}^J$,

$$T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\boldsymbol{\nu}} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\boldsymbol{\nu}}, \tag{B.16}$$

which will complete the proof. To this end, we consider two types of response patterns $\boldsymbol{r} = (r_1, \ldots, r_J)$ respectively in the following: (a) $r_1 = 0$; and (b) $r_1 = 1$.

(a) Firstly, for any $\boldsymbol{r} \in \{0,1\}^J$ such that $r_1 = 0$, $T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^-, \mathbf{0})$, so by our construction,

$$\begin{aligned}
&T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\boldsymbol{\nu}} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\boldsymbol{\nu}} \\
=& \sum_{\boldsymbol{\alpha} \in \mathcal{R}^Q} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\nu}_{[\boldsymbol{\alpha}]} \\
=& \sum_{\boldsymbol{\alpha} \in \mathcal{R}_0} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\nu}_{[\boldsymbol{\alpha}]} + \sum_{\boldsymbol{\alpha} \in \mathcal{R}_1} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\nu}_{[\boldsymbol{\alpha}]} \\
&+ \sum_{\boldsymbol{\alpha} \in \mathcal{R}_c} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\bar{\nu}_{[\boldsymbol{\alpha}]} \\
=& \sum_{\boldsymbol{\alpha} \in \mathcal{R}_0} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) \left(\nu_{[\boldsymbol{\alpha}]} + \left(1 - \frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-}\right) \nu_{[\boldsymbol{\alpha}+\boldsymbol{e}_1]} \right) \\
&+ \sum_{\boldsymbol{\alpha} \in \mathcal{R}_1} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) \left(\frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-} \nu_{[\boldsymbol{\alpha}]} \right) \\
&+ \sum_{\boldsymbol{\alpha} \in \mathcal{R}_c} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\nu_{[\boldsymbol{\alpha}]} \\
:=& I_0 + I_1 + I_c.
\end{aligned}$$

Note that the elements in $\mathcal{R}_0$ and $\mathcal{R}_1$ are paired, and moreover, for any pair of attribute profiles $(\boldsymbol{\alpha}, \boldsymbol{\alpha} + \boldsymbol{e}_1)$ where $\boldsymbol{\alpha} \in \mathcal{R}_0$ and $\boldsymbol{\alpha} + \boldsymbol{e}_1 \in \mathcal{R}_1$, we have

$$T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) = T_{\boldsymbol{r},[\boldsymbol{\alpha}+\boldsymbol{e}_1]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) = \prod_{j: r_j = 1} (\theta_{j,[\boldsymbol{\alpha}]} - \theta_j^-) \tag{B.17}$$

for any type-(a) response pattern $\boldsymbol{r}$, namely $\boldsymbol{r} \in \{0,1\}^J$ such that $r_1 = 0$.

Equation (B.17) leads to

$$
\begin{aligned}
I_1 &= \sum_{\boldsymbol{\alpha}\in\mathcal{R}_0} T_{\boldsymbol{r},[\boldsymbol{\alpha}+\boldsymbol{e}_1]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) \left(\frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-}\nu_{[\boldsymbol{\alpha}+\boldsymbol{e}_1]}\right) \\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{R}_0} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) \left(\frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-}\nu_{[\boldsymbol{\alpha}+\boldsymbol{e}_1]}\right).
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
&I_0 + I_1 \\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{R}_0} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) \Big(\nu_{[\boldsymbol{\alpha}]} + \left(1 - \frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-}\right)\nu_{[\boldsymbol{\alpha}+\boldsymbol{e}_1]} + \frac{\theta_1^+ - \theta_1^-}{\bar{\theta}_1^+ - \theta_1^-}\nu_{[\boldsymbol{\alpha}+\boldsymbol{e}_1]}\Big) \\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{R}_0} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}) \Big(\nu_{[\boldsymbol{\alpha}]} + \nu_{[\boldsymbol{\alpha}+\boldsymbol{e}_1]}\Big) \\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{R}_0} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\nu_{[\boldsymbol{\alpha}]} + \sum_{\boldsymbol{\alpha}\in\mathcal{R}_1} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\nu_{[\boldsymbol{\alpha}]},
\end{aligned}
$$

where the last equality also results from (B.17). This further results in

$$
I_0 + I_1 + I_c = \sum_{\boldsymbol{\alpha}\in\mathcal{R}^Q} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\nu_{[\boldsymbol{\alpha}]} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0})\boldsymbol{\nu}.
$$

This proves that for any $\boldsymbol{r}$ such that $r_1 = 0$, Equation (B.16) holds.

(b) Secondly, consider the type-(b) response pattern, namely those $\boldsymbol{r} = (1, r_2, \ldots, r_J)$. For such $\boldsymbol{r}$, denote $\boldsymbol{r} - \boldsymbol{e}_1 = (0, r_2, \ldots, r_J)$, then

$$
T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^-, \mathbf{0}) = \begin{cases} (\bar{\theta}_1^+ - \theta_1^-)\cdot T_{\boldsymbol{r}-\boldsymbol{e}_1,[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-, \mathbf{0}), & \boldsymbol{\alpha} \succeq (1, \boldsymbol{v}_1); \\ 0, & \boldsymbol{\alpha} \not\succeq (1, \boldsymbol{v}_1), \end{cases}
$$

which indicates $T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^-, \mathbf{0}) = 0$ for all $\boldsymbol{\alpha} \in \mathcal{R}_0 \cup \mathcal{R}_c$. This is because for $\boldsymbol{\alpha} \in \mathcal{R}_0$, $\alpha_1 = 0 \not\succeq 1$; and for $\boldsymbol{\alpha} \in \mathcal{R}_c$, $(\alpha_2, \ldots, \alpha_K) \not\succeq \boldsymbol{v}_1$ by our definitions.

Therefore,

$$
\begin{aligned}
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\theta}}^{+}-\boldsymbol{\theta}^{-},\mathbf{0})\bar{\boldsymbol{\nu}} &= \sum_{\boldsymbol{\alpha}\in\mathcal{R}^{Q}} T_{\boldsymbol{r},[\boldsymbol{\alpha}]}(\bar{\boldsymbol{\theta}}^{+}-\boldsymbol{\theta}^{-},\mathbf{0})\bar{\nu}_{[\boldsymbol{\alpha}]}\\
&= \sum_{\substack{\boldsymbol{\alpha}\in\mathcal{R}^{Q}\\ \boldsymbol{\alpha}\succeq(1,\boldsymbol{v}_1)}} T_{\boldsymbol{r}-\boldsymbol{e}_1,[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^{+}-\boldsymbol{\theta}^{-},\mathbf{0})(\bar{\theta}_1^{+}-\theta_1^{-})\bar{\nu}_{[\boldsymbol{\alpha}]}\\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{R}_1} T_{\boldsymbol{r}-\boldsymbol{e}_1,[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^{+}-\boldsymbol{\theta}^{-},\mathbf{0})(\bar{\theta}_1^{+}-\theta_1^{-})\bar{\nu}_{[\boldsymbol{\alpha}]}\\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{R}_1} T_{\boldsymbol{r}-\boldsymbol{e}_1,[\boldsymbol{\alpha}]}(\boldsymbol{\theta}^{+}-\boldsymbol{\theta}^{-},\mathbf{0})(\theta_1^{+}-\theta_1^{-})\nu_{[\boldsymbol{\alpha}]}\\
&= T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\theta}^{+}-\boldsymbol{\theta}^{-},\mathbf{0})\boldsymbol{\nu},
\end{aligned}
$$

where our previous construction $(\bar{\theta}_1^{+}-\theta_1^{-})\bar{\nu}_{[\boldsymbol{\alpha}]}=(\theta_1^{+}-\theta_1^{-})\nu_{[\boldsymbol{\alpha}]}$ for $\boldsymbol{\alpha}\in\mathcal{R}_1$ defined in (B.15) is used to obtain the last but second equality. This proves that for any $\boldsymbol{r}$ such that $r_1=1$, Equation (B.16) holds.

Now that we have proved Equation (B.16) holds for any $\boldsymbol{r}\in\{0,1\}^J$, we have found two different sets of parameters $(\boldsymbol{\theta}^{+},\boldsymbol{\nu})\neq(\bar{\boldsymbol{\theta}}^{+},\bar{\boldsymbol{\nu}})$ that give $T(\boldsymbol{\theta}^{+},\boldsymbol{\theta}^{-})\boldsymbol{\nu}=T(\bar{\boldsymbol{\theta}}^{+},\boldsymbol{\theta}^{-})\bar{\boldsymbol{\nu}}$. This shows the non-identifiability of the parameters $(\boldsymbol{\theta}^{+},\boldsymbol{\theta}^{-},\boldsymbol{\nu})$, and concludes the proof of part (A).

Proof of Part (b.2). Equation (C.1) is equivalent to

$$
T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\Theta})\boldsymbol{\nu}=T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{\nu}}\quad\text{for all}\quad \boldsymbol{r}=(r_1,\ldots,r_J)^{\top}\in\{0,1\}^J. \tag{B.18}
$$

The detailed form of (C.22) can be written as follows, for any $\boldsymbol{r}\in\{0,1\}^J$,

$$
\sum_{\boldsymbol{\alpha}\in\mathcal{R}^{Q}}\prod_{r_j=1}\theta_{j,[\boldsymbol{\alpha}]}\cdot\nu_{[\boldsymbol{\alpha}]}=\sum_{\boldsymbol{\alpha}\in\mathcal{R}^{Q}}\prod_{r_j=1}\bar{\theta}_{j,[\boldsymbol{\alpha}]}\cdot\bar{\nu}_{[\boldsymbol{\alpha}]}, \tag{B.19}
$$

where $\mathcal{R}^{Q}$ denotes the row-union space of the Q-matrix Q as in (B.11). For any attribute profile $\boldsymbol{\alpha}\in\{0,1\}^K$, $[\boldsymbol{\alpha}]$ denotes the equivalence class containing $\boldsymbol{\alpha}$ that is induced by Q. Let $\boldsymbol{\alpha}_{2:K}$ denote the vector containing last $K-1$ elements of $\boldsymbol{\alpha}$, so $\boldsymbol{\alpha}$ can be written as $\boldsymbol{\alpha} = (\alpha_1, \boldsymbol{\alpha}_{2:K})$ and $[\alpha_1, \boldsymbol{\alpha}_{2:K}]$ represents the equivalence class $\boldsymbol{\alpha}$ belongs to. Recall that we use $\boldsymbol{R} = (R_1, \ldots, R_J)$ to denote a random response vector ranging in $\{0,1\}^J$, and use $\mathbf{A} = (A_1, \ldots, A_K)$ to denote a random attribute profile ranging in the latent class space $\mathcal{A} \subseteq \{0,1\}^K$. Denote $\mathbf{A}_{2:K} := (A_2, \ldots, A_K)$.

Under the assumptions of part (B), the Q-matrix takes the following form

$$Q = \begin{pmatrix} 1 & \boldsymbol{v}_1^\top \\ 1 & \boldsymbol{v}_2^\top \\ \mathbf{0} & Q' \end{pmatrix}.$$

For any two different equivalence classes $[0, \boldsymbol{\alpha}_{2:K}]$ and $[1, \boldsymbol{\alpha}_{2:K}]$ where $\boldsymbol{\alpha}_{2:K} \in \{0,1\}^{K-1}$, their corresponding item parameters to any item $j > 2$ are the same, i.e., for any $j > 2$ and any $\boldsymbol{\alpha}_{2:K} \in \{0,1\}^{K-1}$,

$$\begin{aligned} \mathbb{P}(R_j = 1 \mid \mathbf{A} = (1, \boldsymbol{\alpha}_{2:K})) &= \mathbb{P}(R_j = 1 \mid \mathbf{A} = (0, \boldsymbol{\alpha}_{2:K})) \\ &= \theta_{j,[0,\boldsymbol{\alpha}_{2:K}]}. \end{aligned} \tag{B.20}$$

Therefore for any response pattern in the form $\boldsymbol{r} = (0, 0, r_3, \ldots, r_J)$, (B.19) for such $\boldsymbol{r}$ can be equivalently written as

$$\begin{aligned} &\sum_{\boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}} \prod_{\substack{j>2 \\ r_j=1}} \theta_{j,\,[0,\boldsymbol{\alpha}_{2:K}]} \cdot (\nu_{[0,\boldsymbol{\alpha}_{2:K}]} + \nu_{[1,\boldsymbol{\alpha}_{2:K}]}) \\ =& \sum_{\boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}} \prod_{\substack{j>2 \\ r_j=1}} \bar{\theta}_{j,\,[0,\boldsymbol{\alpha}_{2:K}]} \cdot (\bar{\nu}_{[0,\boldsymbol{\alpha}_{2:K}]} + \bar{\nu}_{[1,\boldsymbol{\alpha}_{2:K}]}), \end{aligned} \tag{B.21}$$

where $\mathcal{R}^{Q'}$ is the row-union space of Q', i.e.,

$$\mathcal{R}^{Q'} = \{\mathbf{0}_{K-1}^\top\} \cup \{\boldsymbol{\alpha} = \vee_{h \in S}\, \boldsymbol{q}'_h : \forall S \subseteq \{3, \ldots, J\}\}.$$

(B.21) involves 2^{J-2} equations with $(r_3,\ldots,r_j)$ freely ranging in $\{0,1\}^{J-2}$, which indicates that $\theta_{j,[0,\boldsymbol{\alpha}_{2:K}]}$ and $(\nu_{[0,\boldsymbol{\alpha}_{2:K}]}+\nu_{[1,\boldsymbol{\alpha}_{2:K}]})$ can be viewed as item parameter and proportion parameter associated with the model under the $(J-2)\times(K-1)$ sub-matrix Q'. Since the sub-matrix Q' satisfies conditions (C1*) and (C2*), Theorem III.1 and the set of equations (B.21) lead to

$$\forall j\geq 3,\quad \theta_{j,[0,\boldsymbol{\alpha}_{2:K}]}=\bar{\theta}_{j,[0,\boldsymbol{\alpha}_{2:K}]},\quad \nu_{[0,\boldsymbol{\alpha}_{2:K}]}+\nu_{[1,\boldsymbol{\alpha}_{2:K}]}=\bar{\nu}_{[0,\boldsymbol{\alpha}_{2:K}]}+\bar{\nu}_{[1,\boldsymbol{\alpha}_{2:K}]}.$$

This implies for any item $j\geq 3$, the item parameters θ_j^+ and θ_j^- associated with the original Q-matrix are identifiable.

Now consider an arbitrary response pattern $\boldsymbol{r}=(r_1,r_2,r_3,\ldots,r_J)$. We claim that (B.19) for $\boldsymbol{r}$ can be equivalently written as

$$\begin{aligned}&\sum_{\boldsymbol{\alpha}_{2:K}\in\mathcal{R}^{Q'}}\prod_{\substack{j>2\\ r_j=1}}\theta_{j,[0,\boldsymbol{\alpha}_{2:K}]}\cdot\mathbb{P}(R_1\geq r_1,\,R_2\geq r_2,\,\mathbf{A}_{2:K}=\boldsymbol{\alpha}_{2:K})\\ =&\sum_{\boldsymbol{\alpha}_{2:K}\in\mathcal{R}^{Q'}}\prod_{\substack{j>2\\ r_j=1}}\bar{\theta}_{j,[0,\boldsymbol{\alpha}_{2:K}]}\cdot\overline{\mathbb{P}}(R_1\geq r_1,\,R_2\geq r_2,\,\mathbf{A}_{2:K}=\boldsymbol{\alpha}_{2:K}),\end{aligned}\tag{B.22}$$

where $\mathbb{P}(R_1\geq r_1,\,R_2\geq r_2,\,\mathbf{A}_{2:K}=\boldsymbol{\alpha}_{2:K})$ represents the probability of $\{R_1\geq r_1,\,R_2\geq r_2\}$ and the attribute profile $\mathbf{A}$ has its last $K-1$ entries being $\boldsymbol{\alpha}_{2:K}$ under the set of model parameters $(\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{\nu})$, while $\overline{\mathbb{P}}(R_1\geq r_1,\,R_2\geq r_2,\,\mathbf{A}_{2:K}=\boldsymbol{\alpha}_{2:K})$ represents that under model parameters $(\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{\theta}}^-,\bar{\boldsymbol{\nu}})$. The reason (B.19) can be equivalently written as (C.32) is that, given any $\boldsymbol{\alpha}_{2:K}\in\mathcal{R}^{Q'}$ and any item $j\in\{3,\ldots,J\}$, the positive response probability of $[\alpha_1,\boldsymbol{\alpha}_{2:K}]$ to item j only depends on $\boldsymbol{\alpha}^*$ part, regardless of the value of α_1, as shown in (B.20). Therefore the terms in $T(\boldsymbol{\Theta})_{\boldsymbol{r},\boldsymbol{\cdot}}\boldsymbol{\nu}$ can be grouped in such a way that it becomes the summation over all the $\boldsymbol{\alpha}_{2:K}\in\mathcal{R}^{Q'}$, exactly as presented in Equation (C.32).

A key observation is that, taking (r_1,r_2) to be $(0,1)$, $(1,0)$, $(1,1)$ in (C.22) respectively, we obtain another three sets of equations expressed in the form of

(C.32), which are exactly in the same form as (B.21) by just replacing $\nu_{[0,\boldsymbol{\alpha}_{2:K}]}$ by $\mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K})$. Actually, taking $(r_1, r_2) = (0, 0)$ gives $\mathbb{P}(R_1 \geq 0,\, R_2 \geq 0, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K}) = \nu_{[0,\boldsymbol{\alpha}_{2:K}]}$. By Theorem III.1, this key observation results in that, for any $(r_1, r_2) \in \{0, 1\}^2$ and any $\boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}$,

$$\begin{aligned} &\mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K}) \\ &= \overline{\mathbb{P}}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K}). \end{aligned} \tag{B.23}$$

We will rely on (C.32) and the above equality (B.23) to proceed with the proof. Now consider two types of combinations of row vectors of Q', categorized based on their relationships with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. In the following proof, write $R_1 \geq r_1, R_2 \geq r_2$ succinctly as $\boldsymbol{R}_{1:2} \succeq \boldsymbol{r}_{1:2}$. We consider the following cases (a*) and (b*).

(a*) In this case, there exists two row vectors $\boldsymbol{v}_0$ and $\boldsymbol{v}'_0$ of Q' s.t. $\boldsymbol{v}_0 \succeq \boldsymbol{v}_1$, $\boldsymbol{v}_0 \nsucceq \boldsymbol{v}_2$, and $\boldsymbol{v}'_0 \nsucceq \boldsymbol{v}_1$, $\boldsymbol{v}'_0 \succeq \boldsymbol{v}_2$.

Consider $\mathbf{A}_{2:K} = \boldsymbol{v}_0$, then $\boldsymbol{v}_0 \succeq \boldsymbol{v}_1$, $\boldsymbol{v}_0 \nsucceq \boldsymbol{v}_2$ imply that

$$\begin{aligned} &\mathbb{P}(\boldsymbol{R}_{1:2} \succeq \boldsymbol{r}_{1:2},\, \mathbf{A}_{2:K} = \boldsymbol{v}_0) \\ &= \begin{cases} \nu_{[0,\boldsymbol{v}_0]} + \nu_{[1,\boldsymbol{v}_0]}, & (r_1, r_2) = (0, 0); \\ \theta_1^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+ \cdot \nu_{[1,\boldsymbol{v}_0]}, & (r_1, r_2) = (1, 0); \\ \theta_2^- \cdot (\nu_{[0,\boldsymbol{v}_0]} + \nu_{[1,\boldsymbol{v}_0]}), & (r_1, r_2) = (0, 1); \\ \theta_2^- \cdot (\theta_1^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+ \cdot \nu_{[1,\boldsymbol{v}_0]}), & (r_1, r_2) = (1, 1). \end{cases} \end{aligned}$$

Note that $\overline{\mathbb{P}}(\boldsymbol{R}_{1:2} \succeq \boldsymbol{r}_{1:2}, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K})$ takes the similar form as $\mathbb{P}(\boldsymbol{R}_{1:2} \succeq \boldsymbol{r}_{1:2}, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K})$, so in order to ensure (B.23) the following equations must hold

$$\begin{cases} \nu_{[0,\boldsymbol{v}_0]} + \nu_{[1,\boldsymbol{v}_0]} = \bar{\nu}_{[0,\boldsymbol{v}_0]} + \bar{\nu}_{[1,\boldsymbol{v}_0]}; \\ \theta_1^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+ \cdot \nu_{[1,\boldsymbol{v}_0]} = \bar{\theta}_1^- \cdot \bar{\nu}_{[0,\boldsymbol{v}_0]} + \bar{\theta}_1^+ \cdot \bar{\nu}_{[1,\boldsymbol{v}_0]}; \\ \theta_2^- \cdot (\nu_{[0,\boldsymbol{v}_0]} + \nu_{[1,\boldsymbol{v}_0]}) = \bar{\theta}_2^- \cdot (\bar{\nu}_{[0,\boldsymbol{v}_0]} + \bar{\nu}_{[1,\boldsymbol{v}_0]}); \\ \theta_2^- \cdot (\theta_1^- \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+ \nu_{[1,\boldsymbol{v}_0]}) = \bar{\theta}_2^- \cdot (\bar{\theta}_1^- \bar{\nu}_{[0,\boldsymbol{v}_0]} + \bar{\theta}_1^+ \bar{\nu}_{[1,\boldsymbol{v}_0]}). \end{cases} \tag{B.24}$$

Taking the ratio of the third and the first equation above gives $\theta_2^- = \bar{\theta}_2^-$. Similarly, $\boldsymbol{v}_0' \not\succeq \boldsymbol{v}_1$, $\boldsymbol{v}_0' \succeq \boldsymbol{v}_2$ also imply $\theta_1^- = \bar{\theta}_1^-$. Plugging $\theta_1^- = \bar{\theta}_1^-$ back to the second equation in (B.24) gives $\theta_1^+ = \bar{\theta}_1^+$, and similarly $\theta_2^+ = \bar{\theta}_2^+$.

(b*) In case (b*), there exist two row vectors $\boldsymbol{v}_0$, $\boldsymbol{v}_0'$ of Q' such that $\boldsymbol{v}_0 \not\succeq \boldsymbol{v}_1$, $\boldsymbol{v}_0 \not\succeq \boldsymbol{v}_2$, and $\boldsymbol{v}_0' \succeq \boldsymbol{v}_1$, $\boldsymbol{v}_0' \succeq \boldsymbol{v}_2$.

Consider $\mathbf{A}_{2:K} = \boldsymbol{v}_0$, then $\boldsymbol{v}_0 \not\succeq \boldsymbol{v}_1$, $\boldsymbol{v}_0 \not\succeq \boldsymbol{v}_2$ imply that the attribute profiles $(1, \boldsymbol{v}_0)$, $(0, \boldsymbol{v}_0)$ both belong to the same equivalence class $[1, \boldsymbol{v}_0]$ induced by Q, and hence

$$\mathbb{P}(\boldsymbol{R}_{1:2} \succeq \boldsymbol{r}_{1:2},\, \mathbf{A}_{2:K} = \boldsymbol{v}_0) = \begin{cases} \nu_{[0,\boldsymbol{v}_0]}, & (r_1, r_2) = (0,0); \\ \theta_1^- \cdot \nu_{[0,\boldsymbol{v}_0]}, & (r_1, r_2) = (1,0); \\ \theta_2^- \cdot \nu_{[0,\boldsymbol{v}_0]}, & (r_1, r_2) = (0,1); \\ \theta_1^- \theta_2^- \cdot \nu_{[0,\boldsymbol{v}_0]}, & (r_1, r_2) = (1,1). \end{cases}$$

With $f_{(r_1,r_2),\boldsymbol{v}_0}$ taking the above form, (B.23) implies $\theta_1^- = \bar{\theta}_1^-$ and $\theta_2^- = \bar{\theta}_2^-$.

Then consider $\mathbf{A}_{2:K} = \boldsymbol{v}_0'$, then $\boldsymbol{v}_0' \succeq \boldsymbol{v}_1$ and $\boldsymbol{v}_0' \succeq \boldsymbol{v}_2$ imply that

$$
\begin{aligned}
&\mathbb{P}(\boldsymbol{R}_{1:2} \succeq \boldsymbol{r}_{1:2},\, \mathbf{A}_{2:K} = \boldsymbol{v}_0') \\
&= \begin{cases}
\nu_{[0,\boldsymbol{v}_0]} + \nu_{[1,\boldsymbol{v}_0]}, & (r_1, r_2) = (0,0); \\
\theta_1^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+ \cdot \nu_{[1,\boldsymbol{v}_0]}, & (r_1, r_2) = (1,0); \\
\theta_2^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_2^+ \cdot \nu_{[1,\boldsymbol{v}_0]}, & (r_1, r_2) = (0,1); \\
\theta_1^-\theta_2^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+\theta_2^+ \cdot \nu_{[1,\boldsymbol{v}_0]}, & (r_1, r_2) = (1,1).
\end{cases}
\end{aligned}
$$

With the above form of $\mathbb{P}(\boldsymbol{R}_{1:2} \succeq \boldsymbol{r}_{1:2},\, \mathbf{A}_{2:K} = \boldsymbol{v}_0')$, (B.23) gives that

$$
\begin{cases}
\nu_{[0,\boldsymbol{v}_0]} + \nu_{[1,\boldsymbol{v}_0]} = \bar{\nu}_{[0,\boldsymbol{v}_0]} + \bar{\nu}_{[1,\boldsymbol{v}_0]}; \\
\theta_1^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+ \cdot \nu_{[1,\boldsymbol{v}_0]} = \theta_1^- \cdot \bar{\nu}_{[0,\boldsymbol{v}_0]} + \bar{\theta}_1^+ \cdot \bar{\nu}_{[1,\boldsymbol{v}_0]}; \\
\theta_2^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_2^+ \cdot \nu_{[1,\boldsymbol{v}_0]} = \theta_2^- \cdot \bar{\nu}_{[0,\boldsymbol{v}_0]} + \bar{\theta}_2^+ \cdot \bar{\nu}_{[1,\boldsymbol{v}_0]}; \\
\theta_1^-\theta_2^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \theta_1^+\theta_2^+ \cdot \nu_{[1,\boldsymbol{v}_0]} = \theta_1^-\theta_2^- \cdot \nu_{[0,\boldsymbol{v}_0]} + \bar{\theta}_1^+\bar{\theta}_2^+ \cdot \nu_{[1,\boldsymbol{v}_0]}.
\end{cases}
$$

where $\theta_1^- = \bar{\theta}_1^-$ and $\theta_2^- = \bar{\theta}_2^-$ are used. Solving the above equations gives $\theta_1^+ = \bar{\theta}_1^+$ and $\theta_2^+ = \bar{\theta}_2^+$.

Based on the above discussion, if Q' contains either of the type-(a*) or type-(b*) combinations of row vectors $\boldsymbol{v}_0$ and $\boldsymbol{v}_0'$, then we have $\theta_1^- = \bar{\theta}_1^-$, $\theta_2^- = \bar{\theta}_2^-$, $\theta_1^+ = \bar{\theta}_1^+$ and $\theta_2^+ = \bar{\theta}_2^+$, and hence by Proposition III.1, the grouped proportion parameters $\boldsymbol{\nu}$ are identifiable.

Note that the arguments in (a*) and (b*) above do not depend on the assumption that $\boldsymbol{v}_0$ or $\boldsymbol{v}_0'$ are single row vectors of Q'. Actually, if there exist two disjoint sets of items $S_1, S_2 \subseteq \{3, \ldots, J\}$ such that

$$\boldsymbol{v}_0 = \vee_{h \in S_1} \boldsymbol{q}_h', \quad \boldsymbol{v}_0' = \vee_{h \in S_2} \boldsymbol{q}_h',$$

and the pair $(\boldsymbol{v}_0, \boldsymbol{v}'_0)$ satisfy either the type-(a*) or the type-(b*) constraint (namely Either $\boldsymbol{v}_0 \succeq \boldsymbol{v}_1$, $\boldsymbol{v}_0 \nsucceq \boldsymbol{v}_2$ and $\boldsymbol{v}'_0 \nsucceq \boldsymbol{v}_1$, $\boldsymbol{v}'_0 \succeq \boldsymbol{v}_2$; Or $\boldsymbol{v}_0 \nsucceq \boldsymbol{v}_1$, $\boldsymbol{v}_0 \nsucceq \boldsymbol{v}_2$ and $\boldsymbol{v}'_0 \succeq \boldsymbol{v}_1$, $\boldsymbol{v}'_0 \succeq \boldsymbol{v}_2$), then the arguments in (a*), (b*) still hold, and the conclusion of partial identifiability follows. Next we show such pair $(\boldsymbol{v}_0, \boldsymbol{v}'_0)$ must exist. The item set $\{3, \ldots, J\}$ can be decomposed as $\{3, \ldots, J\} := S_{00} \cup S_{10} \cup S_{02} \cup S_{12}$ where

$$
\begin{aligned}
S_{00} &= \{3 \leq j \leq J : \boldsymbol{q}'_j \nsucceq \boldsymbol{v}_1,\ \boldsymbol{q}'_j \nsucceq \boldsymbol{v}_2\},\\
S_{10} &= \{3 \leq j \leq J : \boldsymbol{q}'_j \succeq \boldsymbol{v}_1,\ \boldsymbol{q}'_j \nsucceq \boldsymbol{v}_2\},\\
S_{02} &= \{3 \leq j \leq J : \boldsymbol{q}'_j \nsucceq \boldsymbol{v}_1,\ \boldsymbol{q}'_j \succeq \boldsymbol{v}_2\},\\
S_{12} &= \{3, \ldots, J\} \setminus (S_{00} \cup S_{10} \cup S_{02}).
\end{aligned}
$$

The assumption that Q' satisfies condition (C1*), implies that there exists $\boldsymbol{v}'_0 \in \mathcal{R}^{Q'}$ such that $\boldsymbol{v}'_0 \succeq \boldsymbol{v}_1$, $\boldsymbol{v}'_0 \succeq \boldsymbol{v}_2$. So if for $i = 1, 2$, (a) is satisfied, then the type-(b*) combinations of row vectors exist in Q'. While if (a) is not satisfied and (b) is satisfied, then we claim that $S_{10} \neq \varnothing$ and $S_{02} \neq \varnothing$. This is because if $S_{10} = \varnothing$, then together with the fact that $S_{00} = \varnothing$ implied by the failure of (a), we will have $\{3, \ldots, J\} = S_{02} \cup S_{12}$. But this means for any item $j \geq 3$, $\boldsymbol{q}'_j \succeq \boldsymbol{v}_2$, contradictory to the assumption of case (b). So $S_{10} \neq \varnothing$ must hold, and similarly $S_{02} \neq \varnothing$ must hold. This ensures the type-(b*) combinations of row vectors exist in Q'. In either scenarios, Q' contains at least one of type-(a*) or type-(b*) combinations of row vectors, so we obtain the identifiability of all the item parameters. Applying Proposition III.1 gives the identifiability of the grouped proportion parameters $\boldsymbol{\nu}$, which completes the proof of part (B.2).

Proof of Part (b.1). Under the assumptions in part (B.1), the Q-matrix takes the following form

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \boldsymbol{v}^\top \\ \mathbf{0} & Q' \end{pmatrix}. \tag{B.25}$$

Since there exists a single-attribute item with $\boldsymbol{q}$-vector being $(1, \mathbf{0}^\top)$, for any $\boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}$ we have $[0, \boldsymbol{\alpha}_{2:K}] \neq [1, \boldsymbol{\alpha}_{2:K}]$, where the equivalence class notation $[\cdot]$ represents that induced by the $J \times K$ Q-matrix Q. Then following the similar arguments as in the proof of part (B.2), Equation (C.22) hold as long as the following set of equations hold

$$\begin{cases} \nu_{[0,\boldsymbol{\alpha}_{2:K}]} + \nu_{[1,\boldsymbol{\alpha}_{2:K}]} = \bar{\nu}_{[0,\boldsymbol{\alpha}_{2:K}]} + \bar{\nu}_{[1,\boldsymbol{\alpha}_{2:K}]}, & \forall \boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}; \\ \theta_1^- \cdot \nu_{[0,\boldsymbol{\alpha}_{2:K}]} + \theta_1^+ \cdot \nu_{[1,\boldsymbol{\alpha}_{2:K}]} & \\ \qquad = \theta_1^- \cdot \bar{\nu}_{[0,\boldsymbol{\alpha}_{2:K}]} + \bar{\theta}_1^+ \cdot \bar{\nu}_{[1,\boldsymbol{\alpha}_{2:K}]}, & \forall \boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}; \\ \theta_2^- \cdot \nu_{[0,\boldsymbol{\alpha}_{2:K}]} + \theta_2^+ \cdot \nu_{[1,\boldsymbol{\alpha}_{2:K}]} & \\ \qquad = \theta_2^- \cdot \bar{\nu}_{[0,\boldsymbol{\alpha}_{2:K}]} + \bar{\theta}_2^+ \cdot \bar{\nu}_{[1,\boldsymbol{\alpha}_{2:K}]}, & \forall \boldsymbol{\alpha}_{2:K} \succeq \boldsymbol{v},\ \boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}; \\ \theta_1^-\theta_2^- \cdot \nu_{[0,\boldsymbol{\alpha}_{2:K}]} + \theta_1^+\theta_2^+ \cdot \nu_{[1,\boldsymbol{\alpha}_{2:K}]} & \\ \qquad = \theta_1^-\theta_2^- \cdot \nu_{[0,\boldsymbol{\alpha}_{2:K}]} + \bar{\theta}_1^+\bar{\theta}_2^+ \cdot \nu_{[1,\boldsymbol{\alpha}_{2:K}]}, & \forall \boldsymbol{\alpha}_{2:K} \succeq \boldsymbol{v},\ \boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}. \end{cases} \tag{B.26}$$

Now consider a set of parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$ such that $\nu_{[0,\boldsymbol{\alpha}_{2:K}]} = \rho \cdot \nu_{[1,\boldsymbol{\alpha}_{2:K}]}$ for any $\boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}$, where ρ is a positive constant. Setting $\theta_1^+ = \bar{\theta}_1^+$, $\theta_2^- = \bar{\theta}_2^-$, $\theta_j^+ = \bar{\theta}_j^+$ and $\theta_j^- = \bar{\theta}_j^-$ for $j = 3, \ldots, J$ and freely choosing any valid $\bar{\theta}_1^-$ which is not equal to θ_1^-, we construct the remaining parameters $(\bar{\theta}_2^+, \bar{\boldsymbol{\nu}})$ as follows. Let

$$\bar{\theta}_2^+ = \frac{(\theta_1^+ - \bar{\theta}_1^-)(\theta_2^+ - \theta_2^-)}{(\theta_1^+ - \bar{\theta}_1^-) + \rho(\theta_1^- - \bar{\theta}_1^-)} + \bar{\theta}_2^-,$$

and for any $\boldsymbol{\alpha}_{2:K} \in \mathcal{R}^{Q'}$ let

$$\bar{\nu}_{[1,\boldsymbol{\alpha}_{2:K}]} = \frac{(\theta_1^+ - \bar{\theta}_1^-) + \rho(\theta_1^- - \bar{\theta}_1^-)}{\theta_1^+ - \bar{\theta}_1^-}\nu_{[1,\boldsymbol{\alpha}_{2:K}]},$$
$$\bar{\nu}_{[0,\boldsymbol{\alpha}_{2:K}]} = \nu_{[0,\boldsymbol{\alpha}_{2:K}]} + \nu_{[1,\boldsymbol{\alpha}_{2:K}]} - \bar{\nu}_{[1,\boldsymbol{\alpha}_{2:K}]},$$

then by direct calculations one can check (B.26) hold. Therefore, we have found another set of parameters $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{\nu}})$ such that $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{\nu}}) \neq (\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$ and $T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{\nu} = T(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{\nu}}$, which shows the non-identifiability of the model parameters under the Q in the form of (B.25). This completes the proof of part (B.1). $\square$

Proof of Theorem III.3. Without loss of generality, we again focus on the proof of the conclusion for the two-parameter conjunctive models since all the arguments also hold for the compensatory models, following the similar argument in the proof of Proposition III.3. Suppose condition (C1*) holds. Without loss of generality, suppose condition (C2**) does not hold for some basis item j, and suppose that the first K_1 entries of the row vector $\boldsymbol{q}_j$ in the Q-matrix corresponding to this basis item are 1's and the remaining $K - K_1$ entries of $\boldsymbol{q}$ are 0's, i.e.

$$\boldsymbol{q}_j = (\underbrace{1, \ldots, 1,}_{\text{columns } 1,\ldots,K_1} \quad 0, \ldots, 0).$$

Denote $S_{-j} = \{1, \ldots, J\} \setminus \{j\}$. Since j is a basis item, any item in S_{-j} requires some attribute not required by j, i.e.

$$\forall h \in S_{-j}, \quad q_{h,k} = 1 \text{ for some } k \in \{K_1 + 1, \ldots, K\}.$$

We claim, the assumption that (C2**) does not hold for item h, implies that row vectors of items in S_{-j} can be arranged in a way $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{J-1}\}$ such that for any $2 \leq i \leq J-1$, $\boldsymbol{u}_i$ requires at least one more attribute in $\{K_1 + 1, \ldots, K\}$ that is not required by $\cup_{1\leq s\leq i-1}\{\boldsymbol{u}_s\}$. This claim is true since otherwise for some $h \in S_{-j}$ and $S_0 \subseteq S_{-j}\backslash\{h\}$, the difference of attributes required by $\{h\}$ and S_0 are only among $\{1,\ldots,K_1\}$, then taking $S_j^- = S_0$ and $S_j^+ = S_0 \cup \{h\}$ makes (C2**) hold for item j. In other words, for some $1 \leq k_1 < k_2 < \ldots < k_{J-1} \leq K - K_1$ we have that

$$\boldsymbol{w}_1 = \boldsymbol{v}_{k_1}, \quad \boldsymbol{w}_2 = \boldsymbol{v}_{k_2}, \quad \ldots, \quad \boldsymbol{w}_{J-1} = \boldsymbol{v}_{k_{J-1}},$$

where $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ takes the form as follows

$$\begin{array}{r|ccc|ccccc}
\boldsymbol{q}_j: & 1 & \cdots & 1 & 0 & 0 & \cdots & \cdots & 0 \\
\hline
\boldsymbol{v}_1: & * & \cdots & * & 1 & 0 & \cdots & \cdots & 0 \\
\boldsymbol{v}_2: & * & \cdots & * & * & 1 & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\boldsymbol{v}_{K-K_1-1}: & * & \cdots & * & * & * & * & 1 & 0 \\
\boldsymbol{v}_{K-K_1}: & * & \cdots & * & * & * & * & * & 1
\end{array}, \tag{B.27}$$

Now we are ready to construct two different sets of parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu}) \neq (\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{\nu}})$ that give (C.1), i.e.

$$T(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)\boldsymbol{\nu} = T(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{\nu}}.$$

Given $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$, condition (C1*) guarantees $\boldsymbol{\theta}^+ = \bar{\boldsymbol{\theta}}^+$ and $\theta_j^- = \bar{\theta}_j^-$ for $j \in S_{non}$. Equation (C.1) holds if for another set of parameters $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{\theta}}^-, \bar{\boldsymbol{\nu}})$, the following equations hold for any $\boldsymbol{w}_i$ such that $\boldsymbol{w}_i \overset{\Gamma}{\nsim} \boldsymbol{w}_i \vee \boldsymbol{q}_j$

$$\begin{cases}
\nu_{[\boldsymbol{w}_i]} + \nu_{[\boldsymbol{q}_j \vee \boldsymbol{w}_i]} = \bar{\nu}_{[\boldsymbol{w}_i]} + \bar{\nu}_{[\boldsymbol{q}_j \vee \boldsymbol{w}_i]}; \\
\theta_j^- \cdot \nu_{[\boldsymbol{w}_i]} + \theta_j^+ \cdot \nu_{[\boldsymbol{q}_j \vee \boldsymbol{w}_i]} = \bar{\theta}_j^- \cdot \nu_{[\boldsymbol{w}_i]} + \theta_j^+ \cdot \nu_{[\boldsymbol{q}_j \vee \boldsymbol{w}_i]},
\end{cases} \tag{B.28}$$

with any other parameter not specified in (B.28) equal to its counterpart in the original set of parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{\nu})$. Denote the cardinality of the set $\mathcal{W} = \{\boldsymbol{w}_i :$

$\boldsymbol{w}_i \stackrel{\Gamma}{\nsim} \boldsymbol{w}_i \vee \boldsymbol{q}_j\}$ by $|\mathcal{W}|$. The set $\mathcal{W}$ is nonempty since $\mathbf{0} \stackrel{\Gamma}{\nsim} \boldsymbol{q}_j$ and $\mathbf{0} \in \mathcal{W}$, where Γ is the Γ-matrix corresponding to the saturated latent class space $\mathcal{A} = \{0,1\}^K$. Note that (B.28) involve $2|\mathcal{W}|+1$ free parameters $\{\bar{\theta}_j^-\} \cup \{\bar{\nu}_{[\boldsymbol{w}_i]},\ \bar{\nu}_{[\boldsymbol{w}_i \vee \boldsymbol{q}]} : \boldsymbol{w}_i \in \mathcal{W}\}$ while only contain $2|\mathcal{W}|$ equations, so there are infinitely many solutions to (B.28). This proves the non-identifiability of the model parameters. $\square$

Proof of Theorem III.4. First prove the claim that conditions (C1*) and (C2*) are equivalent to conditions (C1′) and (C2′) under the assumption that the Q-matrix is complete and $p_{\boldsymbol{\alpha}} > 0$ for any $\boldsymbol{\alpha} \in \{0,1\}^K$. Theorem 1 in Gu and Xu (2019b) established that if Q is complete and $p_{\boldsymbol{\alpha}} > 0$ for any $\boldsymbol{\alpha} \in \{0,1\}^K$, then conditions (C1′) and (C2′) combined is sufficient and necessary for the identifiability of the DINA model parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$. Since (C1*) and (C2*) are sufficient conditions for identifiability, they must imply the necessary conditions (C1′) and (C2′). In the following we prove the other direction, i.e., conditions (C1′) and (C2′) imply conditions (C1*) and (C2*).

When Q is complete, if condition (C1′) holds that attribute k is required by at least three items in the Q-matrix, then for each unit vector $\boldsymbol{e}_k$ as the $\boldsymbol{q}$-vector, there must exist two other items j_k^1 and j_k^2 that also measure attribute k. Let $S_k^i = \{j_k^i\}$, $i = 1,2$, then S_k^1 and S_k^2 are the two disjoint item sets that satisfy condition (C1*) that $\boldsymbol{e}_k = \boldsymbol{q}_k \preceq \vee_{h \in S_k^i} \boldsymbol{q}_h = \boldsymbol{q}_{j_k^i}$ for $i = 1$ and 2. This shows (C1′) implies (C1*).

Assume without loss of generality that Q takes the form

$$Q = \begin{pmatrix} \mathcal{I}_K \\ Q' \end{pmatrix}. \tag{B.29}$$

If condition (C2′) is satisfied, we next explicitly construct a procedure that sequentially expands the separator set S_{sep} until $S_{sep} = \mathcal{S}$ finally, which by Theorem III.1 would establish identifiability of all the model parameters. The existence of such sequential procedure would ensure the Sequentially Differentiable Condition (C2*) holds. Theorem III.1 has already established that condition (C1*) suffices for the identifiability of all the slipping parameters, and that of the guessing parameters of the non-basis items. Specifically for the complete Q-matrix in the form of (B.29), this conclusion implies $\theta_j^+ = \bar{\theta}_j^+$ for all $j = 1, \ldots, J$ and $\theta_j^- = \bar{\theta}_j^-$ for all $j = K+1, \ldots, J$, because any item $j > K$ must be a non-basis item in the sense that there always exists some item $k \in \{1, \ldots, K\}$ such that $\boldsymbol{q}_k = \boldsymbol{e}_k \preceq \boldsymbol{q}_j$. It remains to show the guessing parameters of the first K items are identifiable, i.e., $\theta_k^- = \bar{\theta}_k^-$ for $k = 1, \ldots, K$. For any binary vectors $\boldsymbol{a} = (a_1, \ldots, a_L)$, $\boldsymbol{b} = (b_1, \ldots, b_L)$ of the same length, we say $\boldsymbol{a}$ is lexicographically smaller than $\boldsymbol{b}$, denoted by $\boldsymbol{a} \prec_{\text{lex}} \boldsymbol{b}$, if either $a_1 < b_1$; or there exists some $2 \leq i \leq l$ such that $a_i < b_i$ and $a_j = b_j$ for all $j < i$. Now that the K column vectors of Q' are mutually distinct, there is a unique permutation $(m_1, m_2, \ldots, m_K)$ of $(1, 2, \ldots, K)$ such that $Q'_{\bullet,m_1} \prec_{\text{lex}} Q'_{\bullet,m_2} \prec_{\text{lex}} \cdots \prec_{\text{lex}} Q'_{\bullet,m_K}$. For any $1 \leq i < j \leq K$, since $Q'_{\bullet,m_i} \prec_{\text{lex}} Q'_{\bullet,m_j}$, we must have $Q'_{\bullet,m_i} \not\succeq Q'_{\bullet,m_j}$. This fact will be useful in the following proof.

We start with the initial separator set $S_{sep} := S_0 = \{K+1, \ldots, J\}$. Note that at this starting stage $S_{sep} \subseteq S_{non}$. We next argue that item m_1 is S_0-differentiable, and further, m_i is $(S_0 \cup \{m_1, \ldots, m_{i-1}\})$-differentiable for all $i = 2, \ldots, K$. Noting that $Q_{\bullet,m_1}$ is of the smallest lexicographic order among all the column vectors of the submatrix Q', define

$$S_{m_1}^- = \{j \in S_0 : q_{j,m_1} = 0\},$$

then $\vee_{h \in S_0} \boldsymbol{q}_h$ equals the all-one vector under condition (C1$'$) while $\vee_{h \in S_{m_1}^-} \boldsymbol{q}_h$ equals the vector that is zero in the m_1th entry and one otherwise, i.e.,

$$\begin{aligned}
\vee_{h \in S_0} \boldsymbol{q}_h &= (1, \ldots, 1), \\
\vee_{h \in S_{m_1}^-} \boldsymbol{q}_h &= (1, \ldots, 1, \underbrace{0}_{\text{column } m_1}, 1, \ldots, 1),
\end{aligned}$$

so $\vee_{h\in S_0}\boldsymbol{q}_h - \vee_{h\in S^-_{m_1}}\boldsymbol{q}_h = \boldsymbol{e}^\top_{m_1} = \boldsymbol{q}_{m_1}$. By definition of S-differentiable, this means item m_1 is S_0-differentiable. Then expand the separator set by including item m_1 in it, i.e. let $S_{sep} := S_0 \cup \{m_1\}$. Now further define $S^-_{m_2} = \{j \in S_0 : Q_{j,m_2} = 0\} \cup \{m_1\}$, then $S^-_{m_2} \subseteq S_{sep}$. Similarly it is easy to check

$$\begin{aligned}
\vee_{h\in S_{sep}}\boldsymbol{q}_h &= (1,\dots,1),\\
\vee_{h\in S^-_{m_2}}\boldsymbol{q}_h &= (1,\dots,1,\underbrace{0}_{\text{column } m_2},1,\dots,1),
\end{aligned}$$

and this implies item m_2 is S_{sep}-differentiable. The similar argument would give that m_i is $(S_0 \cup \{m_1,\dots,m_{i-1}\})$-differentiable for all $i = 2,\dots,K$, so the sequential expanding procedure ends up with $S_{sep} = \{1,\dots,J\} = \mathcal{S}$. Note that we start with an initial separator set S_0 that is a subset of S_{non} and in each expanding step we included exactly one more item into S_{sep} even if we might have included more (all the items that are S_{sep}-differentiable could be included, which can be more than one), the fact that in our procedure S_{sep} finally equals $\mathcal{S}$ actually proves a stronger conclusion than the existence of a sequential procedure described in condition (C2*), so the Sequentially Differentiable Condition (C2*) holds. By now we have shown conditions (C1$'$) and (C2$'$) also imply conditions (C1*) and (C2*).

Since (C1$'$) and (C2$'$) combined is necessary, (C1*) and (C2*) combined is also necessary. This completes the proof of the theorem that (C1*) and (C2*) are sufficient and necessary for strict identifiability of the two-parameter model when the Q-matrix is complete and $p_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$. □

B.3 Proof of Main Results in Section 3.3

We introduce a useful lemma before proving Theorem III.5 and Theorem III.7, the results of strict identifiability of multi-parameter restricted latent class models.

The proof of the following lemma is given in Section B.4. For notational simplicity, we denote $\theta_{j,\mathbf{1}} := \max_{\boldsymbol{\alpha}:\Gamma_{j,\boldsymbol{\alpha}}=1} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\Gamma_{j,\boldsymbol{\alpha}}=1} \theta_{j,\boldsymbol{\alpha}}$ in the following discussion.

Lemma B.3. *For an arbitrary restricted latent class model satisfying constraints* (3.2)*, if Equation (*C.1*) holds, then for any $j \in S_1 \cup S_2$ and any $\boldsymbol{\alpha}$ such that $\Gamma_{j,\boldsymbol{\alpha}} = 0$,*

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \neq \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}, \quad \theta_{\boldsymbol{e}_j,\mathbf{1}} \neq \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}.$$

To prove Theorem III.5, we also need the following lemma, whose proof is given in Section B.4.

Lemma B.4. *Under the assumptions of Theorem III.5, for any $\boldsymbol{\alpha}$ there exists vectors $\boldsymbol{u}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ such that*

$$\begin{aligned}
&\{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\}_{\boldsymbol{\alpha}} \neq 0; \qquad \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\}_{\boldsymbol{\alpha}'} = 0, \quad \forall \boldsymbol{\alpha}' \precneqq_{S_1} \boldsymbol{\alpha}. \\
&\{\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\}_{\boldsymbol{\alpha}} \neq 0; \qquad \{\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\}_{\boldsymbol{\alpha}'} = 0, \quad \forall \boldsymbol{\alpha}' \precneqq_{S_2} \boldsymbol{\alpha}.
\end{aligned} \tag{B.30}$$

Proof of Theorem III.5. Equipped with Lemmas B.3 and D.1, we prove Theorem III.5 in the following three steps. Without loss of generality, assume $S_1 = \{1, \ldots, M_1\}$ and $S_2 = \{M_1 + 1, \ldots, M_1 + M_2\}$, namely item set S_1 contains the first M_1 items and item set S_2 contains the next M_2 items.

Step 1: $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}$ for $j > M_1 + M_2$.

Step 2: $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ for $j > M_1 + M_2$ and any $\boldsymbol{\alpha}$.

Step 3: $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ and $p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}$ for $1 \leq j \leq M_1 + M_2$ and any $\boldsymbol{\alpha}$.

Now we start the proof of the result step by step.

Step 1. Define $\boldsymbol{\theta}^* \in \mathbb{R}^J$ to be

$$\boldsymbol{\theta}^* = (\bar{\theta}_{\boldsymbol{e}_1,\mathbf{1}}, \ldots, \bar{\theta}_{\boldsymbol{e}_{M_1},\mathbf{1}}, \theta_{\boldsymbol{e}_{M_1+1},\mathbf{1}}, \ldots, \theta_{\boldsymbol{e}_{M_1+M_2},\mathbf{1}}, \mathbf{0}_{J-M_1-M_2})^\top,$$

and consider the row vector of the transformed T-matrix $T(\mathbf{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)$ corresponding to $\boldsymbol{r}=\sum_{k=1}^{M_1+M_2}\boldsymbol{e}_k$ is

$$\begin{aligned}T_{\sum_{k=1}^{M_1+M_2}\boldsymbol{e}_k,\cdot}(\mathbf{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top) &= \bigodot_{k=1}^{M_1+M_2} T_{\boldsymbol{e}_k,\cdot}(\mathbf{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\\ &=\left(\prod_{k=1}^{M_1}(\theta_{\boldsymbol{e}_k,\boldsymbol{\alpha}_0}-\bar{\theta}_{\boldsymbol{e}_k,\mathbf{1}})\prod_{k=1}^{M_2}(\theta_{\boldsymbol{e}_{M_1+k},\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_{M_1+k},\mathbf{1}}),\mathbf{0}_M^\top\right),\end{aligned}$$

where the last M elements of this row vector are all zero. By Lemma B.3, the first element is nonzero, i.e.,

$$\prod_{k=1}^{M_1}(\theta_{\boldsymbol{e}_k,\boldsymbol{\alpha}_0}-\bar{\theta}_{\boldsymbol{e}_k,\mathbf{1}})\prod_{k=1}^{M_2}(\theta_{\boldsymbol{e}_{M_1+k},\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_{M_1+k},\mathbf{1}})\neq 0.$$

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

$$\begin{aligned}&T_{\sum_{k=1}^{M_1+M_2}\boldsymbol{e}_k}(\bar{\mathbf{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top)\\ &=\left(\prod_{k=1}^{M_1}(\bar{\theta}_{\boldsymbol{e}_k,\boldsymbol{\alpha}_0}-\bar{\theta}_{\boldsymbol{e}_k,\mathbf{1}})\prod_{k=1}^{M_2}(\bar{\theta}_{\boldsymbol{e}_{M_1+k},\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_{M_1+k},\mathbf{1}}),\mathbf{0}_M^\top\right)\end{aligned}$$

and

$$\prod_{k=1}^{M_1}(\bar{\theta}_{\boldsymbol{e}_k,\boldsymbol{\alpha}_0}-\bar{\theta}_{\boldsymbol{e}_k,\mathbf{1}})\prod_{k=1}^{M_2}(\bar{\theta}_{\boldsymbol{e}_{M_1+k},\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_{M_1+k},\mathbf{1}})\neq 0.$$

Now consider $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}$ for any $j>M_1+M_2$. The row vectors of $T(\mathbf{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)$ and $T(\bar{\mathbf{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top)$ corresponding to the response pattern $\boldsymbol{r}=\sum_{k=1}^{M_1+M_2}\boldsymbol{e}_k+\boldsymbol{e}_j$ are

$$\begin{aligned}&T_{\sum_{k=1}^{2M}\boldsymbol{e}_k+\boldsymbol{e}_j,\cdot}(\mathbf{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\\ &=\left(\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}\prod_{k=1}^{M_1}(\theta_{\boldsymbol{e}_k,\boldsymbol{\alpha}_0}-\bar{\theta}_{\boldsymbol{e}_k,\mathbf{1}})\prod_{k=1}^{M_2}(\theta_{\boldsymbol{e}_{M_1+k},\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_{M_1+k},\mathbf{1}}),\mathbf{0}_M^\top\right),\end{aligned}$$

and

$$T_{\sum_{k=1}^{2M} \boldsymbol{e}_k + \boldsymbol{e}_j, \cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)$$
$$= \left(\bar{\theta}_{\boldsymbol{e}_j, \boldsymbol{\alpha}_0} \prod_{k=1}^{M_1} (\bar{\theta}_{\boldsymbol{e}_k, \boldsymbol{\alpha}_0} - \bar{\theta}_{\boldsymbol{e}_k, \mathbf{1}}) \prod_{k=1}^{M_2} (\bar{\theta}_{\boldsymbol{e}_{M_1+k}, \boldsymbol{\alpha}_0} - \theta_{\boldsymbol{e}_{M_1+k}, \mathbf{1}}), \mathbf{0}_M^\top \right),$$

respectively. Note Equation (C.1) implies that

$$\begin{aligned} \theta_{\boldsymbol{e}_j, \boldsymbol{\alpha}_0} &= \frac{T_{\sum_{k=1}^{M_1+M_2} \boldsymbol{e}_k + \boldsymbol{e}_j, \cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)\boldsymbol{p}}{T_{\sum_{k=1}^{M_1+M_2} \boldsymbol{e}_k, \cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)\boldsymbol{p}} \\ &= \frac{T_{\sum_{k=1}^{M_1+M_2} \boldsymbol{e}_k + \boldsymbol{e}_j, \cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)\bar{\boldsymbol{p}}}{T_{\sum_{k=1}^{M_1+M_2} \boldsymbol{e}_k, \cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)\bar{\boldsymbol{p}}} = \bar{\theta}_{\boldsymbol{e}_j, \boldsymbol{\alpha}_0}. \end{aligned}$$

Step 2. First consider any $j \in (S_1 \cup S_2)^c$. For any $\boldsymbol{\alpha}$, define

$$\boldsymbol{\theta}_{\boldsymbol{\alpha}} = \sum_{h \in S_1 : \Gamma_{h, \boldsymbol{\alpha}} = 0} \theta_{\boldsymbol{e}_h, \mathbf{1}} \boldsymbol{e}_h + \sum_{h \in S_2 : \Gamma_{h, \boldsymbol{\alpha}} = 0} \bar{\theta}_{\boldsymbol{e}_h, \mathbf{1}} \boldsymbol{e}_h,$$

and consider the row vector corresponding to response pattern $\boldsymbol{r} = \sum_{h \in S_1} \boldsymbol{e}_h$ in the transformed T-matrix, then we have

$$\begin{aligned} T_{\sum_{h \in S_1} \boldsymbol{e}_h, \boldsymbol{\alpha}'}(\boldsymbol{\Theta} - \boldsymbol{\theta}_{\boldsymbol{\alpha}} \mathbf{1}^\top) \neq 0 \quad \text{iff} \quad \boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}, \\ T_{\sum_{h \in S_2} \boldsymbol{e}_h, \boldsymbol{\alpha}'}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}_{\boldsymbol{\alpha}} \mathbf{1}^\top) \neq 0 \quad \text{iff} \quad \boldsymbol{\alpha}' \preceq_{S_2} \boldsymbol{\alpha}. \end{aligned}$$

We only prove the first inequality above and the second is just similar. Note

$$T_{\sum_{h \in S_1} \boldsymbol{e}_h, \boldsymbol{\alpha}'}(\boldsymbol{\Theta} - \boldsymbol{\theta}_{\boldsymbol{\alpha}} \mathbf{1}^\top) = \prod_{h \in S_1 : \Gamma_{h, \boldsymbol{\alpha}} = 0} (\theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}'} - \theta_{\boldsymbol{e}_h, \mathbf{1}}), \tag{B.31}$$

and if $\boldsymbol{\alpha}' \npreceq \boldsymbol{\alpha}$, then there exists some h such that $\Gamma_{h, \boldsymbol{\alpha}'} = 1$, $\Gamma_{h, \boldsymbol{\alpha}} = 0$ and hence $\theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}'} - \theta_{\boldsymbol{e}_h, \mathbf{1}} = 0$, which makes the product in (B.31) equal to 0; while if $\boldsymbol{\alpha}' \preceq \boldsymbol{\alpha}$, then for all $h \in S_1$ such that $\Gamma_{h, \boldsymbol{\alpha}} = 0$, we have $\Gamma_{h, \boldsymbol{\alpha}'} \leq \Gamma_{h, \boldsymbol{\alpha}} = 0$ and hence $\theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}'} - \theta_{\boldsymbol{e}_h, \mathbf{1}} \neq 0$, so the product in (B.31) is nonzero.

Then we use the properties of $\boldsymbol{u}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ to continue with the proof. First note that the existence of $\boldsymbol{u}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ satisfying (B.30) only rely on the full-column-rank property of $T(\boldsymbol{\Theta}_{S_i})$ and $T(\bar{\boldsymbol{\Theta}}_{S_i})$, so for some full-rank linear transformation matrix A there still exists some $\boldsymbol{u}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\alpha}}$ such that

$$\begin{aligned}
\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot A \cdot T(\bar{\boldsymbol{\Theta}}_{S_2}) &= (\mathbf{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}),\\
\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot A \cdot T(\boldsymbol{\Theta}_{S_1}) &= (\mathbf{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}),
\end{aligned}$$

and

$$\begin{aligned}
&\{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot A \cdot T(\boldsymbol{\Theta}_{S_2})\}_{\boldsymbol{\alpha}} \neq 0; \quad \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot A \cdot T(\boldsymbol{\Theta}_{S_2})\}_{\boldsymbol{\alpha}'} = 0, \ \forall \boldsymbol{\alpha}' \gneqq_{S_1} \boldsymbol{\alpha}; \qquad \text{(B.32)}\\
&\{\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot A \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\}_{\boldsymbol{\alpha}} \neq 0; \quad \{\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot A \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\}_{\boldsymbol{\alpha}'} = 0, \ \forall \boldsymbol{\alpha}' \gneqq_{S_2} \boldsymbol{\alpha}.
\end{aligned}$$

Now note that $T_{\sum_{h\in S_2}\boldsymbol{e}_h,\cdot}(\boldsymbol{\Theta}-\boldsymbol{\theta}_{\boldsymbol{\alpha}}\mathbf{1}^\top)$ can just be expressed as $D(\boldsymbol{\theta}_{\boldsymbol{\alpha}})\cdot T(\boldsymbol{\Theta}_{S_2})$ indicated by Proposition B.2, so we have

$$\begin{aligned}
&\{\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot T_{\sum_{h\in S_1}\boldsymbol{e}_h,\boldsymbol{\cdot}}(\boldsymbol{\Theta}-\boldsymbol{\theta}_{\boldsymbol{\alpha}}\mathbf{1}^\top)\} \odot \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T_{\sum_{h\in S_2}\boldsymbol{e}_h,\boldsymbol{\cdot}}(\boldsymbol{\Theta}-\boldsymbol{\theta}_{\boldsymbol{\alpha}}\mathbf{1}^\top)\} \qquad \text{(B.33)}\\
&= (\mathbf{0}, \underbrace{x_{\boldsymbol{\alpha}}}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}), \qquad \text{with } x_{\boldsymbol{\alpha}} \neq 0,
\end{aligned}$$

$$\begin{aligned}
&\{\boldsymbol{u}_{\boldsymbol{\alpha}}^\top \cdot T_{\sum_{h\in S_1}\boldsymbol{e}_h,\boldsymbol{\cdot}}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}_{\boldsymbol{\alpha}}\mathbf{1}^\top)\} \odot \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T_{\sum_{h\in S_2}\boldsymbol{e}_h,\boldsymbol{\cdot}}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}_{\boldsymbol{\alpha}}\mathbf{1}^\top)\} \qquad \text{(B.34)}\\
&= (\mathbf{0}, \underbrace{\bar{y}_{\boldsymbol{\alpha}}}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}), \qquad \text{with } \bar{y}_{\boldsymbol{\alpha}} \neq 0.
\end{aligned}$$

Note that the left hand sides of equations (B.33) and (B.34) are both row transformations of the T-matrix, namely there exists a matrix M_1 such that

$$\text{(B.33)} = M_1 \cdot T(\boldsymbol{\Theta}), \quad \text{(B.34)} = M_1 \cdot T(\bar{\boldsymbol{\Theta}}),$$

so by Equation (C.1), we have (B.33) $\cdot\, \boldsymbol{p}$ = (B.34) $\cdot\, \bar{\boldsymbol{p}} \neq 0$. Now consider any item $j \in (S_1 \cup S_2)^c$, since (B.33) and (B.34) involve rows of the T-matrices only with respect to items included in $S_1 \cup S_2$, Equation (C.1) further implies $\{T_{\boldsymbol{e}_j,\boldsymbol{\alpha}}(\boldsymbol{\Theta}) \odot \text{(B.33)}\} \cdot \boldsymbol{p} = \{T_{\boldsymbol{e}_j,\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}) \odot \text{(B.34)}\} \cdot \bar{\boldsymbol{p}}$, therefore we have the equality

$$\frac{\{T_{\boldsymbol{e}_j,\boldsymbol{\alpha}}(\boldsymbol{\Theta}) \odot \text{(B.33)}\} \cdot \boldsymbol{p}}{\text{(B.33)} \cdot \boldsymbol{p}} = \frac{\{T_{\boldsymbol{e}_j,\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}) \odot \text{(B.34)}\} \cdot \bar{\boldsymbol{p}}}{\text{(B.34)} \cdot \bar{\boldsymbol{p}}}. \tag{B.35}$$

Note that the left and right hand sides of the above equation can be written as

$$\begin{aligned} \text{LHS of (B.35)} &= \frac{\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot \text{(B.33)} \cdot \bar{\boldsymbol{p}}}{\text{(B.33)} \cdot \bar{\boldsymbol{p}}} = \theta_{j,\boldsymbol{\alpha}}, \\ \text{RHS of (B.35)} &= \frac{\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot \text{(B.34)} \cdot \bar{\boldsymbol{p}}}{\text{(B.34)} \cdot \bar{\boldsymbol{p}}} = \bar{\theta}_{j,\boldsymbol{\alpha}}, \end{aligned}$$

so $\theta_{j,\boldsymbol{\alpha}} = \bar{\theta}_{j,\boldsymbol{\alpha}}$.

Step 3. *First we prove $\theta_{\boldsymbol{e}_j,\boldsymbol{1}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{1}}$ for any $j \in S_1 \cup S_2$.* Given $\boldsymbol{\alpha}$, define

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

Note that if for some $\boldsymbol{\alpha}$, $\Gamma_{h,\boldsymbol{\alpha}} = 1$ for all $h \in S_1$, then $\boldsymbol{\theta}^*$ is defined to be the zero vector. With $\boldsymbol{\theta}^*$, the row vector corresponding to $\boldsymbol{r}^* = \sum_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 0} \boldsymbol{e}_h$ in the transformed T-matrix takes the following form

$$\begin{aligned} & T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \boldsymbol{1}^\top) \\ &= \Big(\prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0} - \theta_{\boldsymbol{e}_h,\boldsymbol{1}}), *, \dots, *, \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{1}}), 0, \dots, 0 \Big), \end{aligned}$$

and satisfies that

$$T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \boldsymbol{1}^\top) \neq 0; \quad T_{\boldsymbol{r}^*,\boldsymbol{\alpha}'}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \boldsymbol{1}^\top) = 0, \quad \forall \boldsymbol{\alpha}' \not\preceq_{S_1} \boldsymbol{\alpha}.$$

From previous constructions we have

$$\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2}) = (\mathbf{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0})^\top,$$

and denote the value in column $\boldsymbol{\alpha}$ of $\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_2})$ by $b_{\boldsymbol{v},\boldsymbol{\alpha}}$. Consider any $j \in S_1 \cup S_2$ such that $\Gamma_{j,\boldsymbol{\alpha}} = 1$, then obviously $\boldsymbol{e}_j$ is not included in the sum in the previously defined response pattern $\boldsymbol{r}^*$, because $\boldsymbol{r}^*$ only contains those items that $\boldsymbol{\alpha}$ is not capable of. So we have

$$T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\}$$
$$= \Big(\mathbf{0}^\top, \underbrace{b_{\boldsymbol{v},\boldsymbol{\alpha}} \cdot \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}_k}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}), \mathbf{0}^\top}_{\text{column } \boldsymbol{\alpha}}\Big), \tag{B.36}$$

$$T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\}$$
$$= \Big(\mathbf{0}^\top, \underbrace{\theta_{\boldsymbol{e}_j,\mathbf{1}} \cdot b_{\boldsymbol{v},\boldsymbol{\alpha}} \cdot \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}), \mathbf{0}^\top}_{\text{column } \boldsymbol{\alpha}}\Big). \tag{B.37}$$

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

$$T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\}$$
$$= \Big(\mathbf{0}^\top, \underbrace{\prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}), \mathbf{0}^\top}_{\text{column } \boldsymbol{\alpha}}\Big), \tag{B.38}$$

$$T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\}$$
$$= \Big(\mathbf{0}^\top, \underbrace{\bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}} \cdot \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}), \mathbf{0}^\top}_{\text{column } \boldsymbol{\alpha}}\Big). \tag{B.39}$$

Equation (C.1) implies (D.8) $\cdot\, \boldsymbol{p} =$ (D.10) $\cdot\, \bar{\boldsymbol{p}}$, and since (D.10) $\cdot\, \bar{\boldsymbol{p}} \neq 0$, we must also have (D.8) $\cdot\, \bar{\boldsymbol{p}} \neq 0$, which indicates $b_{\boldsymbol{v},\boldsymbol{\alpha}} \neq 0$. The above four equations along with

(C.1) give that

$$\theta_{\boldsymbol{e}_j,\mathbf{1}} = \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \frac{(\text{D.9})\cdot \boldsymbol{p}}{(\text{D.8})\cdot \boldsymbol{p}} = \frac{(\text{D.11})\cdot \bar{\boldsymbol{p}}}{(\text{D.10})\cdot \bar{\boldsymbol{p}}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}, \quad \forall j \in S_2.$$

Note that the above equality $\theta_{\boldsymbol{e}_j,\mathbf{1}} = \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}$ holds for any $\boldsymbol{\alpha}$ and any item j such that $\Gamma_{j,\boldsymbol{\alpha}} = 1$. Therefore we have shown $\theta_{\boldsymbol{e}_j,\mathbf{1}} = \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}$ holds for any $j \in S_1 \cup S_2$. Similarly we also have $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}$. In summary,

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}, \quad \theta_{\boldsymbol{e}_j,\mathbf{1}} = \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}, \quad \forall j \in S_1 \cup S_2.$$

For $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ define

$$\boldsymbol{\theta}^* = \sum_{h \in S_1} \theta_{\boldsymbol{e}_h,\mathbf{1}} \boldsymbol{e}_h,$$

then $T_{\sum_{h\in S_1} \boldsymbol{e}_h}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)\boldsymbol{p} = T_{\sum_{h\in S_1} \boldsymbol{e}_h}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)\bar{\boldsymbol{p}}$ gives

$$\prod_{h\in S_1} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) p_{\boldsymbol{\alpha}_0} = \prod_{h\in S_1} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \bar{p}_{\boldsymbol{\alpha}_0},$$

so $p_{\boldsymbol{\alpha}_0} = \bar{p}_{\boldsymbol{\alpha}_0}$.

Next we show $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ *for any* $\boldsymbol{\alpha}$ *and* $j \in S_1 \cup S_2$, *where* $\Gamma_{j,\boldsymbol{\alpha}} = 0$. We use the induction method to show that for any $\boldsymbol{\alpha} \in \mathcal{C}$,

$$\forall j \in S_1 \cup S_2, \quad \theta_{j,\boldsymbol{\alpha}} = \bar{\theta}_{j,\boldsymbol{\alpha}}, \quad p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}. \tag{B.40}$$

Firstly, we prove (D.12) hold for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1$, where $\boldsymbol{\alpha}_1$ denotes the latent class with the smallest lexicographical order among $\mathcal{C} \setminus \{\boldsymbol{\alpha}_0\}$. For $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1$, define

$$\boldsymbol{\theta}^* = \sum_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=0} \theta_{\boldsymbol{e}_h,\mathbf{1}} \boldsymbol{e}_h + \sum_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=1} \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0} \boldsymbol{e}_h, \tag{B.41}$$

then the row vectors of $\boldsymbol{r}^* = \sum_{h\in S_1} \boldsymbol{e}_h$ in the transformed T-matrices only contain one nonzero element corresponding to column $\boldsymbol{\alpha}_1$ as follows

$$
\begin{aligned}
&T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top) \\
=&\Big(\mathbf{0}^\top, \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),\mathbf{0}^\top\Big),
\end{aligned} \tag{B.42}
$$

$$
\begin{aligned}
&T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top) \\
=&\Big(\mathbf{0}^\top, \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),\mathbf{0}^\top\Big),
\end{aligned} \tag{B.43}
$$

and this is because for any other latent class $\boldsymbol{\alpha}'\neq\boldsymbol{\alpha}_1$, the $\boldsymbol{\alpha}'$ is capable of at least one item in S_1 that $\boldsymbol{\alpha}_1$ is not capable of. Now consider the row vector corresponding to response pattern $\boldsymbol{r}+\boldsymbol{e}_j$ for $j\in S_2$ in the transformed T-matrices, and we have

$$
\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top) \\
=&\Big(\mathbf{0}^\top, \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}}\cdot\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),\mathbf{0}^\top\Big),
\end{aligned}
$$

and

$$
\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top) \\
=&\Big(\mathbf{0}^\top, \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1}\cdot\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}_1}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}_1}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),\mathbf{0}^\top\Big).
\end{aligned}
$$

The above four equations along with Equation (C.1) indicate for $j\in S_2$ we have

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1}.$$

Similarly for $j\in S_1$ we also have $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1}$. Plugging $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1}$ into the equation (B.42)$\boldsymbol{p}$ = (B.43)$\bar{\boldsymbol{p}}$ gives

$$p_{\boldsymbol{\alpha}_1} = \bar{p}_{\boldsymbol{\alpha}_1}.$$

So now we have shown (D.12) holds for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1$.

Then as the induction assumption, suppose for any given $\boldsymbol{\alpha} \in \mathcal{C}$, we have

$$\forall \boldsymbol{\alpha}' \text{ s.t. } \boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}, \quad \forall j \in S_1 \cup S_2, \quad \theta_{\boldsymbol{e}_j, \boldsymbol{\alpha}'} = \bar{\theta}_{\boldsymbol{e}_j, \boldsymbol{\alpha}'}, \quad p_{\boldsymbol{\alpha}'} = \bar{p}_{\boldsymbol{\alpha}'}.$$

Recall that $\boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}$ if and only if $\boldsymbol{\alpha}' \preceq_{S_2} \boldsymbol{\alpha}$. Define $\boldsymbol{\theta}^*$ as

$$\boldsymbol{\theta}^* = \sum_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 0} \theta_{\boldsymbol{e}_h, \mathbf{1}} \boldsymbol{e}_h + \sum_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 1} \theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}_0} \boldsymbol{e}_h,$$

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

$$\begin{aligned} T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)\boldsymbol{p} = & \sum_{\boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}} t_{\boldsymbol{r}^*, \boldsymbol{\alpha}'} \cdot p_{\boldsymbol{\alpha}'} \\ & + \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 0} (\theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h, \mathbf{1}}) \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 1} (\theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}_0}) \cdot p_{\boldsymbol{\alpha}}, \end{aligned} \tag{B.44}$$

$$\begin{aligned} T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)\bar{\boldsymbol{p}} = & \sum_{\boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}} \bar{t}_{\boldsymbol{r}^*, \boldsymbol{\alpha}'} \cdot \bar{p}_{\boldsymbol{\alpha}'} \\ & + \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 0} (\bar{\theta}_{\boldsymbol{e}_h, \boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h, \mathbf{1}}) \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}} = 1} (\bar{\theta}_{\boldsymbol{e}_h, \boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h, \boldsymbol{\alpha}_0}) \cdot \bar{p}_{\boldsymbol{\alpha}}, \end{aligned} \tag{B.45}$$

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

$$t_{\boldsymbol{r}^*,\boldsymbol{\alpha}'} = \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}'} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),$$
$$\bar{t}_{\boldsymbol{r}^*,\boldsymbol{\alpha}'} = \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}'} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}).$$

Note that by induction assumption we have $\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}$ for any $\boldsymbol{\alpha}'$ such that $\boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}$. This implies $t_{\boldsymbol{r}^*,\boldsymbol{\alpha}'} = \bar{t}_{\boldsymbol{r}^*,\boldsymbol{\alpha}'}$ and further implies

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

So (D.18) = (D.19) gives

$$\begin{aligned}&\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}) \cdot p_{\boldsymbol{\alpha}}\\ =&\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}) \cdot \bar{p}_{\boldsymbol{\alpha}},\end{aligned} \tag{B.46}$$

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

$$\begin{aligned}&\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}) \cdot p_{\boldsymbol{\alpha}}\\ =&\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}) \cdot \bar{p}_{\boldsymbol{\alpha}}.\end{aligned} \tag{B.47}$$

Taking the ratio of the above two equations (D.21) and (D.20) gives

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}, \quad \forall j \notin S_1.$$

Redefining $\boldsymbol{r}^* := \sum_{h\in S_2}\boldsymbol{e}_h$ similarly as above we have $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ for any $j \in S_1$.

Plug $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ for all $j \in S_1$ into (D.20), then we have $p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}$. Now we have shown (D.12) hold for this particular $\boldsymbol{\alpha}$. Then the induction argument gives

$$\forall \boldsymbol{\alpha} \in \mathcal{C}, \quad \forall j \in S_1 \cup S_2, \quad \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}, \quad p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}.$$

Combined with the results in Step 1 and 2, all the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ are identifiable and the proof of Theorem III.5 is complete. □

Proof of Proposition III.5. Without loss of generality, assume $S_1 = \{1, \dots, M_1\}$ and $S_2 = \{M_1 + 1, \dots, M_1 + M_2\}$. Recall that $\mathcal{B}_{S_1} = \mathcal{B}_{S_2}$ under condition (C3*). The outline of the proof is as follows.

Step 1: $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}$ for $j > M_1 + M_2$.

Step 2: $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ for $j > M_1 + M_2$ and $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$.

Step 3: $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ and $p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}$ for $1 \le j \le M_1 + M_2$, $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ or $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$.

Step 4: $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ and $p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}$ for $1 \le j \le J$ and for all $\boldsymbol{\alpha}$.

Next we start the proof of the theorem.

Step 1. The proof is exactly the same as Step 1 of Theorem III.5.

Step 2. First consider basis latent classes $\boldsymbol{\alpha}$ under both S_1 and S_2. For $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$, define

$$\begin{aligned}\boldsymbol{\theta}^* = &\sum_{j \in S_1 : \Gamma_{j,\boldsymbol{\alpha}} = 1} \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} \boldsymbol{e}_j + \sum_{j \in S_1 : \Gamma_{j,\boldsymbol{\alpha}} = 0} \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}} \boldsymbol{e}_j \\ &+ \sum_{j \in S_2 : \Gamma_{j,\boldsymbol{\alpha}} = 1} \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} \boldsymbol{e}_j + \sum_{j \in S_2 : \Gamma_{j,\boldsymbol{\alpha}} = 0} \theta_{\boldsymbol{e}_j,\mathbf{1}} \boldsymbol{e}_j,\end{aligned}$$

then the row vectors $\boldsymbol{r}^* = \sum_{j=1}^{M_1+M_2} \boldsymbol{e}_j$ in the transformed T-matrices only contain one potentially nonzero element, corresponding to $\boldsymbol{\alpha}$, as follows

$$
\begin{aligned}
&T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) \\
&= \Big(\mathbf{0}^\top, \prod_{j\in S_1:\Gamma_{j,\boldsymbol{\alpha}}=1} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}) \prod_{j\in S_1:\Gamma_{j,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}) \\
&\qquad \times \prod_{j\in S_2:\Gamma_{j,\boldsymbol{\alpha}}=1} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}) \prod_{j\in S_2:\Gamma_{j,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\mathbf{1}}),\ \mathbf{0}^\top\Big),
\end{aligned}
\tag{B.48}
$$

and

$$
\begin{aligned}
&T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) \\
&= \Big(\mathbf{0}^\top, \prod_{j\in S_1:\Gamma_{j,\boldsymbol{\alpha}}=1} (\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}) \prod_{j\in S_1:\Gamma_{j,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}) \\
&\qquad \times \prod_{j\in S_2:\Gamma_{j,\boldsymbol{\alpha}}=1} (\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}) \prod_{j\in S_2:\Gamma_{j,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\mathbf{1}}),\ \mathbf{0}^\top\Big).
\end{aligned}
\tag{B.49}
$$

Lemma B.3 implies the product elements in (B.48) and (B.49) are both nonzero. Then consider any $j > M_1 + M_2$, the row vector corresponding to the response pattern $\boldsymbol{r}^* + \boldsymbol{e}_j$ in the transformed T-matrices are

$$
\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) \\
&= \Big(\mathbf{0}^\top, \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \bar{\theta}_{\boldsymbol{e}_h,\mathbf{1}}) \\
&\qquad \times \prod_{h\in S_2:\Gamma_{h,\boldsymbol{\alpha}}=1} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}) \prod_{h\in S_2:\Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}),\ \mathbf{0}^\top\Big),
\end{aligned}
\tag{B.50}
$$

and

$$
\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\\
=&\Big(\mathbf{0}^\top, \bar\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}}\cdot \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\bar\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\bar\theta_{\boldsymbol{e}_h,\mathbf{1}})\\
&\quad\times\prod_{h\in S_2:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\prod_{h\in S_2:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}}),\ \mathbf{0}^\top\Big).
\end{aligned}
\tag{B.51}
$$

So we have

$$
\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}}=\frac{(\text{B.50})\cdot\boldsymbol{p}}{(\text{B.48})\cdot\boldsymbol{p}}=\frac{(\text{B.51})\cdot\bar{\boldsymbol{p}}}{(\text{B.49})\cdot\bar{\boldsymbol{p}}}=\bar\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}},\quad\forall\boldsymbol{\alpha}\in\mathcal{B}_{S_1},\quad\forall j>M_1+M_2.
$$

Step 3. *We first prove $\theta_{\boldsymbol{e}_j,\mathbf{1}}=\bar\theta_{\boldsymbol{e}_j,\mathbf{1}}$ for any $j\in S_1\cup S_2$.* Given $\boldsymbol{\alpha}\in\mathcal{B}_{S_1}$, define

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

then the row vector corresponding to $\boldsymbol{r}^*=\sum_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}\boldsymbol{e}_h$ in the transformed T-matrix takes the following form

$$
\begin{aligned}
&T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\\
=&\Big(\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_h,\mathbf{1}}),*,\dots,*,\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}}),0,\dots,0\Big).
\end{aligned}
$$

Condition (C4*) implies that $(\theta_{j,\boldsymbol{\alpha}},j\in(S_1\cup S_2)^c)\neq(\theta_{j,\boldsymbol{\alpha}_0},j\in(S_1\cup S_2)^c)$ for any basis latent class $\boldsymbol{\alpha}\in\mathcal{B}_{S_1}$. So there exist a C-dimensional vector $\boldsymbol{m}$ such that the element in $\boldsymbol{m}^\top\cdot T(\boldsymbol{\Theta}_{(M_1+M_2+1):J})$ corresponding to $\boldsymbol{\alpha}_0$ is 0 and the element corresponding to $\boldsymbol{\alpha}$ is 1, i.e.,

$$
\boldsymbol{m}^\top\cdot T(\boldsymbol{\Theta}_{(M_1+M_2+1):J})=(0,*,\dots,*,\underbrace{1}_{\text{column }\boldsymbol{\alpha}},*,\dots,*),
$$

and based on the conclusions of Step 2, we also have

$$\boldsymbol{m}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{(M_1+M_2+1):J}) = (0, *, \ldots, *, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, *, \ldots, *).$$

By Lemma B.1 $T(\bar{\boldsymbol{\Theta}}_{S_2})$ has full column rank C, hence there exists a vector $\boldsymbol{v}$ such that

$$\boldsymbol{v}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2}) = (\mathbf{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0})^\top,$$

and denote the value in column $\boldsymbol{\alpha}$ of $\boldsymbol{v}^\top \cdot T(\boldsymbol{\Theta}_{S_2})$ by $b_{\boldsymbol{v},\boldsymbol{\alpha}}$. Consider any $j \in S_1 \cup S_2$ such that $\Gamma_{j,\boldsymbol{\alpha}} = 1$, then obviously $\boldsymbol{e}_j$ is not included in the sum in the previously defined response pattern $\boldsymbol{r}^*$, because $\boldsymbol{r}^*$ only contains those items that $\boldsymbol{\alpha}$ is not capable of. So we have

$$\begin{aligned}
&T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{m}^\top \cdot T(\boldsymbol{\Theta}_{(M_1+M_2+1):J})\} \odot \{\boldsymbol{v}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\} \\
&= \Big(\mathbf{0}^\top, \underbrace{b_{\boldsymbol{v},\boldsymbol{\alpha}} \cdot \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}_k}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}})}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}^\top\Big),
\end{aligned} \tag{B.52}$$

$$\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{m}^\top \cdot T(\boldsymbol{\Theta}_{(M_1+M_2+1):J})\} \odot \{\boldsymbol{v}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\} \\
&= \Big(\mathbf{0}^\top, \underbrace{\theta_{\boldsymbol{e}_j,\mathbf{1}} \cdot b_{\boldsymbol{v},\boldsymbol{\alpha}} \cdot \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}})}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}^\top\Big).
\end{aligned} \tag{B.53}$$

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

$$\begin{aligned}
&T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{m}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{(M_1+M_2+1):J})\} \odot \{\boldsymbol{v}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\} \\
&= \Big(\mathbf{0}^\top, \underbrace{\prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}})}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}^\top\Big),
\end{aligned} \tag{B.54}$$

$$
\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top)\odot\{\boldsymbol{m}^\top\cdot T(\bar{\boldsymbol{\Theta}}_{(M_1+M_2+1):J})\}\odot\{\boldsymbol{v}^\top\cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\}\\
=&\Big(\mathbf{0}^\top,\underbrace{\bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}\cdot\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})}_{\text{column }\boldsymbol{\alpha}},\mathbf{0}^\top\Big).
\end{aligned}
\tag{B.55}
$$

Equation (C.1) implies (D.8) $\cdot\, \boldsymbol{p} =$ (D.10) $\cdot\, \bar{\boldsymbol{p}}$, and since (D.10) $\cdot\, \bar{\boldsymbol{p}} \neq 0$, we must also have (D.8) $\cdot\, \bar{\boldsymbol{p}} \neq 0$, which indicates $b_{\boldsymbol{v},\boldsymbol{\alpha}} \neq 0$. The above four equations along with (C.1) give that

$$
\theta_{\boldsymbol{e}_j,\mathbf{1}}=\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}}=\frac{\text{(D.9)}\cdot\boldsymbol{p}}{\text{(D.8)}\cdot\boldsymbol{p}}=\frac{\text{(D.11)}\cdot\bar{\boldsymbol{p}}}{\text{(D.10)}\cdot\bar{\boldsymbol{p}}}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}=\bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}},\quad\forall j\in S_2.
$$

Note that the above equality $\theta_{\boldsymbol{e}_j,\mathbf{1}} = \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}$ holds for any $\boldsymbol{\alpha}$ and any item j such that $\Gamma_{j,\boldsymbol{\alpha}} = 1$. Therefore we have shown $\theta_{\boldsymbol{e}_j,\mathbf{1}} = \bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}}$ holds for any $j \in S_1 \cup S_2$. Similarly we also have $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}$. In summary,

$$
\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0},\quad\theta_{\boldsymbol{e}_j,\mathbf{1}}=\bar{\theta}_{\boldsymbol{e}_j,\mathbf{1}},\quad\forall j\in S_1\cup S_2.
$$

For $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ define

$$
\boldsymbol{\theta}^*=\sum_{h\in S_1}\theta_{\boldsymbol{e}_h,\mathbf{1}}\boldsymbol{e}_h,
$$

then $T_{\sum_{h\in S_1}\boldsymbol{e}_h}(\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p} = T_{\sum_{h\in S_1}\boldsymbol{e}_h}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}}$ gives

$$
\prod_{h\in S_1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_h,\mathbf{1}})p_{\boldsymbol{\alpha}_0}=\prod_{h\in S_1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\bar{p}_{\boldsymbol{\alpha}_0},
$$

so we also have $p_{\boldsymbol{\alpha}_0} = \bar{p}_{\boldsymbol{\alpha}_0}$.

Next we show $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ for any $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$ and $j \in S_1 \cup S_2$, where $\Gamma_{j,\boldsymbol{\alpha}} = 0$. Given $\boldsymbol{\alpha}$, define

$$
\boldsymbol{\theta}^*=\sum_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}\theta_{\boldsymbol{e}_h,\mathbf{1}}\boldsymbol{e}_h+\sum_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}\boldsymbol{e}_h,
\tag{B.56}
$$

then the row vectors of $\boldsymbol{r}^* = \sum_{h\in S_1} \boldsymbol{e}_h$ in the transformed T-matrices only contain one nonzero element corresponding to column $\boldsymbol{\alpha}$ as follows

$$T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(\mathbf{0}^\top, \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}), \mathbf{0}^\top\Big),$$

$$T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(\mathbf{0}^\top, \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}), \mathbf{0}^\top\Big).$$

Now consider the row vectors of $\boldsymbol{r} + \boldsymbol{e}_j$ for $j \in S_2$ in the transformed T-matrices, we have

$$\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)\\
&= \Big(\mathbf{0}^\top,\ \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),\ \mathbf{0}^\top\Big),
\end{aligned}$$

and

$$\begin{aligned}
&T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top)\\
&= \Big(\mathbf{0}^\top,\ \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\mathbf{1}}) \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),\ \mathbf{0}^\top\Big).
\end{aligned}$$

The above four equations along with Equation (C.1) indicate for $j \in S_2$ we have

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}.$$

Similarly for $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$, $j \in S_1$ we also have $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$. In summary, we have

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}, \quad \forall \boldsymbol{\alpha} \in \mathcal{B}_{S_1}, \quad \forall j \in S_1 \cup S_2.$$

Now for $\boldsymbol{\alpha} \in \mathcal{B}_{S_1}$ define

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

then $T_{\sum_{h\in S_1}\boldsymbol{e}_h}(\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p} = T_{\sum_{h\in S_1}\boldsymbol{e}_h}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}}$ gives

$$\begin{aligned}
&\prod_{h\in S_1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_h,\mathbf{1}})p_{\boldsymbol{\alpha}_0}+\prod_{h\in S_1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})p_{\boldsymbol{\alpha}}\\
=&\prod_{h\in S_1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}-\theta_{\boldsymbol{e}_h,\mathbf{1}})p_{\boldsymbol{\alpha}_0}+\prod_{h\in S_1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\bar{p}_{\boldsymbol{\alpha}},
\end{aligned}$$

which implies $p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}$. This completes the proof of Step 3.

Step 4. We use the induction method to prove the conclusions for those $\boldsymbol{\alpha}\notin\mathcal{B}_{S_1}$. In previous steps we already established

$$p_{\boldsymbol{\alpha}_0}=\bar{p}_{\boldsymbol{\alpha}_0},\quad \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0},\quad \forall j\in\{1,\ldots,J\},$$

and

$$p_{\boldsymbol{\alpha}}=\bar{p}_{\boldsymbol{\alpha}},\quad \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}},\quad \forall\boldsymbol{\alpha}\in\mathcal{B}_{S_1},\quad \forall j\in\{1,\ldots,J\}.$$

So as the induction assumption, suppose for any given $\boldsymbol{\alpha}\notin\mathcal{B}_{S_1}$, we have

$$p_{\boldsymbol{\alpha}'}=\bar{p}_{\boldsymbol{\alpha}'},\quad \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}'}=\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}'},\quad \forall\boldsymbol{\alpha}'\text{ s.t. }\boldsymbol{\alpha}'\preceq_{S_1}\boldsymbol{\alpha}\quad \forall j\in\{1,\ldots,J\}.$$

Recall that $\boldsymbol{\alpha}'\preceq_{S_1}\boldsymbol{\alpha}$ if and only if $\boldsymbol{\alpha}'\preceq_{S_2}\boldsymbol{\alpha}$. Define $\boldsymbol{\theta}^*$ as that in (B.56)

$$\boldsymbol{\theta}^*=\sum_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}\theta_{\boldsymbol{e}_h,\mathbf{1}}\boldsymbol{e}_h+\sum_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}\boldsymbol{e}_h,$$

then the row vector corresponding to $\boldsymbol{r}^*=\sum_{h\in S_1}\boldsymbol{e}_h$ in the transformed T-matrix takes the form

$$\begin{aligned}
T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p}=&\sum_{\boldsymbol{\alpha}'\preceq_{S_1}\boldsymbol{\alpha}}t_{\boldsymbol{r}^*,\boldsymbol{\alpha}'}\cdot p_{\boldsymbol{\alpha}'} \qquad\qquad \text{(B.57)}\\
&+\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\cdot p_{\boldsymbol{\alpha}},
\end{aligned}$$

$$T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}} = \sum_{\boldsymbol{\alpha}'\preceq_{S_1}\boldsymbol{\alpha}} \bar{t}_{\boldsymbol{r}^*,\boldsymbol{\alpha}'}\cdot\bar{p}_{\boldsymbol{\alpha}'} + \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\cdot\bar{p}_{\boldsymbol{\alpha}}, \tag{B.58}$$

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

$$t_{\boldsymbol{r}^*,\boldsymbol{\alpha}'} = \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}'}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}),$$
$$\bar{t}_{\boldsymbol{r}^*,\boldsymbol{\alpha}'} = \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}'}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0}).$$

Note that by induction assumption we have $\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}$ for any $\boldsymbol{\alpha}'$ such that $\boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}$. This implies $t_{\boldsymbol{r}^*,\boldsymbol{\alpha}'} = \bar{t}_{\boldsymbol{r}^*,\boldsymbol{\alpha}'}$ and further implies

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

So (E.12) = (B.58) gives

$$\begin{aligned}&\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\cdot p_{\boldsymbol{\alpha}}\\ =&\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\cdot\bar{p}_{\boldsymbol{\alpha}}.\end{aligned} \tag{B.59}$$

Consider any $j\notin S_1$ and similarly we have

$$\begin{aligned}&\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}}\cdot\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\cdot p_{\boldsymbol{\alpha}}\\ =&\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}\cdot\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\mathbf{1}})\prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=1}(\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}}-\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}_0})\cdot\bar{p}_{\boldsymbol{\alpha}}.\end{aligned} \tag{B.60}$$

Taking the ratio of the above two equations gives

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}, \quad \forall j\notin S_1.$$

Similarly we have $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ for any $j \in S_1$. Plug in $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}}$ for all $j \in S_1$ into (B.59), then we have

$$p_{\boldsymbol{\alpha}} = \bar{p}_{\boldsymbol{\alpha}}.$$

This completes the proof of Proposition III.5. $\square$

Proof of Theorem III.6. Without loss of generality, we show the generic identifiability statement holds on the parameter space $\mathcal{T}$

$$\mathcal{T} = \Big\{(\boldsymbol{\Theta}, \boldsymbol{p}) : \forall\ j, \max_{\boldsymbol{\alpha}:\Gamma_{j,\boldsymbol{\alpha}}=1} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\Gamma_{j,\boldsymbol{\alpha}}=1} \theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'} \geq \theta_{j,\boldsymbol{\alpha}_0},\ \forall\ \Gamma_{j,\boldsymbol{\alpha}'} = 0\Big\}.$$

On $\mathcal{T}$, altering some entries of zero to one in the Γ-matrix is equivalently imposing more affine constraints on the parameters and force them to be in a subset $\mathcal{T}^*$ of $\mathcal{T}$. Since Condition (C3) holds for model parameters belonging to the space $\mathcal{T}^*$, the proof of Theorem III.5 gives that the matrix $T(\boldsymbol{\Theta}_{S_i})$ has full column rank C for $i = 1, 2$ for $(\boldsymbol{\Theta}_{S_i}, \boldsymbol{p}) \in \mathcal{T}^*$. Note that saying the $2^{|S_i|} \times C$ matrix $T(\boldsymbol{\Theta}_{S_i})$ has full column rank is equivalently saying the map sending $T(\boldsymbol{\Theta}_{S_i})$ to all its $\binom{2^{|S_i|}}{C}$ possible $C \times C$ minors $A_1^i, A_2^i, \ldots, A_{2^{|S_i|}}^i$ yields at least one nonzero minor, where $A_1^i, A_2^i, \ldots, A_{2^{|S_i|}}^i$ are all polynomials of the item parameters $\boldsymbol{\Theta}_{S_i}$. Define

$$\mathcal{V} = \bigcup_{i=1,2} \Big\{ \bigcap_{l=1}^{2^{|S_i|}} \{(\boldsymbol{\Theta}, \boldsymbol{p}) \in \mathcal{T} : A_l^i(\boldsymbol{\Theta}_{S_i}) = 0\} \Big\},$$

then $\mathcal{V}$ is a algebraic variety defined by polynomials of the model parameters. Moreover, $\mathcal{V}$ is a proper subvariety of $\mathcal{T}$, since the fact $T(\boldsymbol{\Theta}_{S_i})$ has full column rank C for $i = 1, 2$ for one particular set of $(\boldsymbol{\Theta}, \boldsymbol{p}) \in \mathcal{T}^*$ ensures that there exists one particular set of model parameters that give nonzero values when plugged into the polynomials defining $\mathcal{V}$, which indicates that the polynomials defining $\mathcal{V}$ are not all zero polynomials.

This implies for generic choices of $(\mathbf{\Theta}, \boldsymbol{p})$ in the space $\mathcal{T}$, $T(\mathbf{\Theta}_{S_i})$ has full column rank for $i = 1, 2$. Together with the assumption that (C4) holds for Γ, we obtain generic identifiability of the model parameters. This completes the proof of Theorem III.6. $\square$

Proof of Theorem III.7. In the following, we say some statement "generically" holds, if the subset of the parameter space where the statement does not hold is of Lebesgue measure zero. Without loss of generality, assume Q takes the form

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q' \end{pmatrix},$$

where under the assumptions of Theorem III.7, Q_1 and Q_2 are $K \times K$ square matrices with diagonal elements all equal to 1. With a slight abuse of notation, for a $J_i \times K$ submatrix Q_i of Q, let $T(Q_i, \mathbf{\Theta}_{Q_i})$ denote the $2^{J_i} \times 2^K$ T-matrix. We consider the saturated model where all the main effect and interaction effect terms are included in modeling the item parameters, namely the positive response probability for attribute profile $\boldsymbol{\alpha}$ and item j takes the form

$$\begin{aligned} \theta_{j,\boldsymbol{\alpha}} =& f\Big(\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k} q_{jk}\alpha_k + \sum_{k'=k+1}^{K} \sum_{k=1}^{K-1} \beta_{j,kk'}(q_{jk}\alpha_k)(q_{jk'}\alpha_{k'}) \\ &+ \cdots + \beta_{j,12\cdots K} \prod_k (q_{jk}\alpha_k)\Big), \end{aligned} \tag{B.61}$$

where $f(\cdot)$ is the link function, which can be the identify link, log link, or the logistic link. Note that taking those β-coefficients of the interaction terms to be zero, one is left with a main-effect model. Since the following arguments only rely on the main effect coefficients, the conclusion of the theorem applies to any multi-parameter restricted latent class model.

First prove that under condition (C5), the T-matrices $T(Q_1, \mathbf{\Theta}_{Q_1})$ and $T(Q_2, \mathbf{\Theta}_{Q_2})$ corresponding to Q_1 and Q_2 are both generically of full rank 2^K. To show generic identifiability, it suffices to find one specific set of item parameters $\mathbf{\Theta}$ satisfying the constraints imposed by the Q-matrix that make the T-matrices $T(Q_1, \mathbf{\Theta}_{Q_1})$ and $T(Q_2, \mathbf{\Theta}_{Q_2})$ have full rank. In the following we focus on $T(Q_1, \mathbf{\Theta}_{Q_1})$ only. For $k = 1, \ldots, K$, set the k'th main effect parameter of the k'th item to be 1, i.e., set $\beta_{k,k} = 1$, and all the other main effect and interaction effect parameters to be zero, then the T-matrix $T(Q_1, \mathbf{\Theta}_{Q_1})$ now becomes exactly the same as the T-matrix $T(\mathcal{I}_K, \widetilde{\mathbf{\Theta}}_{\mathcal{I}_K})$ under the identity Q-matrix $\mathcal{I}_K$ with the item parameters being

$$\tilde{\theta}_{\boldsymbol{e}_k, \mathbf{0}} = \beta_{k,0} \quad \text{and} \quad \tilde{\theta}_{\boldsymbol{e}_k, \boldsymbol{e}_k} = \tilde{\theta}_{\boldsymbol{e}_k, \mathbf{1}} = \beta_{k,0} + \beta_{k,k} \text{ for } k \in \{1, \ldots, K\}.$$

Moreover, defining $\tilde{\boldsymbol{\theta}}^* = (\tilde{\theta}_{\boldsymbol{e}_1, \mathbf{1}}, \ldots, \tilde{\theta}_{\boldsymbol{e}_K, \mathbf{1}})^T$ and following a similar argument as in the proof of Lemma B.1, we have that $T(\mathcal{I}_K, \widetilde{\mathbf{\Theta}} - \tilde{\boldsymbol{\theta}}^* \mathbf{1}^\top)$ takes an botomn-left triangular form with nonzero diagonal entries, thus Proposition B.2 gives that $T(\mathcal{I}_K, \widetilde{\mathbf{\Theta}}_{\mathcal{I}_K})$ is full-rank. Therefore $T(Q_1, \mathbf{\Theta}_{Q_1})$ is generically full-rank. Similarly $T(Q_2, \mathbf{\Theta}_{Q_2})$ is also generically full-rank.

We next show that if condition (C6) additionally holds, then any two different columns indexed by attribute profiles $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ of $T(Q', \mathbf{\Theta}_{Q'})$ are generically distinct. For distinct $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}' \in \{0,1\}^K$, they at least differ in one attribute k. Without loss of generality, assume $\alpha_k = 1 > 0 = \alpha'_k$. Condition (C6) ensures that there exists some item $j > 2K$ such that $q_{j,k} = 1$. Under the model considered here with $\theta_{j,\boldsymbol{\alpha}}$ in the form of (B.61), this implies $\theta_{j,\boldsymbol{\alpha}} \neq \theta_{j,\boldsymbol{\alpha}'}$ generically.

Next we introduce a result of uniqueness of three-way tensor decomposition to facilitate our proof. Following Kruskal (1977), the Kruskal rank of a matrix is the the largest number I such that every I columns of the matrix are independent. For a matrix M, let $\text{rank}_K(M)$ denote its Kruskal rank. From Kruskal (1977), Rhodes

(2010) has the following result.

Lemma B.5 (Rhodes, 2010)**.** *For a matrix M_i, denote the jth column of it by $\boldsymbol{m}_j^i$. Given matrices M_i of size $s_i \times c$, let the matrix triple product $[M_1, M_2, M_3]$ be an $s_1 \times s_2 \times s_3$ tensor defined as a sum of r rank-1 tensors by*

$$[M_1, M_2, M_3] = \sum_{j=1}^{c} \boldsymbol{m}_j^1 \otimes \boldsymbol{m}_j^2 \otimes \boldsymbol{m}_j^3.$$

Suppose $\text{rank}_K(M_1) = \text{rank}_K(M_2) = c$ and $\text{rank}_K(M_3) \geq 2$; N_1, N_2, N_3 are matrices with c columns and $[M_1, M_2, M_3] = [N_1, N_2, N_3]$. Then there exists some permutation matrix P and invertible diagonal matrices D_i with $D_1 D_2 D_3 = \mathcal{I}_c$ such that $N_i = M_i D_i P$.

Now we consider three T-matrices, $T(Q_1, \boldsymbol{\Theta}_{Q_1})$, $T(Q_2, \boldsymbol{\Theta}_{Q_2})$ and $T(Q', \boldsymbol{\Theta}_{Q'})$, which are of size $2^K \times 2^K$, $2^K \times 2^K$ and $2^{J-2K} \times 2^K$. The rows of the three matrices are indexed by possible item combinations in the three item sets $\{1, \ldots, K\}$, $\{K + 1, \ldots, 2K\}$ and $\{2K + 1, \ldots, J\}$ respectively. We use $\text{Diag}(\boldsymbol{p})$ to denote a diagonal matrix with the diagonal entries being elements of $\boldsymbol{p}$, then it is not hard to see that $T(\boldsymbol{\Theta})\boldsymbol{p}$ is given by the matrix triple product $[T(Q_1, \boldsymbol{\Theta}_{Q_1}), T(Q_2, \boldsymbol{\Theta}_{Q_2}), T(Q', \boldsymbol{\Theta}_{Q'}) \cdot \text{Diag}(\boldsymbol{p})]$, namely the matrix triple product of the three matrices exactly characterizes the distribution of the response vector $\boldsymbol{R}$. Clearly if a matrix has full column rank, then its Kruskal rank equals its rank, thus our previous arguments already established that $\text{rank}_K\{T(Q_1, \boldsymbol{\Theta}_{Q_1})\} = \text{rank}_K\{T(Q_2, \boldsymbol{\Theta}_{Q_2})\} = 2^K$ and $\text{rank}_K\{T(Q', \boldsymbol{\Theta}_{Q'})\} \geq 2$ hold generically. Moreover, we claim $\text{rank}_K\{T(Q', \boldsymbol{\Theta}_{Q'}) \cdot \text{Diag}(\boldsymbol{p})\} \geq 2$ also holds generically. This is because if all the entries of $\boldsymbol{p}$ are positive, which is a generic requirement, then multiplying the invertible diagonal matrix $\text{Diag}(\boldsymbol{p})$ by the matrix $T(Q', \boldsymbol{\Theta}_{Q'})$ would not change the Kruskcal rank of the latter. Now apply Lemma B.5 and follow a similar argument as the proof of Theorem 4 in Allman et al. (2009), we have the conclusion that the model is generically identifiable up to label swapping.

Specifically, the label swapping would happen only between those latent classes which have identical ideal response vectors, namely the labels of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ could possibly be swapped *only if* $\Gamma_{\bullet,\boldsymbol{\alpha}_1} = \Gamma_{\bullet,\boldsymbol{\alpha}_2}$. This is because otherwise, the constraints (2.1) introduced by the Γ-matrix would fail to hold. $\square$

Proof of Theorem III.8. We first prove the conclusion of the part (a), then that of the part (b).

Proof of (a): Without loss of generality assume the Q-matrix takes the following form

$$Q = \begin{pmatrix} 1 & \boldsymbol{v} \\ \mathbf{0} & Q^* \end{pmatrix}, \tag{B.62}$$

then given any set of valid parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$, one can construct another set of model parameters $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ as follows. First set all the item parameters associated items $j \geq 2$ to be the same as the true parameters for this second set of parameters. For any $\boldsymbol{\alpha}' := \boldsymbol{\alpha}_{2:K} \in \{0,1\}^{K-1}$, choose $\bar{\theta}_{1,(1,\boldsymbol{\alpha}')} \neq \theta_{1,(1,\boldsymbol{\alpha}')}$ to be any reasonable value in a small neighborhood of $\theta_{1,(1,\boldsymbol{\alpha}')}$. Set $\bar{\theta}_{1,(0,\boldsymbol{\alpha}')} = \theta_{1,(0,\boldsymbol{\alpha}')}$ and

$$\begin{cases} \bar{p}_{(0,\boldsymbol{\alpha}')} = p_{(0,\boldsymbol{\alpha}')} + \left(1 - \frac{\theta_{1,(1,\boldsymbol{\alpha}')}}{\bar{\theta}_{1,(1,\boldsymbol{\alpha}')}}\right) p_{(1,\boldsymbol{\alpha}')}; \\ \bar{p}_{(1,\boldsymbol{\alpha}')} = \frac{\theta_{1,(1,\boldsymbol{\alpha}')}}{\bar{\theta}_{1,(1,\boldsymbol{\alpha}')}} p_{(1,\boldsymbol{\alpha}')}, \end{cases}$$

then we have

$$\begin{cases} \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')} = p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')}; \\ \bar{\theta}_{1,(0,\boldsymbol{\alpha}')}\bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,(1,\boldsymbol{\alpha}')}\bar{p}_{(1,\boldsymbol{\alpha}')} = \theta_{1,(0,\boldsymbol{\alpha}')}p_{(0,\boldsymbol{\alpha}')} + \theta_{1,(1,\boldsymbol{\alpha}')}p_{(1,\boldsymbol{\alpha}')}. \end{cases}$$

With these two equations, for any $\boldsymbol{r} \in \{0,1\}^J$

$$
\begin{aligned}
& T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}} \\
= & \sum_{\boldsymbol{\alpha}' \in \{0,1\}^{K-1}} \prod_{j>1: r_j=1} \bar{\theta}^{r_j}_{j,(0,\boldsymbol{\alpha}')} \Big[\bar{\theta}^{r_1}_{1,(0,\boldsymbol{\alpha}')} \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}^{r_1}_{1,(1,\boldsymbol{\alpha}')} \bar{p}_{(1,\boldsymbol{\alpha}')} \Big] \\
= & \sum_{\boldsymbol{\alpha}' \in \{0,1\}^{K-1}} \prod_{j>1: r_j=1} \theta^{r_j}_{j,(0,\boldsymbol{\alpha}')} \cdot \begin{cases} \big[\bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')} \big] & \text{if } r_1 = 0; \\ \big[\bar{\theta}_{1,(0,\boldsymbol{\alpha}')} \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,(1,\boldsymbol{\alpha}')} \bar{p}_{(1,\boldsymbol{\alpha}')} \big] & \text{if } r_1 = 1, \end{cases} \\
= & T_{\boldsymbol{r},\boldsymbol{\cdot}}(\boldsymbol{\Theta})\boldsymbol{p}.
\end{aligned}
$$

This proves the model associated with Q in the form of (B.62) is not generically identifiable, since for any valid set of true parameters there exist another set of parameters resulting in the same distribution of the observed responses $\boldsymbol{R}$.

Proof of (b): Part of the proof idea is similar to that of Theorem III.2. Since the $(J-2)\times(K-1)$ sub-matrix Q' satisfies conditions (C5) and (C6), Theorem III.7 gives that, for generic choice of true parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ in the parameter space, if another set of parameters $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ satisfy $T(\boldsymbol{\Theta})\boldsymbol{p} = T(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}}$, then

$$
\forall j \geq 3, \quad \theta_{j,(0,\boldsymbol{\alpha}_{2:K})} = \bar{\theta}_{j,(0,\boldsymbol{\alpha}_{2:K})}, \quad p_{(0,\boldsymbol{\alpha}_{2:K})} + p_{(1,\boldsymbol{\alpha}_{2:K})} = \bar{p}_{(0,\boldsymbol{\alpha}_{2:K})} + \bar{p}_{(1,\boldsymbol{\alpha}_{2:K})}.
$$

For any response pattern $\boldsymbol{r} = (r_1, r_2, r_3, \ldots, r_j) \in \{0,1\}^J$, (C.22) for $\boldsymbol{r}$ can be equivalently written as

$$
\begin{aligned}
& \sum_{\boldsymbol{\alpha}_{2:K} \in \{0,1\}^{K-1}} \prod_{j>2:\, r_j=1} \theta_{j,(0,\boldsymbol{\alpha}_{2:K})} \cdot \mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K}) \qquad \text{(B.63)} \\
= & \sum_{\boldsymbol{\alpha}_{2:K} \in \{0,1\}^{K-1}} \prod_{j>2:\, r_j=1} \bar{\theta}_{j,(0,\boldsymbol{\alpha}_{2:K})} \cdot \overline{\mathbb{P}}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K}).
\end{aligned}
$$

Note that the difference of (B.63) and (C.32) is that $\mathcal{R}^{Q'}$ is replaced by $\{0,1\}^{K-1}$, which is because when considering generic identifiability of multi-parameter Q-restricted models, all the 2^{K-1} possible proportion parameters resulting from the Q' part are generically identifiable under conditions (C5) and (C6).

Following the same reasoning as in the proof of Theorem III.2, part (B.2), we have

$$\begin{aligned}&\mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K}) \\ &= \overline{\mathbb{P}}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}_{2:K}),\end{aligned} \tag{B.64}$$

and this yields that for any $\boldsymbol{\alpha}' := \boldsymbol{\alpha}_{2:K} \in \{0,1\}^{K-1}$,

$$\begin{cases} p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} = \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}; \\ \theta_{1,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{1,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} = \bar{\theta}_{1,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}; \\ \theta_{2,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{2,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} = \bar{\theta}_{2,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{2,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}; \\ \theta_{1,(0,\boldsymbol{\alpha}')}\theta_{2,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{1,(1,\boldsymbol{\alpha}')}\theta_{2,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} \\ \qquad\qquad = \bar{\theta}_{1,(0,\boldsymbol{\alpha}')}\bar{\theta}_{2,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,(1,\boldsymbol{\alpha}')}\bar{\theta}_{2,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}. \end{cases} \tag{B.65}$$

First we show that if there exist $\boldsymbol{\alpha}_1',\, \boldsymbol{\alpha}_2' \in \{0,1\}^{K-1}$, $\boldsymbol{\alpha}_1' \neq \boldsymbol{\alpha}_2'$ such that

$$\begin{cases} \theta_{j,(\alpha_1,\boldsymbol{\alpha}_1')} = \theta_{j,(\alpha_1,\boldsymbol{\alpha}_2')} \text{ and } \bar{\theta}_{j,(\alpha_1,\boldsymbol{\alpha}_1')} = \bar{\theta}_{j,(\alpha_1,\boldsymbol{\alpha}_2')}, \quad \forall j = 1,2,\ \forall \alpha_1 = 0,1; \\ \frac{p_{(1,\boldsymbol{\alpha}_1')}}{p_{(0,\boldsymbol{\alpha}_1')}} := s_1 \neq s_2 =: \frac{p_{(1,\boldsymbol{\alpha}_2')}}{p_{(0,\boldsymbol{\alpha}_2')}}. \end{cases} \tag{B.66}$$

then one must have

$$\theta_{j,(\alpha_1,\boldsymbol{\alpha}_1')} = \bar{\theta}_{j,(\alpha_1,\boldsymbol{\alpha}_1')}, \quad \forall j = 1,2,\ \forall \alpha_1 = 0,1. \tag{B.67}$$

After some transformations, the system of equations (C.40) yields

$$\begin{cases} (\theta_{1,(0,\boldsymbol{\alpha}')} - \theta_{1,(1,\boldsymbol{\alpha}')}) \cdot (\theta_{2,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')}) \cdot p_{(0,\boldsymbol{\alpha}')} \\ \qquad = (\bar{\theta}_{1,(0,\boldsymbol{\alpha}')} - \theta_{1,(1,\boldsymbol{\alpha}')}) \cdot (\bar{\theta}_{2,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')}) \cdot \bar{p}_{(0,\boldsymbol{\alpha}')}; \\ (\theta_{2,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')}) \cdot p_{(0,\boldsymbol{\alpha}')} + (\theta_{2,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')}) \cdot \bar{p}_{(1,\boldsymbol{\alpha}')} \\ \qquad = (\bar{\theta}_{2,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')}) \cdot \bar{p}_{(0,\boldsymbol{\alpha}')}. \end{cases}$$

Under a multi-parameter model, $q_{1,1} = q_{2,1} = 1$ yields that for generic parameters, $\theta_{i,(0,\boldsymbol{\alpha}')} \neq \bar{\theta}_{i,(1,\boldsymbol{\alpha}')}$, $i = 1, 2$, so the left (right) hand side of the first equation above is nonzero. And obviously the right hand side of the second equation above is nonzero. Taking the ratio of the above two equations gives

$$\begin{aligned} &\frac{(\theta_{1,(0,\boldsymbol{\alpha}')} - \theta_{1,(1,\boldsymbol{\alpha}')}) \cdot (\theta_{2,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')})}{(\theta_{2,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')}) + (\theta_{2,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}')}) \cdot p_{(1,\boldsymbol{\alpha}')}/p_{(0,\boldsymbol{\alpha}')}} \\ &= (\bar{\theta}_{1,(0,\boldsymbol{\alpha}')} - \theta_{1,(1,\boldsymbol{\alpha}')}) := f(\boldsymbol{\alpha}'). \end{aligned}$$

The right hand side of the above equation does not involve any proportion parameter $\boldsymbol{p}$ or $\bar{\boldsymbol{p}}$. So for $\boldsymbol{\alpha}'_1$, $\boldsymbol{\alpha}'_2$ satisfying (C.35), $f(\boldsymbol{\alpha}'\boldsymbol{\alpha}'_1) = f(\boldsymbol{\alpha}'_2)$. Note that the left hand side of the above equation involves a ratio $p_{(1,\boldsymbol{\alpha}')}/p_{(0,\boldsymbol{\alpha}')}$ depending on $\boldsymbol{\alpha}'$. Equality $f(\boldsymbol{\alpha}'_1) = f(\boldsymbol{\alpha}'_2)$ along with (C.35) imply

$$\begin{aligned} (\theta_{2,(1,\boldsymbol{\alpha}'_1)} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}'_1)}) \cdot \frac{p_{(1,\boldsymbol{\alpha}'_1)}}{p_{(0,\boldsymbol{\alpha}'_1)}} \;&=\; (\theta_{2,(1,\boldsymbol{\alpha}'_2)} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}'_2)}) \cdot \frac{p_{(1,\boldsymbol{\alpha}'_2)}}{p_{(0,\boldsymbol{\alpha}'_2)}} \\ = (\theta_{2,(1,\boldsymbol{\alpha}'_1)} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}'_1)}) \cdot \frac{p_{(1,\boldsymbol{\alpha}'_2)}}{p_{(0,\boldsymbol{\alpha}'_2)}}, \end{aligned}$$

and

$$(\theta_{2,(1,\boldsymbol{\alpha}'_1)} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}'_1)}) \cdot \left(\frac{p_{(1,\boldsymbol{\alpha}'_1)}}{p_{(0,\boldsymbol{\alpha}'_1)}} - \frac{p_{(1,\boldsymbol{\alpha}'_2)}}{p_{(0,\boldsymbol{\alpha}'_2)}} \right) = 0,$$

then since $p_{(1,\boldsymbol{\alpha}_1')}/p_{(0,\boldsymbol{\alpha}_1')} = s_1 \neq s_2 = p_{(1,\boldsymbol{\alpha}_2')}/p_{(0,\boldsymbol{\alpha}_2')}$, by assumption (C.35), we have

$$\theta_{2,(1,\boldsymbol{\alpha}_1')} - \bar{\theta}_{2,(1,\boldsymbol{\alpha}_1')} = 0.$$

By symmetry of the four item parameters $\theta_{1,(0,\boldsymbol{\alpha}')}$, $\theta_{1,(1,\boldsymbol{\alpha}')}$, $\theta_{2,(0,\boldsymbol{\alpha}')}$ and $\theta_{2,(1,\boldsymbol{\alpha}')}$ in (C.40), equalities (C.36) therefore hold following a similar argument.

Next we show that under the condition of the theorem, the conclusions obtained so far give the generic identifiability of all the item parameters associated with the first two items, and hence proved the generic identifiability of all the model parameters. Since $\boldsymbol{v}_1 \vee \boldsymbol{v}_2 \neq \mathbf{1}$, there must exist some attribute k, $k \neq 1$, that is not required by the first two items. Then for any item parameter $\theta_{j,\boldsymbol{\alpha}}$ corresponding to item j, $j = 1, 2$ and attribute profile $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, define $\boldsymbol{\alpha}_1' = (\alpha_2, \ldots, \alpha_K)$, and $\boldsymbol{\alpha}_2' = (\alpha_2', \ldots, \alpha_K')$, $\alpha_l' = \alpha_l$ for any $l \neq k$ and $\alpha_k' = 1 - \alpha_k$, then

$$\theta_{j,(\alpha_1,\boldsymbol{\alpha}_1)} = \theta_{j,(\alpha_1,\boldsymbol{\alpha}_2')} \text{ and } \bar{\theta}_{j,(\alpha_1,\boldsymbol{\alpha}_1')} = \bar{\theta}_{j,(\alpha_1,\boldsymbol{\alpha}_2')}, \quad \forall j = 1, 2, \ \forall \alpha_1 = 0, 1.$$

This means we have found $\boldsymbol{\alpha}_1' \neq \boldsymbol{\alpha}_2'$ that satisfy the first equation in (C.35), then as long as $p_{(1,\boldsymbol{\alpha}_1')}/p_{(0,\boldsymbol{\alpha}_1')} \neq p_{(1,\boldsymbol{\alpha}_2')}/p_{(0,\boldsymbol{\alpha}_2')}$ then $\theta_{j,\boldsymbol{\alpha}} = \bar{\theta}_{j,\boldsymbol{\alpha}}$ follows for $j = 1, 2$. Since this inequality constraint of the true parameters is a generic constraint, i.e. the parameters not satisfying this constraint falls in a Lebesgue measure zero set of the parameter space, the generic identifiability of all the item parameters holds. Considering the fact $\theta_{j,(0,\boldsymbol{\alpha}')} \neq \theta_{j,(1,\boldsymbol{\alpha}')}$ generically, identifiability of the item parameters combined with (C.40) further gives the generic identifiability of the proportion parameters $\boldsymbol{p}$. This completes the proof of part (b). $\square$

Proof of Proposition III.4 and Proposition III.6. To prove Proposition III.6, suppose the identifiability conditions for $\boldsymbol{p}$-partial identifiability are satisfied. We introduce a

2^J-dimensional empirical response vector

$$\boldsymbol{\gamma} = \Big\{1,\ N^{-1}\sum_{i=1}^N I(\boldsymbol{R}_i \succeq \boldsymbol{e}_1),\ \cdots,\ N^{-1}\sum_{i=1}^N I(\boldsymbol{R}_i \succeq \boldsymbol{e}_J),$$
$$N^{-1}\sum_{i=1}^N I(\boldsymbol{R}_i \succeq \boldsymbol{e}_1+\boldsymbol{e}_2),\ \cdots,\ N^{-1}\sum_{i=1}^N I(\boldsymbol{R}_i \succeq \sum_{j=1}^J \boldsymbol{e}_j)\Big\}^\top,$$

where elements of $\boldsymbol{\gamma}$ are indexed by all the $|\{0,1\}^J| = 2^J$ possible response patterns and they are in the same order as that of the columns of the T-matrix. First, by the definition of the T-matrix and the strong law of large numbers, we have $\boldsymbol{\gamma} \to T(\boldsymbol{\Theta}^0)\boldsymbol{p}^0$ almost surely as $N \to \infty$. Second, the maximum likelihood estimators $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{p}}$ satisfy $\|\boldsymbol{\gamma} - T(\widehat{\boldsymbol{\Theta}})\widehat{\boldsymbol{\nu}}\| \to 0$, where $\|\cdot\|$ denotes the L_2 norm. Therefore, combining these two gives

$$\|T(\boldsymbol{\Theta}^0)\boldsymbol{\nu}^0 - T(\widehat{\boldsymbol{\Theta}})\widehat{\boldsymbol{\nu}}\| \to 0$$

almost surely as $N \to \infty$. Then since the identifiability conditions are satisfied, we have that $T(\boldsymbol{\Theta}^0)\boldsymbol{\nu}^0 = T(\widehat{\boldsymbol{\Theta}})\widehat{\boldsymbol{\nu}}$ indicates $(\boldsymbol{\Theta}^0, \boldsymbol{\nu}^0) = (\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\nu}})$. Therefore we obtain the consistency result that $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{\nu}}) \to (\boldsymbol{\Theta}^0, \boldsymbol{\nu}^0)$ almost surely as $N \to \infty$. This proves Proposition III.4.

To prove Proposition III.6, suppose the identifiability conditions for generic identifiability are satisfied. Then according to Definition IV.2 of generic identifiability, there exists a proper algebraic subvariety $\mathcal{V}$ of $\mathcal{T}$, such that $(\boldsymbol{\Theta}, \boldsymbol{p})$ are strictly identifiable on $\mathcal{T} \setminus \mathcal{V}$, and subvariety $\mathcal{V}$ has Lebesgue measure zero in the parameter space. If the true parameters $(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ belong to $\mathcal{T} \setminus \mathcal{V}$, then for any other valid set of parameters $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$, the equalities $T(\boldsymbol{\Theta}^0)\boldsymbol{p}^0 = T(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}}$ indicate $(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) = (\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$. Similarly to the proof of Proposition III.4 in the last paragraph, we have

$$\|T(\boldsymbol{\Theta}^0)\boldsymbol{p}^0 - T(\widehat{\boldsymbol{\Theta}})\widehat{\boldsymbol{p}}\| \to 0$$

almost surely as $N \to \infty$. And the identifiability of $(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) \in \mathcal{T} \setminus \mathcal{V}$ guarantees that $(\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}}) \to (\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ almost surely as $N \to \infty$. This proves Proposition III.6. $\square$

B.4 Proof of Results in Section 3.4

Proof of Corollary III.2. If the Γ-matrix constructed as in the corollary is separable and contains distinct columns, then each attribute pattern $\boldsymbol{\alpha} \in \mathcal{A}$ corresponds to a unique equivalence class and $\boldsymbol{\nu} = \boldsymbol{p}$, where $\boldsymbol{\nu}$ represents the grouped proportion parameters of Γ-matrix-induced equivalence classes introduced in Section 3.2. Further, the general constraints (3.2) are satisfied for each item j. Note that the proof of Theorem 1 only use the information that each item j has two levels of item parameters θ_j^+, θ_j^- which satisfy (3.2), and that proof does not depend on whether each item is specified as conjunctive (DINA) or disjunctive (DINO). Therefore Theorem 1 can be directly applied here. Given that Γ is separable, conditions (C1) and (C2) lead to strict identifiability of the model parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$. This concludes the proof. $\square$

Proof of Corollary III.3 (a). To prove part (a), we first point out that the Γ-matrix defined in part (a) ensures the model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ satisfy the general constraints (3.2) for each item $j \in \mathcal{S}$. The constraint set $\mathcal{C}_j$ is just defined as $\mathcal{C}_j = \{\boldsymbol{\alpha} \in \mathcal{A} : \Gamma_{j,\boldsymbol{\alpha}} = 1\}$. Then because the proofs of Theorem III.5 and Proposition III.5 do not depend on the specific model assumption of each item, but only use the information that the constraints (3.2) are satisfied for each j, the conclusions of Theorem III.5 and Proposition III.5 still hold in the currently considered scenario. This proves part (a). $\square$

Statement and Proof of Corollary III.3 (b). We first introduce the condition (E2) needed in part (b). For a binary vector $\boldsymbol{a}$, we say another binary vector $\boldsymbol{b}$ of the same length is a unit-shrinkage of $\boldsymbol{a}$, if $\boldsymbol{b} \preceq \boldsymbol{a}$ and $\boldsymbol{b} = \boldsymbol{e}_k$ for some k. Further, for a binary matrix Q, we say another binary matrix $\widetilde{Q}$ of the same size is a unit-shrinkage of Q, if for each j, the jth row vector of $\widetilde{Q}$ is either equal to, or a unit-shrinkage of, the jth row vector of Q. The following condition (E2) ensures the generic identifiability of $(\boldsymbol{\Theta}, \boldsymbol{p})$.

(E2) There exists a decomposition of $Q_{mult} = (Q_{mult,1}^\top, Q_{mult,2}^\top)^\top$ such that the submatrices $Q_{mult,1}$ and $Q_{mult,2}$ satisfy the following conditions.

  (E2.a) There exists a "unit-shrinkage" $\widetilde{Q}_{mult,1}$ of $Q_{mult,1}$ such that the matrix $\widetilde{\Gamma} = (\Gamma^{disj}(Q_{disj}, \mathcal{A})^\top,\ \Gamma^{conj}(Q_{conj}, \mathcal{A})^\top,\ \Gamma^{conj}(\widetilde{Q}_{mult,1}, \mathcal{A})^\top)$ contains two disjoint separable submatrices Γ_1 and Γ_2.

  (E2.b) Each attribute is required by at least one item in $Q_{mult,2}$.

Before proving Corollary III.3 (b), we use an example to illustrate how to check its conditions (E2).

Example B.1. Consider the following Q-matrix with items 1, 4 being two-parameter conjunctive, items 2, 5 being two-parameter disjunctive, and items 3, 6, 7 being multi-parameter.

$$Q = \begin{array}{c} conj\\ disj\\ mult\\ conj\\ disj\\ mult\\ mult \end{array} \left(\begin{array}{cc} 1 & 0\\ 1 & 1\\ 1 & 1\\ \hline 1 & 0\\ 1 & 1\\ 1 & 1\\ \hline 1 & 1 \end{array}\right) \Rightarrow \widetilde{Q} = \left(\begin{array}{cc} 1 & 0\\ 1 & 1\\ \mathbf{0} & 1\\ \hline 1 & 0\\ 1 & 1\\ \mathbf{0} & 1\\ \hline 1 & 1 \end{array}\right) \Rightarrow \widetilde{\Gamma} = \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array}\\ \left(\begin{array}{cccc} 0 & 0 & 1 & 1\\ 0 & 1 & 1 & 1\\ 0 & 1 & 0 & 1\\ \hline 0 & 0 & 1 & 1\\ 0 & 1 & 1 & 1\\ 0 & 1 & 0 & 1\\ \hline 0 & 0 & 0 & 1 \end{array}\right) \end{array}.$$

Then $\widetilde{Q}$ is a unit-shrinkage of Q, and $\widetilde{\Gamma}$ corresponds to $\widetilde{Q}$. We can see that in $\widetilde{Q}$ items

1, 2, 3 give a separable Γ_1; items 4, 5, 6 give a separable Γ_2; and item 7 alone forms $Q_{multi,2}$ which requires both attributes. So (E2.a) and (E2.b) are satisfied and $(\boldsymbol{\Theta}, \boldsymbol{p})$ are generically identifiable.

Proof of Corollary III.3 (*b*). First consider those multi-parameter items in the model. If item j conforms to a multi-parameter model, then by our definition in the end of Section 1.2, it could be a main-effect model or an all-effect model. Whichever multi-parameter model item j follows, the item parameters $\theta_{j,\boldsymbol{\alpha}}$ depend on the main effects of those required attributes of item j, so $\theta_{j,\boldsymbol{\alpha}}$ can be written in the form of (B.61) with some link function f. Now under condition (E2.a), since $\widetilde{Q}_{multi,1}$ is a unit-shrinkage of $Q_{multi,1}$, we denote

$$\begin{aligned} S_u = \{j \in \mathcal{S} : j \text{ belongs to the } \widetilde{Q}_{multi,1} \text{ part;} \\ \widetilde{\boldsymbol{q}}_j \text{ in } \widetilde{Q}_{multi,1} \text{ is a unit-shrinkage of } \boldsymbol{q}_j\}. \end{aligned}$$

Then for each $j \in S_u$, there exists some $k_j \in \{1, \ldots, K\}$ such that $\widetilde{q}_{j,k_j} = q_{j,k_j} = \boldsymbol{e}_{k_j}$. We claim that the $J \times |\mathcal{A}|$ matrix $\widetilde{\boldsymbol{\Theta}} = (\widetilde{\boldsymbol{\theta}}_{j,\boldsymbol{\alpha}})$ defined as follows actually give item parameters that form a submodel of the original model being considered.

$$\widetilde{\theta}_{j,\boldsymbol{\alpha}} = \begin{cases} f(\beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k}\, \widetilde{q}_{j,k}\, \alpha_k) = f(\beta_{j,0} + \beta_{j,k_j}\, \alpha_{k_j}), & j \in S_u,\ \boldsymbol{\alpha} \in \mathcal{A}; \\ \theta_{j,\boldsymbol{\alpha}}, & j \notin S_u,\ \boldsymbol{\alpha} \in \mathcal{A}. \end{cases} \tag{B.68}$$

In other words, $(\widetilde{\boldsymbol{\Theta}}, \boldsymbol{p})$ are a valid set of parameters under the original Q-matrix and original model assumption. This is because setting all the interaction-effect coefficients and all the main-effect coefficients in (B.61) other than $\{\beta_{j,k_j} : j \in S_u\}$ to zero gives (C.2). Note that for each item j with $\boldsymbol{q}$-vector $\widetilde{\boldsymbol{q}}_j = \boldsymbol{e}_{k_j}$, (C.2) actually defines a two-parameter conjunctive model for item j, with the two levels of item parameters being $\widetilde{\theta}_j^+ = f(\beta_{j,0} + \beta_{j,k_j})$ and $\widetilde{\theta}_j^- = f(\beta_{j,0})$. Now we claim that given the

$\widetilde{\mathbf{\Theta}}$ constructed in (C.2), and given the two separable matrices Γ_1 and Γ_2 described in (E2.a), $\widetilde{\mathbf{\Theta}}_{\Gamma_1}$ and $\widetilde{\mathbf{\Theta}}_{\Gamma_2}$ both have full column rank.

In summary, the above reasoning from (E2.a) indicates that the two T-matrices $T(\mathbf{\Theta}_{\Gamma_1})$ and $T(\mathbf{\Theta}_{\Gamma_2})$ are both generically full-column-rank. Combining condition (E2.b) that each column contains at least one entry of "1" in the submatrix $Q_{multi,2}$, a similar argument as that in the proof of Theorem III.7 gives that the entire model is generically identifiable. □

Proof of Proposition III.7. We introduce a useful lemma before proving the proposition.

Lemma B.6. *Under a restricted latent class model with categorial responses $\boldsymbol{R} \in \prod_{j=1}^J\{0, 1, \dots, L_j - 1\}$, if two sets of parameters $(\mathbf{\Theta}^{cat}, \boldsymbol{p})$ and $(\bar{\mathbf{\Theta}}^{cat}, \bar{\boldsymbol{p}})$ satisfy*

$$\mathbb{P}(\boldsymbol{R} \mid \mathbf{\Theta}^{cat}, \boldsymbol{p}) = \mathbb{P}(\boldsymbol{R} \mid \bar{\mathbf{\Theta}}^{cat}, \bar{\boldsymbol{p}}), \tag{B.69}$$

then for any response pattern $\boldsymbol{r}^H = (r_1^H, \dots, r_J^H) \in \prod_{j=1}^J\{1, \dots, L_j - 1\}$ that consists of higher-level responses (higher than the basic level-0) to all the items, we have the following 2^J equalities

$$\sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j:\, r_j = r_j^H} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{p}_{\boldsymbol{\alpha}} \prod_{j:\, r_j = r_j^H} \bar{\theta}_{j,\boldsymbol{\alpha}}^{(r_j^H)}, \quad \forall \boldsymbol{r} \in \prod_{j=1}^J \{0, r_j^H\}. \tag{B.70}$$

We now continue with the proof of Proposition III.7. Given any higher-level response pattern $\boldsymbol{r}^H$, we can define a generalized T-matrix $T^{\boldsymbol{r}^H}$ of size $2^J \times m$, with the $(\boldsymbol{r}, \boldsymbol{\alpha})$th entry being

$$\{T^{\boldsymbol{r}^H}(\mathbf{\Theta}^{\text{cat}})\}_{\boldsymbol{r},\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j:\, r_j = r_j^H} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)}, \quad \boldsymbol{r} \in \prod_{j=1}^J \{0, r_j^H\}.$$

Then (B.70) in Lemma B.6 can be rewritten as

$$T^{\boldsymbol{r}^H}(\boldsymbol{\Theta}^{\mathrm{cat}})\boldsymbol{p} = T^{\boldsymbol{r}^H}(\bar{\boldsymbol{\Theta}}^{\mathrm{cat}})\bar{\boldsymbol{p}}. \tag{B.71}$$

which has the same form as (C.1), $T(\boldsymbol{\Theta})\boldsymbol{p} = T(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}}$. Now consider all the proposed sufficient conditions for strict (or $\boldsymbol{p}$-partial, generic) identifiability in Sections 3.2 and 3.3. In those proofs, we always start with assuming (C.1) holds and then show $(\boldsymbol{\Theta}, \boldsymbol{p}) = (\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ under those sufficient conditions. In the current case of categorical responses, under the same set of sufficient conditions as those in Sections 3.2 and 3.3, assuming (B.71) holds leads to $\boldsymbol{p} = \bar{\boldsymbol{p}}$ and

$$\theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} = \bar{\theta}_{j,\boldsymbol{\alpha}}^{(r_j^H)}, \quad \forall \boldsymbol{\alpha} \in \mathcal{A},\ j \in \{1, \ldots, J\},$$

for the specific $\boldsymbol{r}^H$. Since $\boldsymbol{r}^H$ is arbitrary, we obtain

$$\theta_{j,\boldsymbol{\alpha}}^{(r_j)} = \bar{\theta}_{j,\boldsymbol{\alpha}}^{(r_j)}, \quad \forall \boldsymbol{\alpha} \in \mathcal{A},\ j \in \{1, \ldots, J\},\ r_j \in \{1, \ldots, L_j - 1\}.$$

This further gives $\theta_{j,\boldsymbol{\alpha}}^{(0)} = 1 - \sum_{l>0} \theta_{j,\boldsymbol{\alpha}}^{(l)} = 1 - \sum_{l>0} \bar{\theta}_{j,\boldsymbol{\alpha}}^{(l)} = \bar{\theta}_{j,\boldsymbol{\alpha}}^{(0)}$ for any item j. By far we have shown if (B.69) holds and the previously proposed sufficient identifiability conditions are satisfied, then $(\boldsymbol{\Theta}^{\mathrm{cat}}, \boldsymbol{p}) = (\bar{\boldsymbol{\Theta}}^{\mathrm{cat}}, \bar{\boldsymbol{p}})$ hold. This concludes the proof of the proposition. □

Proof of Proposition III.8. We rewrite the probability distribution function of a RBM as

$$\mathbb{P}(\boldsymbol{R}, \boldsymbol{\alpha}^{(1)}, \cdots) = \frac{1}{Z} \exp\Big(- \boldsymbol{R}^\top \boldsymbol{W}^Q \boldsymbol{\alpha}^{(1)} - (\boldsymbol{\alpha}^{(1)})^\top \boldsymbol{U} \boldsymbol{\alpha}^{(2)} - \cdots \Big), \tag{B.72}$$

where the "$\cdots$" part denote deeper latent layers $\boldsymbol{\alpha}^{(2)}$, $\boldsymbol{\alpha}^{(3)}$, etc. The conditional distribution of R_j given $\boldsymbol{\alpha}^{(1)}$ can be written as

$$\mathbb{P}(R_j = 1 \mid \boldsymbol{\alpha}^{(1)}) = \frac{\exp\Big(W^Q_{j,\bullet}\boldsymbol{\alpha}^{(1)} + a_j\Big)}{1 + \exp\Big(W^Q_{j,\bullet}\boldsymbol{\alpha}^{(1)} + a_j\Big)} \tag{B.73}$$
$$= \sigma\Big(W^Q_{j,\bullet}\boldsymbol{\alpha}^{(1)} + a_j\Big),$$

where $\sigma(x) = e^x/(1+e^x)$ denotes the sigmoid function. Denote the length of $\boldsymbol{\alpha}^{(1)}$ by K_1. Since $\boldsymbol{\alpha}^{(1)} \in \{0,1\}^{K_1}$ can be viewed as a latent attribute pattern, we denote $\boldsymbol{\alpha}^{(1)} = (\alpha^{(1)}_1, \ldots, \alpha^{(1)}_{K_1})$ and further write (B.73) as

$$\theta_{j,\boldsymbol{\alpha}^{(1)}} = \sigma\Big(\sum_{k:\, W^Q_{j,k}\neq 0} W^Q_{j,k}\alpha^{(1)}_k\Big).$$

Now it is clear from the above display that the RBM defined in (B.72) can be viewed as a multi-parameter main-effect restricted latent class model with J items and K_1 latent attributes, with a Q-matrix resulting from the sparse bipartite structure $\boldsymbol{W}^Q$. Therefore, part (a) of the theorem follows from the generic identifiability result of the unrestricted latent class models (Allman et al., 2009) that $J \geq 2K_1 + 1$ suffices for generic identifiability of the item parameters $\boldsymbol{\Theta}$, and hence $\boldsymbol{W}^Q$. Also, part (b) of the theorem holds because when Q satisfies the sufficient conditions for strict or generic identifiability under a multi-parameter restricted latent class model, the item parameters $\boldsymbol{\Theta} = (\theta_{j,\boldsymbol{\alpha}})$ are strictly or generically identifiable. This completes the proof of the theorem. $\square$

B.5 Proof of Technical Lemmas in Chapter III

Proof of Lemma B.1 (on page 185, Section B.2). Without loss of generality, assume Γ^S is separable with the item set $S = \{1, \ldots, J\}$. Define $\boldsymbol{\theta}^* = \sum_{j\in S}\theta_{\boldsymbol{e}_j,\mathbf{1}}$. The aim is to find response patterns $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ such that the corresponding row vectors of $\boldsymbol{r}_0 := \mathbf{0}, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ in the transformed $T(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ form a $m \times m$ lower triangular matrix with nonzero diagonal elements, which will prove the conclusion that $T(\boldsymbol{\Theta})$ has full column rank m.

Since Γ^S is separable and every two different column vectors of it are distinct, without loss of generality, assume the m column vectors in the ideal response matrix Γ^S are arranged in a lexicographic order, where the first column is an all-zero column corresponding to the universal least capable class $\boldsymbol{\alpha}_0$. In other words, for any $0 \leq k < h \leq m-1$, $\Gamma^S_{\cdot,\boldsymbol{\alpha}_k}$ is of smaller lexicographical order than $\Gamma^S_{\cdot,\boldsymbol{\alpha}_h}$. In the following proof denote $\Gamma := \Gamma^S$ to simplify notations. Define response patterns $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ to be

$$\boldsymbol{r}_k = \sum_{j:\, \Gamma_{j,\,\boldsymbol{\alpha}_k}=0} \boldsymbol{e}_j, \quad k = 1, \ldots, C-1,$$

and define a sub-matrix T^{sub} of $T(\boldsymbol{\Theta})$ whose m rows corresponding to response patterns $\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ and m columns corresponding to class profiles $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m-1}$. We claim that $T^{sub}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ is a lower triangular square matrix of full rank m. This is because for any $0 \leq k \leq m-1$, the row vector corresponding to $\boldsymbol{r}_k$ in $T^{sub}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ is

$$T^{sub}_{\boldsymbol{r}_k,\boldsymbol{\cdot}}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \bigodot_{j:\, \Gamma_{j,\,\boldsymbol{\alpha}_k}=0} T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}\left(\boldsymbol{\Theta} - \Big(\sum_{j=1}^{J} \theta_{\boldsymbol{e}_j,\mathbf{1}}\boldsymbol{e}_j\Big)\mathbf{1}^\top\right). \tag{B.74}$$

For any $h > k$, there must exist an item j such that $\Gamma_{j,\boldsymbol{\alpha}_k} = 0$ and $\Gamma_{j,\boldsymbol{\alpha}_h} = 1$. Existence of such j means that $\boldsymbol{\alpha}_h$ is capable of at least one item not mastered by $\boldsymbol{\alpha}_k$, and guarantees that the $\boldsymbol{\alpha}_h$-entry of the above row vector (B.74) is zero. We have shown $T^{sub}_{\boldsymbol{r}_k,\boldsymbol{\alpha}_h}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = 0$ for arbitrary $0 \leq k < h \leq m-1$, so $T^{sub}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)$ is a lower triangular square matrix. Moreover, the diagonal entries are

$$T^{sub}_{\boldsymbol{r}_k,\boldsymbol{\alpha}_k}(\boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \prod_{j:\, \Gamma_{j,\boldsymbol{\alpha}_k}=0} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_k} - \theta_{\boldsymbol{e}_j,\mathbf{1}}) \neq 0,$$

so $T^{sub}(\boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)$ is of rank m, with the shape

$$\begin{vmatrix} \prod_{j \in S_{\boldsymbol{\alpha}_0}} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} - \theta_{\boldsymbol{e}_j,\mathbf{1}}) & 0 & \cdots & 0 \\ \prod_{j \in S_{\boldsymbol{\alpha}_1}} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} - \theta_{\boldsymbol{e}_j,\mathbf{1}}) & \prod_{j \in S_{\boldsymbol{\alpha}_1}} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_1} - \theta_{\boldsymbol{e}_j,\mathbf{1}}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \prod_{j \in S_{\boldsymbol{\alpha}_{m-1}}} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_0} - \theta_{\boldsymbol{e}_j,\mathbf{1}}) & * & \cdots & \prod_{j \in S_{\boldsymbol{\alpha}_{m-1}}} (\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}_{m-1}} - \theta_{\boldsymbol{e}_j,\mathbf{1}}) \end{vmatrix}$$

where $S_{\boldsymbol{\alpha}_i} := \{j : \Gamma_{j,\boldsymbol{\alpha}_i} = 0\}$ for $i = 0, 1, \ldots, m-1$. The proof of Lemma B.1 is complete. □

Proof of Lemma B.2 (on page 188, Section B.2). Since $\theta_j^+ > \theta_j^-$ for each item j from the definition constraints of restricted latent class models, $T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}(\boldsymbol{\Theta})\boldsymbol{p} = T_{\boldsymbol{e}_j,\boldsymbol{\cdot}}(\bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}}$ indicates

$$\theta_j^+ = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \theta_j^+ p_{\boldsymbol{\alpha}} \geq \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \theta_{j,\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \bar{\theta}_{j,\boldsymbol{\alpha}} \bar{p}_{\boldsymbol{\alpha}} \geq \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \bar{\theta}_j^- \bar{p}_{\boldsymbol{\alpha}} = \bar{\theta}_j^-,$$

where among the two "$\geq$" there is at least a strict "$>$". This is because the first "$\geq$" is an equality sign only if all the latent classes are capable of item j, namely $\Gamma_{j,\boldsymbol{\alpha}} = 1$ for all $\boldsymbol{\alpha} \in \mathcal{A}$, and in this case, $\sum_{\boldsymbol{\alpha} \in \mathcal{A}} \bar{\theta}_{j,\boldsymbol{\alpha}} \bar{p}_{\boldsymbol{\alpha}} = \bar{\theta}_j^+ > \bar{\theta}_j^-$ and therefore the second "$\geq$" must be a strict "$>$". Similarly, the second "$\geq$" is an equality sign only if all the latent classes are incapable of item j, and in this case, $\theta_j^+ > \theta_j^- = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \theta_{j,\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}}$ and therefore the first "$\geq$" must be a strict "$>$". This proves that $\theta_j^+ > \bar{\theta}_j^-$ for all j, and similarly we have $\theta_j^- < \bar{\theta}_j^+$ for all j. □

Proof of Lemma B.3 (on page 213, Section B.3). Since the sub-matrix $T(\boldsymbol{\Theta}_{S_1})$ has full column rank m, there exists a vector $\boldsymbol{m}_{\boldsymbol{\alpha}}$ such that

$$\boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_1}) = (\mathbf{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}),$$

On the other hand, Equation (C.1) implies $\boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_1})\boldsymbol{p} = \boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\bar{\boldsymbol{p}}$, which further indicates the row vector $\boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})$ also contains at least one nonzero element in some column. Denote such a column by $\boldsymbol{\alpha}^*$ and the nonzero value by $\bar{x}_{\boldsymbol{\alpha}^*}$. Since the sub-matrix $T(\bar{\boldsymbol{\Theta}}_{S_2})$ also has full column rank, there exists another vector $\boldsymbol{n}_{\boldsymbol{\alpha}^*}$ such that

$$\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2}) = (\mathbf{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}^*}, \mathbf{0}),$$

Again from Equation (C.1) we have that the $\boldsymbol{\alpha}$-th entry of the row vector $\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top T(\boldsymbol{\Theta}_{S_2})$ is also nonzero. We denote this nonzero value by $y_{\boldsymbol{\alpha}}$. Then we have

$$\{\boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\} = (\mathbf{0}, \underbrace{y_{\boldsymbol{\alpha}}}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}),$$

$$\{\boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\} = (\mathbf{0}, \underbrace{\bar{x}_{\boldsymbol{\alpha}^*}}_{\text{column } \boldsymbol{\alpha}^*}, \mathbf{0}).$$

Now consider one more row $\boldsymbol{e}_j$ for an arbitrary $j > M_1 + M_2$ in the T-matrix, we have

$$T_{\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta}) \odot \{\boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\} = (\mathbf{0}, \underbrace{\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} \cdot y_{\boldsymbol{\alpha}}}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}),$$

$$T_{\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}}) \odot \{\boldsymbol{m}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\} = (\mathbf{0}, \underbrace{\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*} \cdot \bar{x}_{\boldsymbol{\alpha}^*}}_{\text{column } \boldsymbol{\alpha}^*}, \mathbf{0}).$$

The above four equations along with Equation (C.1) imply

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*}, \quad \forall j > M_1 + M_2. \tag{B.75}$$

Now that $(\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}}, j > M_1 + M_2) = (\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*}, j > M_1 + M_2)$, condition (C4) implies that there exists a vector $\boldsymbol{s}_{\boldsymbol{\alpha}}$ such that

$$\boldsymbol{s}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{(M_1+M_2+1):J}) = (0, *, \dots, *, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, *, \dots, *),$$

$$\boldsymbol{s}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{(M_1+M_2+1):J}) = (0, *, \dots, *, \underbrace{1}_{\text{column } \boldsymbol{\alpha}^*}, *, \dots, *).$$

Next redefine

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

then we have

$$\begin{aligned} &\{\boldsymbol{s}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{(M_1+M_2+1):J})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\} \odot \{T_{\boldsymbol{r}^*,\cdot}(\boldsymbol{\Theta}_{S_1} - \boldsymbol{\theta}^* \boldsymbol{1}^\top)\} \\ &= \Big(\boldsymbol{0}, \underbrace{y_{\boldsymbol{\alpha}} \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\boldsymbol{1}})}_{\text{column } \boldsymbol{\alpha}}, \boldsymbol{0}\Big), \end{aligned} \tag{B.76}$$

and

$$\begin{aligned} &\{\boldsymbol{s}_{\boldsymbol{\alpha}}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{(M_1+M_2+1):J})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\} \odot \{T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}}_{S_1} - \boldsymbol{\theta}^* \boldsymbol{1}^\top)\} \\ &= \Big(\boldsymbol{0}, \underbrace{\prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\boldsymbol{1}})}_{\text{column } \boldsymbol{\alpha}^*}, \boldsymbol{0}\Big). \end{aligned} \tag{B.77}$$

Since the $\boldsymbol{\alpha}$-entry of (B.76) is nonzero, the $\boldsymbol{\alpha}^*$-entry of (B.77) must also be nonzero since by (C.1) we have (B.76) $\cdot \boldsymbol{p} =$ (B.77) $\cdot \bar{\boldsymbol{p}}$. Further consider row $j \in S_1$ such that $\Gamma_{j,\boldsymbol{\alpha}} = 1$. Obviously $\boldsymbol{e}_j$ does not appear in the summation of the previously defined $\boldsymbol{r}^*$, so we have

$$\begin{aligned} &\{\boldsymbol{s}_{\boldsymbol{\alpha}}^\top \cdot T(\boldsymbol{\Theta}_{(M_1+M_2+1):J})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\} \odot \{T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\boldsymbol{\Theta}_{S_1} - \boldsymbol{\theta}^* \boldsymbol{1}^\top)\} \\ &= \Big(\boldsymbol{0}, \underbrace{\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} y_{\boldsymbol{\alpha}} \prod_{h \in S_1 : \Gamma_{h,\boldsymbol{\alpha}}=0} (\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\boldsymbol{1}})}_{\text{column } \boldsymbol{\alpha}}, \boldsymbol{0}\Big), \end{aligned} \tag{B.78}$$

and

$$\{\boldsymbol{s}_{\boldsymbol{\alpha}}^{\top} \cdot T(\bar{\boldsymbol{\Theta}}_{(M_1+M_2+1):J})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^{\top} \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\} \odot \{T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}}_{S_1} - \boldsymbol{\theta}^*\boldsymbol{1}^{\top})\}$$
$$= \Big(\boldsymbol{0}, \underbrace{\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*} \prod_{h\in S_1:\Gamma_{h,\boldsymbol{\alpha}}=0} (\bar{\theta}_{\boldsymbol{e}_h,\boldsymbol{\alpha}} - \theta_{\boldsymbol{e}_j,\boldsymbol{1}}), \boldsymbol{0}}_{\text{column } \boldsymbol{\alpha}^*}\Big), \tag{B.79}$$

and therefore

$$\theta_{\boldsymbol{e}_j,\boldsymbol{1}} = \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \frac{(\text{B.78}) \cdot \boldsymbol{p}}{(\text{B.76}) \cdot \boldsymbol{p}} = \frac{(\text{B.79}) \cdot \bar{\boldsymbol{p}}}{(\text{B.77}) \cdot \bar{\boldsymbol{p}}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*}, \quad \forall j \in S_1 \text{ s.t. } \Gamma_{j,\boldsymbol{\alpha}} = 1.$$

Therefore for any $j \in S_1$ and any $\boldsymbol{\alpha}'$ such that $\Gamma_{j,\boldsymbol{\alpha}'} = 0$, as long as there exists some $\boldsymbol{\alpha}$ such that $\Gamma_{j,\boldsymbol{\alpha}} = 1$, we have $\theta_{\boldsymbol{e}_j,\boldsymbol{1}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*}$ from the above proof. Then the following inequality holds

$$\forall j \in S_1, \quad \forall \boldsymbol{\alpha}', \text{ s.t. } \Gamma_{j,\boldsymbol{\alpha}'=0}, \quad \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}'} < \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \theta_{\boldsymbol{e}_j,\boldsymbol{1}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*} \leq \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{1}}.$$

Similarly we also have $\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}'} < \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{1}}$ for any $j \in S_2$ and $\Gamma_{j,\boldsymbol{\alpha}'} = 0$; and $\theta_{\boldsymbol{e}_j,\boldsymbol{1}} > \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}'}$ for any $j \in S_1 \cup S_2$ and $\Gamma_{j,\boldsymbol{\alpha}'} = 0$. The proof of Lemma B.3 is complete. □

Proof of Lemma D.1 (on page 323, Section B.3). We focus on $T(\boldsymbol{\Theta}_{S_2})$ first. As shown in the proof of Lemma B.3, for any $\boldsymbol{\alpha}$ there exists some $\boldsymbol{\alpha}^*$, which depends on $\boldsymbol{\alpha}$, such that

$$\begin{aligned}
&\boldsymbol{m}_{\boldsymbol{\alpha}}^{\top} \cdot T(\boldsymbol{\Theta}_{S_1}) = (\boldsymbol{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}}, \boldsymbol{0}),\\
&\boldsymbol{n}_{\boldsymbol{\alpha}^*}^{\top} \cdot T(\bar{\boldsymbol{\Theta}}_{S_2}) = (\boldsymbol{0}, \underbrace{1}_{\text{column } \boldsymbol{\alpha}^*}, \boldsymbol{0}),\\
&\{\boldsymbol{m}_{\boldsymbol{\alpha}}^{\top} \cdot T(\boldsymbol{\Theta}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^{\top} \cdot T(\boldsymbol{\Theta}_{S_2})\} = (\boldsymbol{0}, \underbrace{y_{\boldsymbol{\alpha}}}_{\text{column } \boldsymbol{\alpha}}, \boldsymbol{0}), \quad y_{\boldsymbol{\alpha}} \neq 0\\
&\{\boldsymbol{m}_{\boldsymbol{\alpha}}^{\top} \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^{\top} \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\} = (\boldsymbol{0}, \underbrace{\bar{x}_{\boldsymbol{\alpha}^*}}_{\text{column } \boldsymbol{\alpha}^*}, \boldsymbol{0}), \quad \bar{x}_{\boldsymbol{\alpha}^*} \neq 0
\end{aligned}$$

for some vectors $\boldsymbol{m}_{\boldsymbol{\alpha}}$ and $\boldsymbol{n}_{\boldsymbol{\alpha}^*}$. And based on these constructions we proved

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*}, \quad \forall j > M_1 + M_2.$$

Clearly from the constructions we have

$$\{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\}_{\boldsymbol{\alpha}} \neq 0,$$

then we furthermore claim that under condition (C4*), $\boldsymbol{n}_{\boldsymbol{\alpha}^*}$ also has the following property

$$\{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\}_{\boldsymbol{\alpha}'} = 0, \quad \forall \boldsymbol{\alpha}' \precneqq_{S_1} \boldsymbol{\alpha}. \tag{B.80}$$

Since otherwise if $\{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\}_{\boldsymbol{\alpha}'} = 0$ for some $\boldsymbol{\alpha}' \preceq_{S_1} \boldsymbol{\alpha}$, we would have

$$\{\boldsymbol{m}_{\boldsymbol{\alpha}'}^\top \cdot T(\boldsymbol{\Theta}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\boldsymbol{\Theta}_{S_2})\} = (\boldsymbol{0}, \underbrace{s_{\boldsymbol{\alpha}'}}_{\text{column } \boldsymbol{\alpha}'}, \boldsymbol{0}), \quad s_{\boldsymbol{\alpha}'} \neq 0, \tag{B.81}$$

$$\{\boldsymbol{m}_{\boldsymbol{\alpha}'}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_1})\} \odot \{\boldsymbol{n}_{\boldsymbol{\alpha}^*}^\top \cdot T(\bar{\boldsymbol{\Theta}}_{S_2})\} = (\boldsymbol{0}, \underbrace{\bar{t}_{\boldsymbol{\alpha}^*}}_{\text{column } \boldsymbol{\alpha}^*}, \boldsymbol{0}), \quad \bar{t}_{\boldsymbol{\alpha}^*} \neq 0, \tag{B.82}$$

then using similar argument as that in Lemma B.3, for any $j \in (S_1 \cup S_2)^c$ we would have

$$\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}'} = \frac{\theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}'} \cdot (\text{B.81}) \cdot \boldsymbol{p}}{(\text{B.81}) \cdot \boldsymbol{p}} = \frac{\bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}'} \cdot (\text{B.82}) \cdot \bar{\boldsymbol{p}}}{(\text{B.82}) \cdot \bar{\boldsymbol{p}}} = \bar{\theta}_{\boldsymbol{e}_j,\boldsymbol{\alpha}^*} = \theta_{\boldsymbol{e}_j,\boldsymbol{\alpha}},$$

which contradicts Condition (C4) that $\Gamma^{(S_1 \cup S_2)^c}_{\bullet,\boldsymbol{\alpha}} = \Gamma^{(S_1 \cup S_2)^c}_{\bullet,\boldsymbol{\alpha}'}$ for any $\boldsymbol{\alpha}' \precneqq_{S_1} \boldsymbol{\alpha}$, since (C4) naturally leads to $(\theta_{j,\boldsymbol{\alpha}}, j \in (S_1 \cup S_2)^c) \neq (\theta_{j,\boldsymbol{\alpha}'}, j \in (S_1 \cup S_2)^c)$ for any $\boldsymbol{\alpha}' \precneqq_{S_1} \boldsymbol{\alpha}$. So the claim (B.80) must hold. By far we have found $\boldsymbol{v}_{\boldsymbol{\alpha}} := \boldsymbol{n}_{\boldsymbol{\alpha}^*}$ that satisfies the first equation in (B.30) for each $\boldsymbol{\alpha}$. By symmetry between $(\boldsymbol{\Theta}, \boldsymbol{p})$ and $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$, using exactly the same techniques will lead to $\boldsymbol{u}_{\boldsymbol{\alpha}}$ for each $\boldsymbol{\alpha}$ that satisfies the second equation in (B.30). □

Proof of Lemma B.6 (on page 244, Section B.4). Equation (B.70) for $\boldsymbol{R} = \boldsymbol{r}$ can be written as

$$\sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j:\, r_j=r_j^H} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} \prod_{j:\, r_j\neq r_j^H} \theta_{j,\boldsymbol{\alpha}}^{(r_j)} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{p}_{\boldsymbol{\alpha}} \prod_{j:\, r_j=r_j^H} \bar{\theta}_{j,\boldsymbol{\alpha}}^{(r_j^H)} \prod_{j:\, r_j\neq r_j^H} \bar{\theta}_{j,\boldsymbol{\alpha}}^{(r_j)}. \tag{B.83}$$

We denote $\{0,\dots,L_j-1\}$ by $[L_j]$ for simplicity. Now consider an arbitrary item set $S\subseteq\boldsymbol{\mathcal{S}}$, and we write $\boldsymbol{r}_S=\boldsymbol{r}_S^H$ if $r_j=r_j^H$ for any $j\in S$. For this S, we sum (B.83) over all response patterns $\boldsymbol{r}$ for which $r_j=r_j^H$ if and only if $j\in S$ (i.e., $\boldsymbol{r}$ satisfies $\boldsymbol{r}_S=\boldsymbol{r}_S^H$ and $\boldsymbol{r}_{S^c}\in\prod_{j\notin S}[L_j]\setminus[r_j^H]$), then the left hand side (LHS) of the new equation is

$$\begin{aligned}
&\sum_{\substack{\boldsymbol{r}:\,\boldsymbol{r}_S=\boldsymbol{r}_S^H,\\ \boldsymbol{r}_{S^c}\in\prod_{j\neq S}[L_j]\setminus[r_j^H]}} \Big(\sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j:\, r_j=r_j^H} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} \prod_{j:\, r_j\neq r_j^H} \theta_{j,\boldsymbol{\alpha}}^{(r_j)}\Big)\\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j\in S} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} \sum_{\substack{\boldsymbol{r}:\,\boldsymbol{r}_S=\boldsymbol{r}_S^H,\\ \boldsymbol{r}_{S^c}\in\prod_{j\neq S}[L_j]\setminus[r_j^H]}} \prod_{j\notin S} \theta_{j,\boldsymbol{\alpha}}^{(r_j)}\\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j\in S} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} \prod_{j\notin S} \Big(\sum_{r_j\neq r_j^H} \theta_{j,\boldsymbol{\alpha}}^{(r_j)}\Big)\\
&= \sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j\in S} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} \prod_{j\notin S} \Big(1-\theta_{j,\boldsymbol{\alpha}}^{(r_j^H)}\Big),
\end{aligned}$$

so from (B.83) we have

$$\sum_{\boldsymbol{\alpha}\in\mathcal{A}} p_{\boldsymbol{\alpha}} \prod_{j\in S} \theta_{j,\boldsymbol{\alpha}}^{(r_j^H)} \prod_{j\notin S} \Big(1-\theta_{j,\boldsymbol{\alpha}}^{(r_j^H)}\Big) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \bar{p}_{\boldsymbol{\alpha}} \prod_{j\in S} \bar{\theta}_{j,\boldsymbol{\alpha}}^{(r_j^H)} \prod_{j\notin S} \Big(1-\bar{\theta}_{j,\boldsymbol{\alpha}}^{(r_j^H)}\Big) \tag{B.84}$$

holds for any $S\subseteq\boldsymbol{\mathcal{S}}$. By far we have shown the system of 2^J equations (B.84) hold for any $\boldsymbol{r}^H$. Note that (B.84) can be viewed as probability of a response pattern consisting of binary responses, where for each item j and each latent class $\boldsymbol{\alpha}$, there are two possible responses with probabilities $\theta_{j,\boldsymbol{\alpha}}^{(r_j^H)}$ and $1-\theta_{j,\boldsymbol{\alpha}}^{(r_j^H)}$ respectively. Then similar to the proof of Proposition III.1 which establishes equivalence between equality of probability mass functions and equality of marginal probabilities, (B.84) is equivalent to (B.70) in the lemma. This completes the proof of Lemma B.6. □

APPENDIX C

Appendix of Chapter IV

This is the appendix to Chapter IV and it is organized as follows. Appendix C.1 gives the proof of Proposition IV.1. Appendix C.2 presents the proof of Theorem IV.1, one of the main results of this Chapter IV. Appendix C.3 gives the proof of Theorem IV.2. Appendix C.4 gives the proof of Theorem IV.3. Appendix C.5 presents the proof of Theorem IV.4. Appendix C.6 gives the proof of Theorem IV.5. Appendix C.7 gives the proof of Proposition IV.4. Appendix C.8 presents various simulation studies for Chapter IV.

We introduce some additional notations. For a submatrix Q_1 of Q that has size $J_1 \times K$, we denote the item parameter matrix corresponding to these J_1 items by $\mathbf{\Theta}_{Q_1}$, then $\mathbf{\Theta}_{Q_1}$ is a $J_1 \times K$ submatrix of $\mathbf{\Theta}$. Denote Q_1's corresponding T-matrix by $T(Q_1, \mathbf{\Theta}_{Q_1})$, then $T(Q_1, \mathbf{\Theta}_{Q_1})$ has size $2^{J_1} \times 2^K$. For notational simplicity, in the following we denote $\boldsymbol{\theta}^+ \equiv \mathbf{1} - \boldsymbol{s}$ under the DINA model, then $\mathbf{\Theta} = (\mathbf{1} - \boldsymbol{s}, \boldsymbol{g}) = (\boldsymbol{\theta}^+, \boldsymbol{g})$ under DINA. The following useful lemma is in the same spirit as Proposition B.1 and Proposition B.2 and its proof is omitted.

Lemma C.1. *Under a restricted latent class model,* $(Q, \mathbf{\Theta}, \boldsymbol{p})$ *are identifiable if and* *only if for any* $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ *and* $(\bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$*,*

$$T(Q, \boldsymbol{\Theta})\boldsymbol{p} = T(\bar{Q}, \bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}} \tag{C.1}$$

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

$$T(Q, \boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top) = D(\boldsymbol{\theta}^*) T(Q, \boldsymbol{\Theta}). \tag{C.2}$$

We add some remarks on Lemma C.1. First, Equation (C.1) can be written as that, for any response pattern $\boldsymbol{r} \in \{0,1\}^J$, $T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q, \boldsymbol{\Theta})\boldsymbol{p} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}}$. Second, thanks to (C.2), for any $\boldsymbol{\theta}^* = (\theta_1, \ldots, \theta_J)^\top \in \mathbb{R}^J$, equality (C.1) leads to

$$T(Q, \boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)\boldsymbol{p} = T(\bar{Q}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)\bar{\boldsymbol{p}},$$

and further $T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q, \boldsymbol{\Theta} - \boldsymbol{\theta}^* \mathbf{1}^\top)\boldsymbol{p} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top)\bar{\boldsymbol{p}}$ for any $\boldsymbol{r} \in \{0,1\}^J$. Besides, If (C.1) holds, then for any submatrix Q_1 of Q, equality $T(Q_1, \boldsymbol{\Theta}_{Q_1})\boldsymbol{p} = T(\bar{Q}_1, \bar{\boldsymbol{\Theta}}_{\bar{Q}_1})\bar{\boldsymbol{p}}$ also holds.

C.1 Proof of Proposition IV.1

Consider a Q-matrix of size $J \times K$ in the form

$$Q = \begin{pmatrix} Q' \\ \mathbf{0} \end{pmatrix},$$

where Q' is of size $J' \times K$ and contains those nonzero $\boldsymbol{q}$-vectors of Q. For any item $j \in \{J'+1, \ldots, J\}$ which has $\boldsymbol{q}_j = \mathbf{0}$, all the attribute profiles $\boldsymbol{\alpha}$ satisfy $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$, so there is only one item parameter associated with j under Q, and we denote it by θ_j.

Denote the first J' rows of $\boldsymbol{\Theta}$ by $\boldsymbol{\Theta}'$. Denote the $2^{J'} \times 2^K$ T-matrix associated with matrix Q' by $T'(Q', \boldsymbol{\Theta}')$.

First consider the case where $(Q', \boldsymbol{\Theta}', \boldsymbol{p})$ are strictly (or generically) identifiable, and we will show $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ are also strictly (or generically) identifiable. Assume there is a $J \times K$ matrix $\bar{Q}$ and associated parameters $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ such that (C.1) holds. Denote the submatrix of $\bar{Q}$ containing its first J' rows by $\bar{Q}'$, and the submatrix of $\bar{\boldsymbol{\Theta}}$ containing its first J' rows by $\bar{\boldsymbol{\Theta}}'$. Then (C.1) implies $T(Q', \boldsymbol{\Theta}')\boldsymbol{p}' = T(\bar{Q}', \bar{\boldsymbol{\Theta}}')\bar{\boldsymbol{p}}'$, and the strict (or generic) joint identifiability of $(Q', \boldsymbol{\Theta}', \boldsymbol{p})$ gives that $\bar{Q}' \sim Q'$ and $(\bar{\boldsymbol{\Theta}}', \bar{\boldsymbol{p}}) = (\boldsymbol{\Theta}', \boldsymbol{p})$. For an arbitrary RLCM, the strict (or generic) identifiability of $(Q', \boldsymbol{\Theta}', \boldsymbol{p})$ implies that $T(Q', \boldsymbol{\Theta}')$ has full rank 2^K strictly (or generically).

This is because if not so, then the proportion parameters $\boldsymbol{p}$ can not be strictly (or generically) identifiable, in the sense that there exist multiple different $\boldsymbol{p}$ such that $T(Q', \boldsymbol{\Theta}')\boldsymbol{p}$ are all equal. This would contradict the assumption that $(Q', \boldsymbol{\Theta}', \boldsymbol{p})$ are strictly (or generically) identifiable. Therefore $T(Q', \boldsymbol{\Theta}')$ is strictly (or generically) full-rank. Then for each $\boldsymbol{\alpha} \in \{0,1\}^K$ there must exist a 2^K-dimensional vector $\boldsymbol{v}_{\boldsymbol{\alpha}}$ such that

$$\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot} T(Q', \boldsymbol{\Theta}') = \boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot} T(\bar{Q}', \bar{\boldsymbol{\Theta}}') = (\mathbf{0}, \underbrace{x_{\boldsymbol{\alpha}}}_{\text{column } \boldsymbol{\alpha}}, \mathbf{0}), \quad x_{\boldsymbol{\alpha}} \neq 0,$$

and $\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot} T(Q', \boldsymbol{\Theta}')\boldsymbol{p} = \boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot} T(\bar{Q}', \bar{\boldsymbol{\Theta}}')\bar{\boldsymbol{p}} = x_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} \neq 0$. Then again use the property (C.2) and we have the following equality for any $j \in \{J'+1, \ldots, J\}$,

$$\begin{aligned}\theta_{j,\boldsymbol{\alpha}} =& \frac{\{T_{\boldsymbol{e}_j, \boldsymbol{\cdot}}(Q, \boldsymbol{\Theta}) \odot [\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot}\ T(Q', \boldsymbol{\Theta}')]\}\boldsymbol{p}}{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot}\ T(Q', \boldsymbol{\Theta}')\boldsymbol{p}} \\ =& \frac{\{T_{\boldsymbol{e}_j, \boldsymbol{\cdot}}(Q, \boldsymbol{\Theta}) \odot [\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot}\ T(\bar{Q}', \bar{\boldsymbol{\Theta}}')]\}\bar{\boldsymbol{p}}}{\boldsymbol{v}_{\boldsymbol{\alpha}}^\top \boldsymbol{\cdot}\ T(\bar{Q}', \bar{\boldsymbol{\Theta}}')\bar{\boldsymbol{p}}} = \bar{\theta}_{j,\boldsymbol{\alpha}},\end{aligned}$$

where "$\odot$" represents the element-wise product of two vectors. This proves $\boldsymbol{\Theta} = \bar{\boldsymbol{\Theta}}$ and $Q \sim \bar{Q}$. So $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ are strictly (or generically) identifiable.

Next consider the case where $(Q', \mathbf{\Theta}', \boldsymbol{p})$ are **not** strictly (or generically) identifiable, so there exist $(\bar{Q}', \bar{\mathbf{\Theta}}', \bar{\boldsymbol{p}}) \nsim (Q', \mathbf{\Theta}', \boldsymbol{p})$ such that $T'(\bar{Q}', \bar{\mathbf{\Theta}}')\bar{\boldsymbol{p}} = T'(Q', \mathbf{\Theta}')\boldsymbol{p}$. Now extend $\bar{Q}'$ to $\bar{Q}$ of size $J \times K$ by adding $J - J'$ all-zero $\boldsymbol{q}$-vectors, i.e.,

$$\bar{Q} = \begin{pmatrix} \bar{Q}' \\ \mathbf{0} \end{pmatrix},$$

and set $\bar{\theta}_j = \theta_j$ for $j \in \{J'+1, \ldots, J\}$. Then for any $\boldsymbol{r} = (r_1, \ldots, r_{J'}, r_{J'+1}, \ldots, r_J) \in \{0,1\}^J$ and the corresponding $\boldsymbol{r'} = (r_1, \ldots, r_{J'})$,

$$T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q, \mathbf{\Theta})\boldsymbol{p} = \left\{T'_{\boldsymbol{r}',\boldsymbol{\cdot}}(Q', \mathbf{\Theta}')\boldsymbol{p}\right\} \prod_{j>J'} \theta_j^{r_j};$$
$$T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{Q}, \bar{\mathbf{\Theta}})\bar{\boldsymbol{p}} = \left\{T'_{\boldsymbol{r}',\boldsymbol{\cdot}}(\bar{Q}', \bar{\mathbf{\Theta}}')\boldsymbol{p}\right\} \prod_{j>J'} \theta_j^{r_j}.$$

Now that $T(Q, \mathbf{\Theta})\boldsymbol{p} = T(\bar{Q}, \bar{\mathbf{\Theta}})\bar{\boldsymbol{p}}$ but $(\bar{Q}, \bar{\mathbf{\Theta}}, \bar{\boldsymbol{p}}) \nsim (Q, \mathbf{\Theta}, \boldsymbol{p})$, we obtain that $(Q, \mathbf{\Theta}, \boldsymbol{p})$ are not strictly (or generically) identifiable. The proof of the proposition is complete.

C.2 Proof of Theorem IV.1

We first prove the sufficiency, and then show the necessity of the conditions. Under DINA, (C.1) can be equivalently written as that for any $\boldsymbol{r} \in \{0,1\}^J$,

$$T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+, \boldsymbol{g})\boldsymbol{p} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{g}})\bar{\boldsymbol{p}}. \tag{C.3}$$

We first introduce some notations. In the following discussion, for an integer M, we denote $[M] = \{1, \ldots, M\}$. For an item set $S \subseteq [J]$, denote $\boldsymbol{q}_S = \vee_{j\in S}\boldsymbol{q}_j = (\max_{j\in S} q_{j,1}, \max_{j\in S} q_{j,2}, \ldots, \max_{j\in S} q_{j,K})$, then $\boldsymbol{q}_S$ is also a K-dimensional binary vector, and we denote its k element by $q_{S,k}$. Recall

$$Q = \begin{pmatrix} I_K \\ Q^\star \end{pmatrix},$$

and we denote the submatrix of $\bar{Q}$ consisting of its first K row vectors by $\bar{Q}_{1:K,\bullet}$. We next show in five steps that if (C.3) holds, then $\bar{Q} \sim Q$, and also $\boldsymbol{\theta}^+ = \bar{\boldsymbol{\theta}}^+$, $\bar{\boldsymbol{g}} = \boldsymbol{g}$, $\bar{\boldsymbol{p}} = \boldsymbol{p}$.

Step 1. After some column rearrangement, $\bar{Q}_{1:K,\bullet}$ is an upper-triangular matrix with all the diagonal elements being ones.

Step 2. $\bar{c}_j = c_j$ for all $j \in \{K+1, \ldots, J\}$.

Step 3. $\bar{g}_k = g_k$ for all $k \in \{1, \ldots, K\}$.

Step 4. $\bar{Q}_{1:K,\bullet} \sim I_K$

Step 5. $\bar{Q} \sim Q$, $\boldsymbol{\theta}^+ = \bar{\boldsymbol{\theta}}^+$, $\bar{\boldsymbol{g}} = \boldsymbol{g}$, $\bar{\boldsymbol{p}} = \boldsymbol{p}$.

For any item set $S \subseteq \{1, \ldots, J\}$, denote $\boldsymbol{\theta}_S^+ = \sum_{j \in S} c_j \boldsymbol{e}_j$, and denote $\boldsymbol{g}_S$, $\bar{\boldsymbol{\theta}}_S^+$, and $\bar{\boldsymbol{g}}_S$ similarly. Consider the response pattern $\boldsymbol{r}^\star = \sum_{j \in S} \boldsymbol{e}_j$ and any $\boldsymbol{\theta}^\star = \sum_{j \in S} \theta_j^\star \boldsymbol{e}_j$, then Equation (C.3) together with Lemma C.1 imply that

$$T_{\boldsymbol{r}^\star,\bullet}(Q, \boldsymbol{\theta}_S^+ - \boldsymbol{\theta}^\star, \boldsymbol{g}_S - \boldsymbol{\theta}^\star)\boldsymbol{p} = T_{\boldsymbol{r}^\star,\bullet}(\bar{Q}, \bar{\boldsymbol{\theta}}_S^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}}_S - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}}. \qquad \text{(C.4)}$$

We will frequently use (E.3) in the following proof. And when the item set S and response pattern $\boldsymbol{r}^\star$ are clearly implied by the definition of $\boldsymbol{\theta}^\star$, we will omit the subscript S in the above (E.3). We also frequently use the fact that when (E.3) holds, $c_j \neq \bar{g}_j$ and $g_j \neq \bar{c}_j$ for any item j. This is true because if $c_j = \bar{g}_j$, we would have

$$T_{\boldsymbol{e}_j,\bullet}(Q,\boldsymbol{\theta}^+,\boldsymbol{g})\boldsymbol{p} = c_j(\sum_{\boldsymbol{\alpha}\succeq\boldsymbol{q}_j} p_{\boldsymbol{\alpha}}) + g_j(\sum_{\boldsymbol{\alpha}\nsucceq\boldsymbol{q}_j} p_{\boldsymbol{\alpha}}) < c_j = \bar{g}_j$$
$$\leq \bar{c}_j(\sum_{\boldsymbol{\alpha}\succeq\boldsymbol{q}_j} \bar{p}_{\boldsymbol{\alpha}}) + \bar{g}_j(\sum_{\boldsymbol{\alpha}\nsucceq\boldsymbol{q}_j} \bar{p}_{\boldsymbol{\alpha}}) = T_{\boldsymbol{e}_j,\bullet}(\bar{Q},\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{g}})\bar{\boldsymbol{p}},$$

which contradicts (C.3). So $c_j \neq \bar{g}_j$ and similarly $g_j \neq \bar{c}_j$ for each j. As stated in the main text, we assume without loss of generality that there is no all-zero row vector in true Q-matrix. If, however, the jth row vector of $\bar{Q}$ equals $\mathbf{0}$, then $\bar{c}_j$ would equal $\bar{g}_j$, and we denote this value by $\bar{\theta}_j$. Equation (C.3) gives

$$\bar{\theta}_j = c_j\Big(\sum_{\boldsymbol{\alpha}:\,\boldsymbol{\alpha}\succeq\boldsymbol{q}_j} p_{\boldsymbol{\alpha}}\Big) + g_j\Big(\sum_{\boldsymbol{\alpha}:\,\boldsymbol{\alpha}\nsucceq\boldsymbol{q}_j} p_{\boldsymbol{\alpha}}\Big),$$

and hence $g_j < \bar{\theta}_j < c_j$ holds for this j.

Step 1. In this step we prove that $\bar{Q}_{1:K,\bullet}$ must take the following form after some column rearrangement,

$$\bar{Q}_{1:K,\bullet} \sim \begin{pmatrix} 1 & * & \dots & * \\ 0 & 1 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}. \tag{C.5}$$

Namely, after properly rearranging the columns of $\bar{Q}_{1:K,\bullet}$, we have $\bar{Q}_{k,k} = 1$ and $\bar{Q}_{k,h} = 0$ for any $k > h$.

We first introduce the following useful lemma.

Lemma C.2. *Suppose the true Q satisfies Condition A that $Q_{1:K} = I_K$. If there exists an item set $S \subseteq \{K+1,\dots,J\}$ such that*

$$\max_{m\in S} q_{m,h} = 0, \quad \max_{m\in S} q_{m,j} = 1\ \forall j \in \mathcal{J}$$

for some attributes $h \in [K]$ and a set of attributes $\mathcal{J} \subseteq [K] \setminus \{h\}$, then

$$\vee_{j\in\mathcal{J}}\, \bar{\boldsymbol{q}}_j \not\succeq \bar{\boldsymbol{q}}_h.$$

Proof of Lemma C.2. We prove by contradiction. Assume there exist attribute $h \in [K]$ and a set of attributes $\mathcal{J} \subseteq [K]\setminus\{h\}$, such that $\vee_{j\in\mathcal{J}}\, \bar{\boldsymbol{q}}_j \not\succeq \bar{\boldsymbol{q}}_h$; and that there exists $S \subseteq \{K+1,\dots,J\}$ such that $\max_{m\in S} q_{m,h} = 0$ and $\max_{m\in S} q_{m,j} = 1$. Define

$$\boldsymbol{\theta}^\star = \bar{c}_h \boldsymbol{e}_h + \sum_{j\in\mathcal{J}} \bar{g}_j \boldsymbol{e}_j + \sum_{m=K+1}^{J} g_m \boldsymbol{e}_m, \quad \boldsymbol{r}^\star = \boldsymbol{e}_h + \sum_{j\in\mathcal{J}} \boldsymbol{e}_j + \sum_{m=K+1}^{J} \boldsymbol{e}_m,$$

and we claim that $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)$ is an all-zero vector. This is because for any $\boldsymbol{\alpha} \in \{0,1\}^K$, the corresponding element in $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)$ contains a factor $F_{\boldsymbol{\alpha}} = (\bar{\theta}_{h,\boldsymbol{\alpha}} - \bar{c}_h)\prod_{j\in\mathcal{J}}(\bar{\theta}_{j,\boldsymbol{\alpha}} - \bar{g}_j)$. While this factor $F_{\boldsymbol{\alpha}} \neq 0$ only if $\bar{\theta}_{h,\boldsymbol{\alpha}} = \bar{g}_h$ and $\bar{\theta}_{j,\boldsymbol{\alpha}} = \bar{c}_j$ for all $j \in \mathcal{J}$, which happens if and only if $\boldsymbol{\alpha} \not\succeq \bar{\boldsymbol{q}}_h$ and $\boldsymbol{\alpha} \succeq \bar{\boldsymbol{q}}_j$ for all $j \in \mathcal{J}$, which is impossible because $\vee_{j\in\mathcal{J}}\bar{\boldsymbol{q}}_j \succeq \bar{\boldsymbol{q}}_h$ by our assumption. So the claim $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star) = \boldsymbol{0}$ is proved, and further $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = 0$. Equality (E.3) becomes $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = 0$, which leads to

$$0 = T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)\boldsymbol{p} = p_{\boldsymbol{1}}(c_h - \bar{c}_h)\prod_{j\in\mathcal{J}}(c_j - \bar{g}_j)\prod_{m>K}(c_m - g_m),$$

which is because for any $\boldsymbol{\alpha} \neq \boldsymbol{1}$, we must have $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_m$ for some $m > K$ under Condition C, and hence the element $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)$ contains a factor $(g_m - g_m) = 0$. Since $c_m - g_m > 0$ for $m > K$ and $c_j - \bar{g}_j \neq 0$, we obtain $c_h = \bar{c}_h$.

We remark here that $c_h = \bar{c}_h$ also implies $\bar{\boldsymbol{q}}_h \neq \boldsymbol{0}$, because otherwise we would have $\bar{\theta}_h = \bar{c}_h = c_h$, which contradicts the $g_h < \bar{\theta}_h < c_h$ proved before the current Step 1. This indicates the $\bar{Q}_{1:K,\boldsymbol{\cdot}}$ can not contain any all-zero row vector, because otherwise $\bar{\boldsymbol{q}}_j \succeq \bar{\boldsymbol{q}}_h$ for the all-zero row vector $\bar{\boldsymbol{q}}_h$, which we showed is impossible.

Consider the item set S in the lemma that satisfies $S \subseteq \{K+1, \dots, J\}$ such that $\max_{m \in S} q_{m,h} = 0$ and $\max_{m \in S} q_{m,j} = 1$ for all $j \in \mathcal{J}$. Define

$$\boldsymbol{\theta}^\star = \bar{c}_h \boldsymbol{e}_h + \sum_{j \in \mathcal{J}} \bar{g}_j \boldsymbol{e}_j + \sum_{m \in S} g_m \boldsymbol{e}_m.$$

Note that $c_h = \bar{c}_h$. The RHS of (E.3) is zero, and so is the LHS of it. The row vector $T_{\boldsymbol{r}^\star, \boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)$ has the following property

$$\begin{aligned}
&T_{\boldsymbol{r}^\star, \boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star) \\
&= \begin{cases} (g_h - \bar{c}_h) \prod_{j \in \mathcal{J}} (c_j - \bar{g}_j) \prod_{m \in S} (c_m - g_m), & \boldsymbol{\alpha} \not\succeq \boldsymbol{q}_h,\ \boldsymbol{\alpha} \succeq \boldsymbol{q}_{\mathcal{J}},\ \boldsymbol{\alpha} \succeq \boldsymbol{q}_S; \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

An important observation is that $\{\boldsymbol{\alpha} \in \{0,1\}^K : \boldsymbol{\alpha} \not\succeq \boldsymbol{q}_h,\ \boldsymbol{\alpha} \succeq \boldsymbol{q}_{\mathcal{J}},\ \boldsymbol{\alpha} \succeq \boldsymbol{q}_S\} = \mathcal{A} \neq \varnothing$. This is because $q_{S,h} = 0$ and $q_{S,j} = 1$ for all $j \in \mathcal{J}$ hold, and we can just choose $\boldsymbol{\alpha}$ for which $\alpha_h = 0$ and $\alpha_k = 1$ for all $q_{S,k} = 1$, then such $\boldsymbol{\alpha}$ belongs to the set $\mathcal{A}$. Therefore we have

$$\begin{aligned}
&T_{\boldsymbol{r}^\star, \boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star) \boldsymbol{p} \\
&= (g_h - \bar{c}_h) \prod_{j \in \mathcal{J}} (c_j - \bar{g}_j) \prod_{m \in S} (c_m - g_m) \Big(\sum_{\boldsymbol{\alpha} \in \mathcal{A}} p_{\boldsymbol{\alpha}} \Big) = 0,
\end{aligned}$$

which leads to a contradiction since $g_h - \bar{c}_h \neq 0$, $c_j - \bar{g}_j \neq 0$, $c_m - g_m \neq 0$ and $\sum_{\boldsymbol{\alpha} \in \mathcal{A}} p_{\boldsymbol{\alpha}} > 0$, i.e., every factor in the above product is nonzero. This completes the proof of Lemma C.2. $\square$

We now proceed with the proof of Step 1 using an induction argument. We first introduce the definition of *lexicographic order* between two binary vectors of the same length. Specifically, for two binary vectors $\boldsymbol{a} = (a_1, \dots, a_L)^\top$ and $\boldsymbol{b} = (b_1, \dots, b_L)^\top$ both of length L, we say $\boldsymbol{a}$ is of smaller lexicographic order than $\boldsymbol{b}$ and denote $\boldsymbol{a} \prec_{\text{lex}} \boldsymbol{b}$, if either $a_1 < b_1$, or there exists a integer $l \in \{2, \ldots, L\}$ such that $a_l < b_l$ and $a_m = b_m$ for all $m = 1, \ldots, l-1$. It is not hard to see when Condition B that $Q^\star$ contains K distinct column vectors is satisfied, the K columns of $Q^\star$ can be arranged in an increasing lexicographic order. Namely, under Condition B, there exists a permutation map $\sigma(\cdot) : [K] \to [K]$ such that

$$Q^\star_{,\sigma(1)} \prec_{\text{lex}} Q^\star_{,\sigma(2)} \prec_{\text{lex}} \cdots \prec_{\text{lex}} Q^\star_{,\sigma(K)}. \tag{C.6}$$

Without loss of generality, next we consider the case where $\sigma(\cdot)$ is the identity map, i.e., $\sigma(k) = k$ for all $k \in [K]$.

We first consider attribute 1. Since $Q^\star_{,1}$ has the smallest lexicographic order among the columns of $Q^\star$, we have the conclusion that there must exist an item set $S \subseteq \{K+1, \ldots, J\}$ such that

$$q_{S,1} = 0, \quad q_{S,\ell} = 1 \; \forall \ell = 2, \ldots, K.$$

We apply Lemma C.2 to obtain $\vee_{\ell \in \{2,\ldots,K\}} \bar{\boldsymbol{q}}_\ell \not\succeq \bar{\boldsymbol{q}}_1$, which means

$$\begin{aligned}
&\big(\max_{m \in \{2,\ldots,K\}} \bar{q}_{\ell,1}, \max_{m \in \{2,\ldots,K\}} \bar{q}_{\ell,2}, \ldots, \max_{m \in \{2,\ldots,K\}} \bar{q}_{\ell,K}\big) \\
&\not\succeq (\bar{q}_{1,1}, \;\ldots,\; \bar{q}_{1,K}).
\end{aligned}$$

This implies there must exist an attribute $m_1 \in [K]$ such that

$$\max_{k \in [K]\setminus\{1\}} \bar{q}_{k,m_1} = 0, \quad \bar{q}_{1,m_1} = 1, \tag{C.7}$$

which exactly says the m_1-th column vector of $\bar{Q}_{1:K,\boldsymbol{\cdot}}$ must equal the basis vector $(\underbrace{1}_{\text{column 1}}, \mathbf{0})^\top = \boldsymbol{e}_1$, i.e., we have $\bar{Q}_{1:K,m_1} = \boldsymbol{e}_1$.

Now we assume as the inductive hypothesis that for $h \in [K]$ and $h > 1$, we have a distinct set of attributes $\{m_1, \ldots, m_{h-1}\} \subseteq [K]$ such that their corresponding column vectors in $\bar{Q}_{1:K,\bullet}$ satisfy

$$\forall i = 1, \ldots, h-1, \quad \bar{Q}_{1:K,m_i} = (*, \ldots, *, \underbrace{1}_{\text{column } i}, 0, \ldots, 0)^\top. \tag{C.8}$$

Now we focus on attribute h. By (E.5), the column vector $Q^\star_{,h}$ has the smallest lexicographic order among the $K-h-1$ columns in $\{Q^\star_{\bullet,h},\, Q^\star_{\bullet,h+1},\, \ldots,\, Q^\star_{\bullet,K}\}$, therefore similar to the argument in the previous paragraph, there must exist an item set $S \subseteq \{K+1, \ldots, J\}$ such that

$$q_{S,h} = 0, \quad q_{S,\ell} = 1 \; \forall \ell = h+1, \ldots, K. \tag{C.9}$$

Therefore Lemma C.2 implies $\vee_{\ell\in\{h+1,\ldots,K\}} \bar{\boldsymbol{q}}_\ell \nsucceq \bar{\boldsymbol{q}}_1$, and further leads to

$$\max_{\ell\in\{h+1,\ldots,K\}} \bar{q}_{\ell,m_h} = 0, \quad \bar{q}_{h,m_h} = 1. \tag{C.10}$$

We point out that $m_h \notin \{m_1, \ldots, m_{h-1}\}$, because by the induction hypothesis (E.6) we have $\bar{q}_{h,m_i} = 0$ for $i = 1, \ldots, h-1$. So $\{m_1, \ldots, m_{h-1}, m_h\}$ contains h distinct attributes. Furthermore, (E.8) gives that $\bar{Q}_{\bullet,m_h} = (*, \ldots, *, \underbrace{1}_{\text{column } h}, 0, \ldots, 0)^\top$, which generalizes (E.6) by extending $h-1$ there to h. Therefore, we use the induction argument to obtain

$$\forall k \in \{1, \ldots, K-1\}, \quad \bar{Q}_{1:K,m_k} = (*, \ldots, *, \underbrace{1}_{\text{column } k}, 0, \ldots, 0)^\top.$$

Furthermore, when considering the last attribute K, the Kth item must have $\boldsymbol{q}$-vector taking the form of $\bar{\boldsymbol{q}}_K = (0, \ldots, 0, \underbrace{*}_{\text{column } m_K}, 0, \ldots, 0)$, where the "$*$" in $\bar{\boldsymbol{q}}_K$ is the only element unspecified. Since previously we have shown in the proof of Lemma C.2 that $\bar{\boldsymbol{q}}_j = \mathbf{0}$ can not happen for any item j, there must be $\bar{\boldsymbol{q}}_K =$

$(0, \ldots, 0, \underbrace{1}_{\text{column } m_K}, 0, \ldots, 0)$. Now we have essentially obtained

$$\bar{Q}_{1:K,\,(m_1,\ldots,m_K)} = \begin{pmatrix} 1 & * & \ldots & * \\ 0 & 1 & \ldots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \tag{C.11}$$

and the conclusion of Step 1 in (C.5) is proved.

Step 2. In this step we prove $c_j = \bar{c}_j$ for $j = K+1, \ldots, J$. For an arbitrary item $j \in \{K+1, \ldots, J\}$, define a response vector $\boldsymbol{r}^* = \sum_{h:\, h \neq j} \boldsymbol{e}_j$ and

$$\boldsymbol{\theta}^* = \sum_{h=1}^{K} \bar{g}_h \boldsymbol{e}_h + \sum_{h>K,\, h\neq j} g_h \boldsymbol{e}_h.$$

We claim that $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)$ contains only one nonzero element corresponding to the all-one attribute pattern $\boldsymbol{\alpha} = \mathbf{1}$. The reasoning is as follows. Under the conclusion of Step 1, $\bar{Q}_{1:K,\boldsymbol{\cdot}}$ takes the form of (C.5), which means each attribute is required by at least one item in $\{\bar{\boldsymbol{q}}_1, \ldots, \bar{\boldsymbol{q}}_K\}$. Then for any $\boldsymbol{\alpha} \neq \mathbf{1}$, there must exist some attribute $k \in [K]$ such that $\boldsymbol{\alpha} \not\succeq \bar{\boldsymbol{q}}_k$, which implies for this particular $\boldsymbol{\alpha}$ the element $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)$ contains a factor $(\bar{g}_h - \bar{g}_h) = 0$. Therefore $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*) \neq 0$ only if $\boldsymbol{\alpha} = \mathbf{1}$. Next consider $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)$. Under Condition A, in the true Q each attribute is required by at least three items, so the row vector corresponding to response pattern $\boldsymbol{r}^*$ in $T(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)$ only contains one nonzero element, in column $\boldsymbol{\alpha} = \mathbf{1}_K^\top$, representing the attribute profile mastering all the K attributes. This is because for any other attribute profile $\boldsymbol{\alpha}'$ that lacks at least one attribute k, there must be some item $h > K$, $h \neq j$ requiring attribute k so that $\boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_h$; and this results in $\theta_{\boldsymbol{e}_h,\boldsymbol{\alpha}'} = g_h$ and $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}'}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) = 0$. In summary,

$$T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}(Q,\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{g}-\boldsymbol{\theta}^*)=\prod_{h=1}^{K}(\theta_{h,\boldsymbol{\alpha}}-\bar{g}_h)\prod_{\substack{h>K:\\ h\neq j}}(\theta_{h,\boldsymbol{\alpha}}-g_h)\neq 0\quad\text{iff}\quad \boldsymbol{\alpha}=\mathbf{1};$$

$$T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}(\bar{Q},\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{g}}-\boldsymbol{\theta}^*)=\prod_{h=1}^{K}(\bar{\theta}_{h,\boldsymbol{\alpha}}-\bar{g}_h)\prod_{\substack{h>K:\\ h\neq j}}(\bar{\theta}_{h,\boldsymbol{\alpha}}-g_h)\neq 0\quad\text{iff}\quad \boldsymbol{\alpha}=\mathbf{1}.$$

Now further consider item j. Since $\mathbf{1}_K^\top\succeq\boldsymbol{q}_j$ and $\mathbf{1}_K^\top\succeq\bar{\boldsymbol{q}}_j$, one must have $\theta_{j,\mathbf{1}_K^\top}=c_j$ and $\bar{\theta}_{j,\mathbf{1}_K^\top}=\bar{c}_j$. Since we assume $p_{\boldsymbol{\alpha}}>0$ for each $\boldsymbol{\alpha}$, we have $T_{\boldsymbol{r}^*,\bullet}(Q,\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{g}-\boldsymbol{\theta}^*)\boldsymbol{p}=T_{\boldsymbol{r}^*,\mathbf{1}_K^\top}(Q,\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{g}-\boldsymbol{\theta}^*)p_{\mathbf{1}_K^\top}\neq 0$. So (C.2) in Lemma C.1 implies that

$$c_j=\frac{T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\bullet}(Q,\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{g}-\boldsymbol{\theta}^*)\boldsymbol{p}}{T_{\boldsymbol{r}^*,\bullet}(Q,\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{g}-\boldsymbol{\theta}^*)\boldsymbol{p}}=\frac{T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\bullet}(\bar{Q},\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{g}}-\boldsymbol{\theta}^*)\bar{\boldsymbol{p}}}{T_{\boldsymbol{r}^*,\bullet}(\bar{Q},\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{g}}-\boldsymbol{\theta}^*)\bar{\boldsymbol{p}}}=\bar{c}_j.$$

In the above argument j is arbitrary, so $c_j=\bar{c}_j$ for any $j=K+1,\ldots,J$.

Step 3. In this step we prove $g_k=\bar{g}_k$ for $k=1,\ldots,K$. Recall that in Step 1 we showed that (E.5) about the lexicographic order holds and assumed $\sigma(k)=k$ for $k\in[K]$ without loss of generality. We now prove $g_1=\bar{g}_1$. Define

$$\boldsymbol{\theta}^*=\sum_{h=1}^{K}\bar{g}_h\boldsymbol{e}_h+\sum_{\substack{h>K:\\ q_{h,1}=0}}g_h\boldsymbol{e}_h+\sum_{\substack{h>K:\\ q_{h,1}=1}}c_h\boldsymbol{e}_h,\tag{C.12}$$

then

$$T_{\sum_h\boldsymbol{e}_h,\boldsymbol{\alpha}}(Q,\boldsymbol{\theta}^+-\boldsymbol{\theta}^*,\boldsymbol{g}-\boldsymbol{\theta}^*)=\prod_{h=1}^{K}(\theta_{h,\boldsymbol{\alpha}}-\bar{g}_h)\prod_{\substack{h>K:\\ q_{h,1}=0}}(\theta_{h,\boldsymbol{\alpha}}-g_h)\prod_{\substack{h>K:\\ q_{h,1}=1}}(\theta_{h,\boldsymbol{\alpha}}-c_h);$$

$$T_{\sum_h\boldsymbol{e}_h,\boldsymbol{\alpha}}(\bar{Q},\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{g}}-\boldsymbol{\theta}^*)=\prod_{h=1}^{K}(\bar{\theta}_{h,\boldsymbol{\alpha}}-\bar{g}_h)\prod_{\substack{h>K:\\ q_{h,1}=0}}(\bar{\theta}_{h,\boldsymbol{\alpha}}-g_h)\prod_{\substack{h>K:\\ q_{h,1}=1}}(\bar{\theta}_{h,\boldsymbol{\alpha}}-c_h).$$

First, the row vector $T_{\sum_{h=1}^J\boldsymbol{e}_h,\bullet}(\bar{Q},\bar{\boldsymbol{\theta}}^+-\boldsymbol{\theta}^*,\bar{\boldsymbol{g}}-\boldsymbol{\theta}^*)$ equals the zero vector. This is because $\bar{Q}_{1:K,\boldsymbol{\cdot}}$ takes the form in (C.5) by Step 1, and any attribute profile $\boldsymbol{\alpha} \neq \mathbf{1}_K^\top$ would have $\bar{\theta}_{h,\boldsymbol{\alpha}} = \bar{g}_h$ for some $h \in \{1, \ldots, K\}$, which makes the corresponding element in the above row vector zero. Furthermore, $T_{\sum_{h=1}^J \boldsymbol{e}_h, \mathbf{1}_K^\top}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)$ is also zero, because $\bar{\theta}_{h,\boldsymbol{\alpha}} = \bar{c}_h = c_h$ for those $h > K$ such that $q_{h,1} = 1$. Since $Q^\star_{\boldsymbol{\cdot},1}$ has the smallest lexicographic order among the columns of $Q^\star$, for any $k \in \{2, \ldots, K\}$, there must exist some item $h \in \{K+1, \ldots, J\}$ that requires attribute 1, as a result

$$\vee_{h>K:\, q_{h,1}=0}\, \boldsymbol{q}_h = (0, 1, \ldots, 1).$$

This ensures $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)$ would equal zero if $\boldsymbol{\alpha}$ lacks any attribute other than the first one. So the nonzero elements in the row vector $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)$ can only correspond to columns $\boldsymbol{\alpha}^1 = (0, 1, \ldots, 1)$ or $\boldsymbol{\alpha}^2 = \mathbf{1}_K^\top$. Further, we claim $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}^2}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) = 0$, this is because $\theta_{h,\boldsymbol{\alpha}} = c_h$ for those h such that $q_{h,1} = 1$. So the row vector $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)$ only contains one potentially nonzero element in column $\boldsymbol{\alpha}_1 = (0, 1, \ldots, 1)$ as follows

$$T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}_1}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) = (g_1 - \bar{g}_1) \prod_{h=2}^K (c_h - \bar{g}_h) \prod_{\substack{h>K:\\ q_{h,1}=0}} (c_h - g_h) \prod_{\substack{h>K:\\ q_{h,1}=1}} (g_h - c_h). \tag{C.13}$$

Using the fact $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*) = \mathbf{0}_{2^K}$, the equality

$$T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}^1}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)\boldsymbol{p} = T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}^1}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)\bar{\boldsymbol{p}} = 0$$

implies the element in (E.11) must also be zero. As shown earlier, $c_h - \bar{g}_h \neq 0$ for any h, so $g_1 = \bar{g}_1$ must hold.

Next we use an induction argument to prove that for $k = 2, \ldots, K$, $g_k = \bar{g}_k$. In particular, suppose for any $1 \le m \le k-1$, we already have $g_m = \bar{g}_m$. Define

$$\boldsymbol{\theta}^* = \sum_{h=1}^{K} \bar{g}_h \boldsymbol{e}_h + \sum_{h>K:\, q_{h,k}=0} g_h \boldsymbol{e}_h + \sum_{h>K:\, q_{h,k}=1} c_h \boldsymbol{e}_h. \tag{C.14}$$

For the similar reason as stated before, $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)$ equals the zero vector. We claim that the row vector $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)$ only contains one potentially nonzero element in column $\boldsymbol{\alpha}' := (1, \ldots, 1, \underbrace{0}_{\text{column } k}, 1, \ldots, 1)$. The reason is as follows. On the one hand, for any attribute profile $\boldsymbol{\alpha}$ that lacks some attribute $l \in \{k+1, \ldots, K\}$, due to the assumption in (E.5) that $Q^*_{\boldsymbol{\cdot},k} \prec_{\text{lex}} Q^*_{\boldsymbol{\cdot},l}$, there must exist some item $h > K$ such that $q_{h,k} = 0$, $q_{h,l} = 1$. So for this particular $\boldsymbol{\alpha}$ we have $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_h$, $\theta_{h,\boldsymbol{\alpha}} = g_h$, which makes $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) = 0$. On the other hand, for any attribute profile $\boldsymbol{\alpha}'$ that lacks some attribute $m \in \{1, \ldots, k-1\}$, one has $\boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_m = \boldsymbol{e}_m$ and $\theta_{m,\boldsymbol{\alpha}'} = g_m = \bar{g}_m$, where the last equality $g_m = \bar{g}_m$ comes from the induction assumption. This results in $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}'}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) = 0$ for all such $\boldsymbol{\alpha}'$. In conclusion, the nonzero elements in this transformed row vector can only be in columns $\boldsymbol{\alpha}'$ or $\boldsymbol{\alpha}_2 = \mathbf{1}_K^\top$. For similar reason as in proving $g_1 = \bar{g}_1$, $T_{\sum_{h=1}^J \boldsymbol{e}_h, \boldsymbol{\alpha}_2}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) = 0$. So the transformed row vector only contains one potentially nonzero entry corresponding to $\boldsymbol{\alpha}'$:

$$\begin{aligned}
&T_{\sum_h \boldsymbol{e}_h, \boldsymbol{\alpha}'}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) \\
&= (g_k - \bar{g}_k) \prod_{\substack{1 \le h \le K: \\ h \ne k}} (c_h - \bar{g}_h) \prod_{\substack{h>K: \\ q_{h,k}=0}} (c_h - g_h) \prod_{\substack{h>K: \\ q_{h,k}=1}} (g_h - c_h).
\end{aligned}$$

The same argument after (E.11) gives $g_k = \bar{g}_k$. In conclusion, the induction method yields $g_k = \bar{g}_k$ for $k = 1, \ldots, K$.

Step 4. In this step we show that $\bar{Q}_{1:K, \boldsymbol{\cdot}} \sim I_K$. Recall that in Step 1 we already obtained (E.9), and now we aim to show that the $\bar{Q}_{1:K, (m_1, \ldots, m_K)}$ in (E.9) can be further written as

$$\bar{Q}_{1:K,(m_1,\dots,m_K)} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

We now claim that $\bar{\boldsymbol{q}}_j \nsucceq \bar{\boldsymbol{q}}_h$ for any $1 \le j < h \le K$. If this claim is true, then $\bar{Q}_{1:K,(m_1,\dots,m_K)} = I_K$ must hold and the conclusion $\bar{Q}_{1:K,\boldsymbol{\cdot}} \sim I_K$ is reached. We next prove that claim by contradiction. If there exist some $1 \le j < h \le K$ such that $\bar{\boldsymbol{q}}_j \succeq \bar{\boldsymbol{q}}_h$, then define

$$\boldsymbol{\theta}^\star = \bar{c}_h \boldsymbol{e}_h + \bar{g}_j \boldsymbol{e}_j + \sum_{m=K+1}^{J} g_m \boldsymbol{e}_m,$$

we have

$$\begin{aligned} 0 &= T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} \\ &= T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)\boldsymbol{p} \\ &= p_{\boldsymbol{1}}(c_h - \bar{c}_h)(c_j - \bar{g}_j) \prod_{m=K+1}^{J} (c_m - g_m), \end{aligned}$$

which implies $c_h = \bar{c}_h$. Note that we have obtained $g_j = \bar{g}_j$ in Step 3, and we next define $\boldsymbol{\theta}^\star = \bar{c}_h \boldsymbol{e}_h + \bar{g}_j \boldsymbol{e}_j$. The equality $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = 0$ still holds and (E.3) gives

$$\begin{aligned} 0 &= T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)\boldsymbol{p} \\ &= (g_h - \bar{c}_h)(c_j - \bar{g}_j)\Big(\sum_{\boldsymbol{\alpha}:\, \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_h, \boldsymbol{\alpha} \succeq \boldsymbol{q}_j} p_{\boldsymbol{\alpha}}\Big) \\ &= (g_h - c_h)(c_j - g_j)\Big(\sum_{\boldsymbol{\alpha}:\, \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_h, \boldsymbol{\alpha} \succeq \boldsymbol{q}_j} p_{\boldsymbol{\alpha}}\Big). \end{aligned} \tag{C.15}$$

Since $Q_{1:K,\boldsymbol{\cdot}} = I_K$, we have that $\boldsymbol{q}_j$ and $\boldsymbol{q}_h$ in the true Q are distinct basis vectors, therefore $\left(\sum_{\boldsymbol{\alpha}:\, \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_h, \boldsymbol{\alpha} \succeq \boldsymbol{q}_j} p_{\boldsymbol{\alpha}}\right) > 0$. Therefore (C.15) leads to a contradiction, and we have proved the claim that $\bar{\boldsymbol{q}}_j \nsucceq \bar{\boldsymbol{q}}_h$ for any $1 \leq j < h \leq K$. As stated earlier, this claim naturally leads to the conclusion of Step 3 that $\bar{Q}_{1:K,\boldsymbol{\cdot}} \sim I_K$.

Step 5. In this step we prove that after reordering the columns in $\bar{Q}$ such that $\bar{Q}_{1:K} = I_K$, we must have $\boldsymbol{q}_j = \bar{\boldsymbol{q}}_j$ for $j = K+1, \ldots, J$. In the following two parts, we first prove $\bar{\boldsymbol{q}}_j \succeq \boldsymbol{q}_j$ for all $j \in \{K+1, \ldots, J\}$ in part (a); and then prove $\bar{\boldsymbol{q}}_j = \boldsymbol{q}_j$ for all $j \in \{K+1, \ldots, J\}$ in part (b).

(a) We next show $\bar{\boldsymbol{q}}_j \succeq \boldsymbol{q}_j$ for all $j \in \{K+1, \ldots, J\}$. We use proof by contradiction, and assume $\bar{\boldsymbol{q}}_j \nsucceq \boldsymbol{q}_j$ for some $j \in \{K+1, \ldots, J\}$. Then $\{\boldsymbol{\alpha} : \boldsymbol{\alpha} \succeq \bar{\boldsymbol{q}}_j, \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_j\} = \mathcal{A} \neq \varnothing$ and $\sum_{\boldsymbol{\alpha} \in \mathcal{A}} p_{\boldsymbol{\alpha}} \neq 0$. Define

$$\boldsymbol{\theta}^* = \sum_{k \in [K] :\, \bar{q}_{j,k} = 1} g_k \boldsymbol{e}_k + c_j \boldsymbol{e}_j, \tag{C.16}$$

then $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*) = \boldsymbol{0}$ and $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)\bar{\boldsymbol{p}} = 0$. However, for any $\boldsymbol{\alpha} \in \mathcal{A}$, one has $\theta_{j,\boldsymbol{\alpha}} = g_j$ and $\theta_{k,\boldsymbol{\alpha}} = c_k$ for any k s.t. $\bar{q}_{j,k} = 1$, so for any $\boldsymbol{\alpha} \in \mathcal{A}$ we have

$$\begin{aligned} T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*) &= \prod_{\substack{1 \leq k \leq K: \\ q_{j,k} = 1}} (\theta_{k,\boldsymbol{\alpha}} - g_k)(\theta_{j,\boldsymbol{\alpha}} - c_j) \\ &= \prod_{\substack{1 \leq k \leq K: \\ q_{j,k} = 1}} (c_k - g_k)(g_j - c_j) \neq 0, \end{aligned}$$

and hence

$$\begin{aligned} T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)\boldsymbol{p} &= \prod_{\substack{1 \leq k \leq K: \\ q_{j,k} = 1}} (c_k - g_k)(g_j - c_j) \sum_{\boldsymbol{\alpha} \in \mathcal{A}} p_{\boldsymbol{\alpha}} \\ &\neq 0 = T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)\bar{\boldsymbol{p}}, \end{aligned}$$

which contradicts (C.3).

(b) Based on (a), we next show $\bar{\boldsymbol{q}}_j = \boldsymbol{q}_j$ for all $j \in \{K+1, \ldots, J\}$ using proof by contradiction. Since part (a) gives $\bar{\boldsymbol{q}}_j \succeq \boldsymbol{q}_j$, if $\bar{\boldsymbol{q}}_j \neq \boldsymbol{q}_j$, then there must exist some attribute $k \in [K]$ such that $\bar{q}_{j,k} = 1$ and $q_{j,k} = 0$. This implies $\bar{\boldsymbol{q}}_j \succeq \bar{\boldsymbol{q}}_k$. Define

$$\boldsymbol{\theta}^\star = \bar{c}_k \boldsymbol{e}_k + \bar{g}_j \boldsymbol{e}_j + \sum_{m>K:\, m\neq j} g_m \boldsymbol{e}_m,$$

then $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = 0$. Since Condition C holds, each attribute is required by at least one item in the set $\{m > K : m \neq j\}$, which implies $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star) \neq 0$ only if $\boldsymbol{\alpha} = \mathbf{1}$. Therefore (E.3) gives that

$$\begin{aligned}
0 &= T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)\boldsymbol{p} \\
&= (c_k - \bar{c}_k)(c_j - \bar{g}_j) \prod_{m>K:\, m\neq j} (c_m - g_m) p_{\mathbf{1}},
\end{aligned}$$

so $c_k = \bar{c}_k$. Now we further define

$$\boldsymbol{\theta}^\star = \bar{c}_k \boldsymbol{e}_k + \bar{g}_j \boldsymbol{e}_j + \sum_{h\in[K]\setminus\{k\}} g_m \boldsymbol{e}_m,$$

then $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = 0$. However, $\boldsymbol{q}_j \not\succeq \boldsymbol{q}_k$ under the true Q, and (E.3) gives

$$T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{g} - \boldsymbol{\theta}^\star)\boldsymbol{p} = (g_k - \bar{c}_k) \prod_{h\in[K]\setminus\{k\}} (c_h - g_h)(c_j - \bar{g}_j) p_{\boldsymbol{\alpha}-\boldsymbol{e}_k},$$

where $\boldsymbol{\alpha} - \boldsymbol{e}_k = (\mathbf{1}, \underbrace{0}_{\text{column } k}, \mathbf{1})$, so the above display is nonzero. This contradicts (E.3), and this means $\bar{\boldsymbol{q}}_j \neq \boldsymbol{q}_j$ can not happen. So we have $\bar{\boldsymbol{q}}_j = \boldsymbol{q}_j$ for $j \in \{K+1, \ldots, J\}$.

Now we have proved $Q \sim \bar{Q}$. Now that $Q \sim \bar{Q}$, Theorem 1 in Gu and Xu (2019b) (Chapter II) gives that Conditions A and B ensure the identifiability of the model parameters ($\boldsymbol{s} := \mathbf{1} - \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}$). This concludes the proof of the sufficiency of the conditions.

In the end we show the necessity of the conditions. By Theorem 1 in Gu and Xu (2019b), Conditions A and B are necessary for identifiability of the model parameters $(\boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ given a known Q, so they are also necessary for identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$.

C.3 Proof of Theorem IV.2

Proof of the necessity of each attribute required by ≥ 2 items. Suppose Q takes the form of

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & Q^\star \end{pmatrix},$$

then for any valid $(\boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ associated with Q, we next construct $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}}) \neq (\boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ such that $T(Q, \boldsymbol{\theta}^+, \boldsymbol{g})\boldsymbol{p} = T(Q, \bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{g}})\bar{\boldsymbol{p}}$ holds. In particular, we arbitrarily choose $\bar{c}_1$ that is not equal to $c_1 = 1 - s_1$ and set

$$\bar{p}_{\boldsymbol{\alpha}} = \begin{cases} (c_1/\bar{c}_1)p_{\boldsymbol{\alpha}}, & \text{if } \alpha_1 = 1, \\ p_{\boldsymbol{\alpha}} + (1 - c_1/\bar{c}_1)p_{\boldsymbol{\alpha}+\boldsymbol{e}_1}, & \text{if } \alpha_1 = 0. \end{cases}$$

Then set $\bar{g}_1 = g_1$, and $\bar{c}_j = c_j$, $\bar{g}_j = g_j$ for $j = 2, \ldots J$. Then it is not hard to check that $T(Q, \boldsymbol{\theta}^+, \boldsymbol{g})\boldsymbol{p} = T(Q, \bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{g}})\bar{\boldsymbol{p}}$. Since $(\boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ are arbitrary, we have shown the non-identifiability set spans the entire parameter space and $(Q, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ are not generically identifiable. Therefore, this proves that $(Q, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ are not generically identifiable if some attribute is required by only one item.

In the following we prove part (a), (b), and (c) when some attribute is required by only two items.

Proof of Part (a). Under the assumption of part (a), Q takes the form

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{1}^\top \\ \mathbf{0} & Q^\star \end{pmatrix}.$$

Given arbitrary DINA model parameters $(\boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ under this Q, we next construct another different set of DINA parameters $(\bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{g}}, \bar{\boldsymbol{p}}) \neq (\boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ also associated with this Q, such that

$$T(Q, \boldsymbol{\theta}^+, \boldsymbol{g})\boldsymbol{p} = T(Q, \bar{\boldsymbol{\theta}}^+, \bar{\boldsymbol{g}})\bar{\boldsymbol{p}}. \tag{C.17}$$

In particular, we set $\bar{c}_j = c_j$ and $\bar{g}_j = g_j$ for all $j = 3, \ldots, J$. Under this construction, (C.17) simplifies to the following two sets of equations

$$\forall \boldsymbol{\alpha}' \in \{0,1\}^{K-1},\ \boldsymbol{\alpha}' \neq \mathbf{1}, \quad \begin{cases} p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} = \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}, \\ g_1 p_{(0,\boldsymbol{\alpha}')} + c_1 p_{(1,\boldsymbol{\alpha}')} = \bar{g}_1 \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{c}_1 \bar{p}_{(1,\boldsymbol{\alpha}')}, \\ g_2 [p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')}] = \bar{g}_2 [\bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}], \\ g_2 [g_1 p_{(0,\boldsymbol{\alpha}')} + c_2 p_{(1,\boldsymbol{\alpha}')}] = \bar{g}_2 [\bar{g}_1 \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{c}_1 \bar{p}_{(1,\boldsymbol{\alpha}')}]; \end{cases} \tag{C.18}$$

and for $\boldsymbol{\alpha}' = \mathbf{1}$,

$$\begin{cases} p_{(0,\mathbf{1})} + p_{(1,\mathbf{1})} = \bar{p}_{(0,\mathbf{1})} + \bar{p}_{(1,\mathbf{1})}, \\ g_1 p_{(0,\mathbf{1})} + c_1 p_{(1,\mathbf{1})} = \bar{g}_1 \bar{p}_{(0,\mathbf{1})} + \bar{c}_1 \bar{p}_{(1,\mathbf{1})}, \\ g_2 p_{(0,\mathbf{1})} + c_2 p_{(1,\mathbf{1})} = \bar{g}_2 \bar{p}_{(0,\mathbf{1})} + \bar{c}_2 \bar{p}_{(1,\mathbf{1})}, \\ g_1 g_2 p_{(0,\mathbf{1})} + c_1 c_2 p_{(1,\mathbf{1})} = \bar{g}_1 \bar{g}_2 \bar{p}_{(0,\mathbf{1})} + \bar{c}_1 \bar{c}_2 \bar{p}_{(1,\mathbf{1})}. \end{cases} \tag{C.19}$$

The above (C.18) obviously leads to $\bar{g}_2 = g_2$, and the last two equations of (C.18) are automatically satisfied if the first two of (C.18) are satisfied. Then the last two equations of (C.19) can be transformed to

$$\begin{cases} (c_2 - g_2)p_{(1,\mathbf{1})} = (\bar{c}_2 - g_2)\bar{p}_{(1,\mathbf{1})}, \\ c_1(c_2 - g_2)p_{(1,\mathbf{1})} = \bar{c}_1(\bar{c}_2 - g_2)\bar{p}_{(1,\mathbf{1})}; \end{cases}$$

which gives $\bar{c}_1 = c_1$. Additionally, when $\bar{c}_1 = c_1$, we also have that the last equality of (C.19) holds as long as the first three equalities of (C.19) hold. In summary, now there are 2^K+2 parameters to be determined, which are $\{\bar{g}_1, \bar{c}_2\} \cup \{\bar{p}_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K\}$, while they only have to satisfy the following $2 \times (2^{K-1} - 1) + 3 = 2^K + 1$ constraints,

$$\forall \boldsymbol{\alpha}' \in \{0,1\}^{K-1}, \text{ for } \boldsymbol{\alpha}' \neq \mathbf{1}, \quad \begin{cases} p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} = \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}, \\ g_1 p_{(0,\boldsymbol{\alpha}')} + c_1 p_{(1,\boldsymbol{\alpha}')} = \bar{g}_1 \bar{p}_{(0,\boldsymbol{\alpha}')} + c_1 \bar{p}_{(1,\boldsymbol{\alpha}')}; \end{cases}$$

$$\text{and for } \boldsymbol{\alpha}' = \mathbf{1}, \quad \begin{cases} p_{(0,\mathbf{1})} + p_{(1,\mathbf{1})} = \bar{p}_{(0,\mathbf{1})} + \bar{p}_{(1,\mathbf{1})}, \\ g_1 p_{(0,\mathbf{1})} + c_1 p_{(1,\mathbf{1})} = \bar{g}_1 \bar{p}_{(0,\mathbf{1})} + c_1 \bar{p}_{(1,\mathbf{1})}, \\ g_2 p_{(0,\mathbf{1})} + c_2 p_{(1,\mathbf{1})} = g_2 \bar{p}_{(0,\mathbf{1})} + \bar{c}_2 \bar{p}_{(1,\mathbf{1})}. \end{cases}$$

Since the number of free variables $2^K + 2$ is greater than the number of constraints 2^K+1, there exist infinitely many different solutions to the above system of equations. This means that the $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are not generically identifiable. In particular, one can arbitrarily choose $\bar{g}_1$ close to but not equal to g_1, then solve for the remaining parameters $\{\bar{p}_{\boldsymbol{\alpha}},\ \boldsymbol{\alpha}\in\{0,1\}^K\}$ and $\bar{c}_2$ as follows,

$$\forall\boldsymbol{\alpha}'\in\{0,1\}^{K-1},\quad \begin{cases}\bar{p}_{(0,\boldsymbol{\alpha}')}=p_{(0,\boldsymbol{\alpha}')}(g_1-c_1)/(\bar{g}_1-c_1),\\[2ex] \bar{p}_{(1,\boldsymbol{\alpha}')}=p_{(0,\boldsymbol{\alpha}')}+p_{(1,\boldsymbol{\alpha}')}-\bar{p}_{(0,\boldsymbol{\alpha}')};\end{cases}$$

$$\bar{c}_2=\frac{g_2[p_{(0,\mathbf{1})}-\bar{p}_{(0,\mathbf{1})}]+c_2p_{(1,\mathbf{1})}}{\bar{p}_{(1,\mathbf{1})}}.$$

This concludes the proof of part (a) of the theorem.

Next we first prove (b.2), i.e. when $Q^\star$ has two submatrices $\mathcal{I}_{K-1}$. In this case, the Q contains a submatrix of the form $(I_K,I_K)^\top$. The proof of (b.1), i.e. when $Q^\star$ satisfies Conditions A, B and C, is combined with the proof of part (c) later.

Proof of Part (b.2). We first give the proof when Q only consists of two I_K's, namely $Q=(I_K,I_K)^\top$. In this case, we first prove that $\bar{Q}\sim Q$ must hold, using an argument similar to Step 1 of the proof of Theorem IV.1. Suppose $T(Q,\boldsymbol{\theta}^+,\boldsymbol{g})\boldsymbol{p}=T(\bar{Q},\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{g}})\bar{\boldsymbol{p}}$. Since $Q_{(K+1):(2K),\bullet}=I_K$, we have that for each attribute $h\in[K]$, there is

$$\max_{\substack{m\in\{K+1,\dots,2K\},\\ m\neq K+h}} q_{m,h}=0,\qquad \max_{m\in\{K+1,\dots,2K\}} q_{m,k}=1\ \forall k\in[K]\setminus\{h\}.$$

Therefore we can apply Lemma C.2 with $S=\{K+1,\dots,2K\}\setminus\{K+h\}$ and $\mathcal{J}=[K]\setminus\{h\}$ to obtain

$$\max_{k\in\mathcal{J}}\bar{\boldsymbol{q}}_k\not\succeq\bar{\boldsymbol{q}}_h.$$

This essentially implies that for an arbitrary $h\in[K]$, there must be a $m_h\in[K]$ such that $\bar{q}_{h,m_h}=0$ and $\bar{q}_{k,m_h}=0$ for all $k\in[K]\setminus\{h\}$. Moreover, the K integers $m_1,m_2,\dots,m_K$ must all be distinct, otherwise it is easy to see $\max_{k\in\mathcal{J}}\bar{\boldsymbol{q}}_k\not\succeq\bar{\boldsymbol{q}}_h$ would fail to hold for some $h\in[K]$. So $(m_1,m_2,\dots,m_K)$ is a permutation of $(1,2,\dots,K)$. Now we have obtained that $\bar{Q}_{1:K,(m_1,\dots,m_K)}$ must be an identity matrix, i.e., $\bar{Q}_{1:K,\bullet}\sim Q_{1:K,\bullet}$. Reasoning in exactly the same way gives $\bar{Q}_{(K+1):(2K),\bullet}\sim Q_{(K+1):(2K),\bullet}$, and we have $\bar{Q} \sim Q$. Now for an arbitrary $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K) \equiv (\alpha_1, \boldsymbol{\alpha}')$, define

$$\begin{aligned}\boldsymbol{\theta}^* =&\bar{g}_1\boldsymbol{e}_1 + \bar{c}_{K+1}\boldsymbol{e}_{K+1} + \sum_{k>1:\,\alpha_k=1} g_k\boldsymbol{e}_k + \sum_{k>1:\,\alpha_k=0} c_k\boldsymbol{e}_k \\ \equiv&\bar{g}_1\boldsymbol{e}_1 + \bar{c}_{K+1}\boldsymbol{e}_{K+1} + \boldsymbol{\theta}^{\boldsymbol{\alpha}}\end{aligned}$$

then $T_{\boldsymbol{e}_1+\boldsymbol{e}_{K+1}}(Q, \bar{\boldsymbol{s}} - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*) = \boldsymbol{0}$, so

$$\begin{aligned}0 =&T_{\boldsymbol{e}_1+\boldsymbol{e}_{K+1}}(Q, \bar{\boldsymbol{s}} - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)\bar{\boldsymbol{p}} = T_{\boldsymbol{e}_1+\boldsymbol{e}_{K+1}}(Q, \boldsymbol{s} - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)\boldsymbol{p} \\ =& \prod_{k>1:\,\alpha_k=1}(c_k - g_k) \times \prod_{k>1:\,\alpha_k=0}(g_k - c_k)\times \\ &\Big[(g_1 - \bar{g}_1)(g_{K+1} - \bar{c}_{K+1})p_{(0,\alpha_2,\ldots,\alpha_K)} + (c_1 - \bar{g}_1)(c_{K+1} - \bar{c}_{K+1})p_{(1,\alpha_2,\ldots,\alpha_K)}\Big].\end{aligned}$$

This implies that for any $\boldsymbol{\alpha}' = (\alpha_2, \ldots, \alpha_K) \in \{0, 1\}^{K-1}$, we have

$$(g_1 - \bar{g}_1)(g_{K+1} - \bar{c}_{K+1})p_{(0,\alpha_2,\ldots,\alpha_K)} + (c_1 - \bar{g}_1)(c_{K+1} - \bar{c}_{K+1})p_{(1,\alpha_2,\ldots,\alpha_K)} = 0.$$

Since $g_{K+1} - \bar{c}_{K+1} \neq 0$, we have that

$$g_1 - \bar{g}_1 = \frac{(c_1 - \bar{g}_1)(c_{K+1} - \bar{c}_{K+1})p_{(1,\boldsymbol{\alpha}')}}{(\bar{c}_{K+1} - g_{K+1})p_{(0,\boldsymbol{\alpha}')}}, \quad \text{for any } \boldsymbol{\alpha}' \in \{0, 1\}^{K-1}.$$

This equality indicates that if there exists $\boldsymbol{\alpha}'_1 \neq \boldsymbol{\alpha}'_2$ such that

$$\frac{p_{(1,\boldsymbol{\alpha}'_1)}}{p_{(0,\boldsymbol{\alpha}'_1)}} \neq \frac{p_{(1,\boldsymbol{\alpha}'_2)}}{p_{(0,\boldsymbol{\alpha}'_2)}}, \tag{C.20}$$

then one must have

$$c_{K+1} - \bar{c}_{K=1} = 0, \quad g_1 - \bar{g}_1 = 0.$$

Redefine $\boldsymbol{\theta}^* = \bar{c}_1\boldsymbol{e}_1 + \bar{g}_{K+1}\boldsymbol{e}_{K+1} + \boldsymbol{\theta}^{\boldsymbol{\alpha}}$, then following the same procedure as above one gets that if $\boldsymbol{p}$ satisfy (C.20), then $g_{K+1} - \bar{g}_{K=1} = 0$ and $c_1 - \bar{c}_1 = 0$.

Similarly as the above procedure for $k=1$, we have that if for any attribute $k \in \{1,\dots,K\}$, there exist two attribute profiles $\boldsymbol{\alpha}^{k,1}, \boldsymbol{\alpha}^{k,2} \in \{0,1\}^{k-1} \times \{0\} \times \{0,1\}^{K-k-1}$ such that

$$\frac{p_{\boldsymbol{\alpha}^{k,1}+\boldsymbol{e}_k}}{p_{\boldsymbol{\alpha}^{k,1}}} \neq \frac{p_{\boldsymbol{\alpha}^{k,2}+\boldsymbol{e}_k}}{p_{\boldsymbol{\alpha}^{k,2}}}, \tag{C.21}$$

then

$$\bar{g}_k = g_k, \quad \bar{c}_k = c_k\ , \quad \bar{g}_{K+k} = g_{K+k}, \quad \bar{c}_{K+k} = c_{K+k} \quad \text{for every } k \in \{1,\dots,K\}.$$

Now that all the item parameters are identified under (C.21), Equation (C.22) gives $\bar{\boldsymbol{p}} = \boldsymbol{p}$. Therefore other than the measure zero set of the parameter space specified by constraints (C.21), $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are identifiable. This means $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$ are generically identifiable.

In particular, if Q takes form of the $Q_{2\times 4}$ in (5.2),

$$Q_{2\times 4} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},$$

then constraints (C.21) just simplify to

$$\frac{p_{(10)}}{p_{(00)}} \neq \frac{p_{(11)}}{p_{(01)}} \quad \text{and} \quad \frac{p_{(01)}}{p_{(00)}} \neq \frac{p_{(11)}}{p_{(10)}},$$

which can be equivalently written as inequality (4.3) that $p_{(01)}p_{(10)} \neq p_{(00)}p_{(11)}$ in the main text.

Next we prove the conclusion when Q contains other rows besides the two identity submatrices, namely $Q = (I_K, I_K, (Q^\star)^\top)^\top$. Using exactly the same arguments as previously we have that generically, all the item parameters of the first $2K$ items as well as all the proportion parameters are satisfied. Now for any $J > 2K$ and $\boldsymbol{\alpha} \in \{0,1\}^K$ define $\boldsymbol{r}^* = \sum_{k=1}^K \boldsymbol{e}_j$ and

$$\boldsymbol{\theta}^* = \sum_{1 \le k \le K:\, \boldsymbol{\alpha} \succeq \boldsymbol{q}_j} g_j \boldsymbol{e}_j + \sum_{1 \le k \le K:\, \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_j} c_j \boldsymbol{e}_j,$$

then (C.2) implies that

$$\theta_{j,\boldsymbol{\alpha}} = \frac{T_{\boldsymbol{r}^*+\boldsymbol{e}_j}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)\boldsymbol{p}}{T_{\boldsymbol{r}^*}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)\boldsymbol{p}} = \frac{T_{\boldsymbol{r}^*+\boldsymbol{e}_j}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)\bar{\boldsymbol{p}}}{T_{\boldsymbol{r}^*}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)\bar{\boldsymbol{p}}} = \bar{\theta}_{j,\boldsymbol{\alpha}}.$$

This proves that any slipping or guessing parameter associated with item $j > 2K$ is identifiable under the generic constraints (C.21), and this completes the proof of part (b.2) of the theorem.

Next we prove (b.1) and (c) in Theorem 2 in four steps.

Proof of Part (b.1) and Part (c).

Step 1. In this step, we aim to show that if

$$T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q, \boldsymbol{s}, \boldsymbol{g})\boldsymbol{p} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{s}}, \bar{\boldsymbol{g}})\bar{\boldsymbol{p}} \quad \text{for every } \boldsymbol{r} \in \{0,1\}^J, \tag{C.22}$$

then $\bar{Q}$ must take the following form up to column permutation

$$\bar{Q} = \begin{pmatrix} 1 & \mathbf{0} \\ \bar{u} & \bar{\boldsymbol{v}} \\ \mathbf{0} & Q^\star \end{pmatrix}. \tag{C.23}$$

Here $(\bar{u}, \bar{\boldsymbol{v}})$ is a K dimensional binary vector. The structure of $(\bar{u}, \bar{\boldsymbol{v}})$ will be studied in Steps 2 and 3.

Since the submatrix $Q^\star$ of Q satisfies Conditions A, B and C, the matrix Q can be written as

$$
Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \boldsymbol{v}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q^{\star\star} \end{pmatrix},
$$

then follow the same procedure as Step 1 in the proof of Theorem IV.1 one has that, up to some column permutation, $\bar{Q}$ takes the form

$$
\bar{Q} = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \bar{u} & \bar{\boldsymbol{v}}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \bar{\boldsymbol{b}} & \bar{Q}^{\star\star} \end{pmatrix}.
$$

For notational convenience and without loss of generality, in the following proof we rearrange the order of the row vectors of Q (and $\bar{Q}$) and rewrite them as follows

$$
Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ 1 & \boldsymbol{v}^\top \\ \mathbf{0} & Q^{\star\star} \end{pmatrix}, \qquad \bar{Q} = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \bar{u} & \bar{\boldsymbol{v}}^\top \\ \bar{\boldsymbol{b}} & \bar{Q}^{\star\star} \end{pmatrix}. \tag{C.24}
$$

Now that each column of $Q^{\star\star}$ contains at least two entries of "1" from the assumption of scenarios (b.1) and (c), following the same procedure as Step 2 in the proof of Theorem IV.1 we can obtain

$$
c_j = \bar{c}_j, \quad \text{for } j = K+2, \ldots, J.
$$

Note that slightly different from Step 2 in the proof of Theorem IV.1, here we *do not have* $c_{K+1} = \bar{c}_{K+1}$ due to the fact that the first attribute is required by only two items.

Now denote the $(J-K)\times(K-1)$ bottom-right submatrix of Q by Q^s and the $(J-K)\times K$ bottom submatrix of Q by Q^l, i.e.,

$$Q^s = \begin{pmatrix} \boldsymbol{v}^\top \\ Q^{\star\star} \end{pmatrix}, \quad Q^l = \begin{pmatrix} 1 & \boldsymbol{v}^\top \\ \mathbf{0} & Q^{\star\star} \end{pmatrix}$$

and assume without loss of generality that the $K-1$ column vectors of Q^s are arranged in the lexicographic order. Specifically, for any $1 \le k_1 < k_2 \le K-1$, assume $Q^s_{\bullet,k_1} \prec_{\text{lex}} Q^s_{\bullet,k_2}$. This implies that the vector $\boldsymbol{v}$ can be written as

$$\boldsymbol{v} = (0,\dots,0,1,\dots,1)$$

Note that in scenario (b.1), $\boldsymbol{v}=\mathbf{0}$ and $k_0 = K-1$. where its first k_0 elements are zero and the remain $K-1-k_0$ elements are one. So $\boldsymbol{q}_2 = (1,\boldsymbol{v}) = (1,0,\dots,0,1,\dots,1)$. We now use an induction method to prove that

$$g_k = \bar{g}_k, \quad \forall k = 2,\ \dots,\ 1+k_0. \tag{C.25}$$

A key observation is that if considering the order of the columns of the larger submatrix Q^l instead of Q^s, then the first column of Q^l, i.e. $Q^l_{\bullet,1}$ is of larger lexicographic order of $Q^l_{\bullet,k}$ for any $k = 2,\dots,1+k_0$. This indicates that we can follow a similar induction argument as Step 3 in the proof of Theorem IV.1 by defining $\boldsymbol{\theta}^*_k$ as (the same form as (C.14))

$$\boldsymbol{\theta}^*_k = \sum_{h=1}^{K} \bar{g}_h \boldsymbol{e}_h + \sum_{h>K:\, q_{h,k}=0} g_h \boldsymbol{e}_h + \sum_{h>K:\, q_{h,k}=1} c_h \boldsymbol{e}_h, \tag{C.26}$$

for $k = 2,\dots,1+k_0$ one after another, to obtain (E.23).

We emphasize here that if $\boldsymbol{v}=\mathbf{0}$, i.e. in scenario (b.1) of the theorem, then

$k_0 = K - 1$ *and by far we have already obtained* $\bar{g}_k = g_k$ *for all* $k = 2, \ldots, K$. So we can directly go to the next step, Step 2 of the proof, without the local condition to appear in (C.29) later. That is why in scenario (b.1) of the theorem, we have *global* generic identifiability of $(Q, \boldsymbol{s}, \boldsymbol{g}, \boldsymbol{p})$.

Next we consider the case $\boldsymbol{v} \neq \mathbf{0}$, i.e. in scenario (c) of the theorem, then $k_0 < K-1$. We will use another induction argument to show $\bar{g}_k = g_k$ for $k = k_0+2, \ldots, K$, under an additional local condition. First we consider $\bar{g}_k$ and g_k for $k = k_0 + 2$. Note that $Q_{\bullet,k} \succ_{\text{lex}} Q_{\bullet,1}$, and $Q_{\bullet,k} \prec_{\text{lex}} Q_{\bullet,m}$ for any $m = k+1, \ldots, K$. Define $\boldsymbol{\theta}^*_k$ the same as in (C.26), then $T_{\boldsymbol{r}^*,\bullet}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*_k, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*_k) = \mathbf{0}$ and $T_{\boldsymbol{r}^*,\bullet}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*_k, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*_k)\bar{\boldsymbol{p}} = 0$, so $T_{\boldsymbol{r}^*,\bullet}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*_k, \boldsymbol{g} - \boldsymbol{\theta}^*_k)\boldsymbol{p} = 0$. We claim that in the the vector $T_{\boldsymbol{r}^*,\bullet}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*_k, \boldsymbol{g} - \boldsymbol{\theta}^*_k)$, denoted by $T_{\boldsymbol{r}^*,\bullet}$ afterwards for notational simplicity, only contains two potentially nonzero elements corresponding to attribute profiles $\boldsymbol{\alpha}_{1k} = \sum_{m=1}^K \boldsymbol{e}_m - \boldsymbol{e}_k = (1, \ldots, 1, \alpha_k = 0, 1, \ldots, 1)$ and $\boldsymbol{\alpha}_{0k} = \boldsymbol{\alpha}_1 - \boldsymbol{e}_1 = (\alpha_1 = 0, 1, \ldots, 1, \alpha_k = 0, 1, \ldots, 1)$. This is because on the one hand, for any attribute profile $\boldsymbol{\alpha}$ that lacks some attribute $m \in \{k+1, \ldots, K\}$, $\theta_{h,\boldsymbol{\alpha}} = g_h$ for some item $h > K$ with $q_{h,k} = 0$, which makes $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 0$; and on the other hand, for any attribute profile that lacks some attribute $m \in \{2, \ldots, k-1\}$, since we already have (E.23), $\theta_{h,m} = g_h = \bar{g}_h$ for some $h \in \{2, \ldots, K\}$, which makes $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 0$. Now $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ would only happen if $\boldsymbol{\alpha} = (\alpha_1, 1, \ldots, 1, \alpha_k, 1, \ldots, 1)$. However, if $\alpha_k = 1$ and $\boldsymbol{\alpha} = (\alpha_1, 1, \ldots, 1)$, then $\theta_{h,\boldsymbol{\alpha}} = c_h$ for some item $h > K$ with $q_{h,k} = 1$, which also makes $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 0$. Now we have proven the claim that $T_{\boldsymbol{r}^*,\bullet}$ has only two potentially nonzero elements corresponding to $\boldsymbol{\alpha}_{1k}$ and $\boldsymbol{\alpha}_{0k}$. Therefore we have for $k = k_0 + 2$,

$$
\begin{aligned}
0 =& T_{\boldsymbol{r}^*,\bullet}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*_k, \boldsymbol{g} - \boldsymbol{\theta}^*_k)\boldsymbol{p} \\
=& \prod_{h=2}^{K}(c_h - \bar{g}_h) \prod_{\substack{h>K:\\ q_{h,k}=0}} (c_h - g_h) \prod_{\substack{h>K:\\ q_{h,k}=1}} (g_h - c_h) \\
& \times \Big[(g_1 - \bar{g}_1)p_{\boldsymbol{\alpha}_{0k}} + (c_1 - \bar{g}_1)p_{\boldsymbol{\alpha}_{1k}}\Big](g_k - \bar{g}_k),
\end{aligned}
$$

which further gives

$$\Big[(g_1-\bar g_1)p_{\boldsymbol{\alpha}_{0k}}+(c_1-\bar g_1)p_{\boldsymbol{\alpha}_{1k}}\Big](g_k-\bar g_k)=0 \text{ for } k=k_0+2. \tag{C.27}$$

Note that if $\bar g_1 = g_1$, then the part in the bracket in the above display becomes $(c_1-g_1)p_{\boldsymbol{\alpha}_1}$, which is nonzero. Therefore, when $\bar g_1$ is sufficiently close to the true parameter g_1, the part in the bracket in (C.27) would be nonzero. We formally write it as

$$\begin{aligned}\text{for } k=k_0+2,\quad \forall \bar g_1\in\mathcal{N}_k,\quad (g_1-\bar g_1)p_{\boldsymbol{\alpha}_{0k}}+(c_1-\bar g_1)p_{\boldsymbol{\alpha}_{1k}}\neq 0,\\ \text{where } \mathcal{N}_k=\{x:0<x<\frac{g_1p_{\boldsymbol{\alpha}_{0k}}+c_1p_{\boldsymbol{\alpha}_{1k}}}{p_{\boldsymbol{\alpha}_{0k}}+p_{\boldsymbol{\alpha}_{1k}}}\}.\end{aligned} \tag{C.28}$$

This indicates that in the neighborhood $\mathcal{N}_k$ of g_1, (C.27) leads to $g_k=\bar g_k$ for $k=k_0+2$.

Then we use induction to prove $g_k=\bar g_k$ for all $k=k_0+3,\dots,K$. As the induction assumption, assume that when $\bar g_1\in\bigcap_{m=k_0+2}^{k-1}\mathcal{N}_m$ holds, we have $g_m=\bar g_m$ for all $m=2,\dots,k-1$. Then define $\boldsymbol{\theta}^*$ the same as in (C.26), and deduce in the same way as in proving $g_{k_0+2}=\bar g_{k_0+2}$, we have

$$\Big[(g_1-\bar g_1)p_{\boldsymbol{\alpha}_{0k}}+(c_1-\bar g_1)p_{\boldsymbol{\alpha}_{1k}}\Big](g_k-\bar g_k)=0,$$

and further for any $\bar g_1\in\mathcal{N}_k$ (more accurately any $\bar g_1\in\big[\cap_{m=k_0+2}^{k-1}\mathcal{N}_m\big]\cap\mathcal{N}_k$), we must have $\bar g_k=g_k$. Here $\mathcal{N}_k$ takes the same form as that in (C.28). Now by induction, we have that if

$$\bar g_1\in\bigcap_{m=k_0+2}^{K}\mathcal{N}_m, \tag{C.29}$$

then $g_k=\bar g_k$ for $k=k_0+2,\dots,K$. Combined with the previous results shown in (E.23), now we have proven that in scenario (c) of the theorem, if the local condition (C.29) is satisfied, then $\bar g_k=g_k$ for $k=2,\dots,K$.

In summary, we have shown $\bar{g}_k = g_k$ for $k = 2, \ldots, K$ (under (C.29) if in scenario (c)) and $\bar{c}_j = c_j$ for $j = K+2, \ldots, J$. Based on these, following similar procedures as in Step 5 of the proof of Theorem IV.1, we obtain that

$$\bar{\boldsymbol{q}}_j = \boldsymbol{q}_j, \quad \forall j = K+2, \ldots, J.$$

Step 2. In this step we show $\bar{u} = 1$ in (C.23). If $\bar{u} = 0$, set

$$\boldsymbol{\theta}^* = c_1 \boldsymbol{e}_1 + \bar{c}_2 \boldsymbol{e}_2 + \sum_{j=3}^{K+3} g_k \boldsymbol{e}_k, \quad \boldsymbol{r}^* = \sum_{j=1}^{K+3} \boldsymbol{e}_j,$$

then

$$\begin{aligned}
&T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}} - \boldsymbol{\theta}^*)\bar{\boldsymbol{p}} = \boldsymbol{0}^\top \cdot \bar{\boldsymbol{p}} = 0, \\
&T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^*, \boldsymbol{g} - \boldsymbol{\theta}^*)\boldsymbol{p} = (g_1 - c_1)(g_2 - \bar{c}_2) \prod_{j=3}^{K+3} (c_j - g_j) p_{(0,1,\ldots,1)} \neq 0,
\end{aligned}$$

which contradicts Equation (C.3). So $\bar{u} = 1$. Now we have obtained

$$Q = \begin{pmatrix} 1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \mathcal{I}_{K-1} \\ 1 & \boldsymbol{v}^\top \\ \boldsymbol{0} & Q^{\star\star} \end{pmatrix}, \qquad \bar{Q} = \begin{pmatrix} 1 & \boldsymbol{0}^\top \\ \boldsymbol{0} & \mathcal{I}_{K-1} \\ 1 & \bar{\boldsymbol{v}}^\top \\ \boldsymbol{0} & Q^{\star\star} \end{pmatrix}. \tag{C.30}$$

Step 3. In this step we show $\bar{\boldsymbol{v}} = \boldsymbol{v}$. For notational simplicity in the following proof, we rearrange the order of the row vectors in Q and $\bar{Q}$ in (C.30) again to the following forms

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \boldsymbol{v}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q^{\star\star} \end{pmatrix}, \qquad \bar{Q} = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \bar{u} & \bar{\boldsymbol{v}}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q^{\star\star} \end{pmatrix}, \tag{C.31}$$

and our conclusions proved so far are $\bar{g}_k = g_k$ for $k = 3, \ldots, K+1$ and $\bar{c}_j = c_j$ for $j = K+2, \ldots, J$ (under the local condition (C.29) if in scenario (b.1)). Given that the last $J-2$ rows of Q and $\bar{Q}$ are equal, we claim that (C.22) for response pattern $\boldsymbol{r}$ can be equivalently written as

$$\begin{aligned} &\sum_{\substack{\boldsymbol{\alpha}'\in \\ \{0,1\}^{K-1}}} \prod_{\substack{j>2 \\ r_j=1}} \theta_{j,(0,\boldsymbol{\alpha}')} \cdot \mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) \qquad \text{(C.32)} \\ =& \sum_{\substack{\boldsymbol{\alpha}'\in \\ \{0,1\}^{K-1}}} \prod_{\substack{j>2 \\ r_j=1}} \bar{\theta}_{j,(0,\boldsymbol{\alpha}')} \cdot \mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}). \end{aligned}$$

Here $\mathbf{A} = (A_1, \ldots, A_K)$ denotes a random attribute profile following a categorical distribution with proportion parameters $\boldsymbol{p}$, and $\mathbf{A}_{2:K}$ denotes the vector consisting of the last $K-1$ elements of $\mathbf{A}$. The reason for the equivalence of (C.32) and (C.22) is stated as follows. Since all items other than the first two do not require the first attribute, we have that for any $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$, the two attribute profiles $(0, \boldsymbol{\alpha}')$ and $(1, \boldsymbol{\alpha}')$ always have the same response probability $\theta_{j,(0,\boldsymbol{\alpha}')}$ to any item $j > 2$. This indicates that the left hand side of (C.22) can be written as

$$T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q, \boldsymbol{s}, \boldsymbol{g})\boldsymbol{p} = \sum_{\substack{\boldsymbol{\alpha}'\in \\ \{0,1\}^{K-1}}} \prod_{\substack{j>2 \\ r_j=1}} \theta_{j,(0,\boldsymbol{\alpha}')} \cdot \mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}),$$

and this further leads to the equivalence between (C.22) and (C.32). In particular, when $(r_1, r_2) = (0,0)$, we have $\mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) = p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')}$. Now for any J-dimensional response pattern $\boldsymbol{r}$ with $(r_1, r_2) = (0,0)$, then the constraint $T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q,\boldsymbol{\theta}^{+},\boldsymbol{g})\boldsymbol{p} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{Q},\bar{\boldsymbol{\theta}}^{+},\bar{\boldsymbol{g}})\bar{\boldsymbol{p}}$ simply becomes

$$\sum_{\substack{\boldsymbol{\alpha}'\in\\ \{0,1\}^{K-1}}} \prod_{\substack{j>2\\ r_j=1}} \theta_{j,(0,\boldsymbol{\alpha}')} \cdot (p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')}) = \sum_{\substack{\boldsymbol{\alpha}'\in\\ \{0,1\}^{K-1}}} \prod_{\substack{j>2\\ r_j=1}} \bar{\theta}_{j,(0,\boldsymbol{\alpha}')} \cdot (\bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}).$$

Since the above equality holds for any $(r_3, r_4, \ldots, r_J) \in \{0,1\}^{J-2}$, we claim that, parameters $\theta_{j,(0,\boldsymbol{\alpha}')}$ and $\bar{\theta}_{j,(0,\boldsymbol{\alpha}')}$ for $j = 3,\ldots,J$ can be equivalently viewed as all the item parameters (slipping or guessing) associated with the submatrix $Q^{\star}$, while grouped proportion parameters $p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')}$ and $\bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}$ can be viewed as all the "proportion parameters" associated with $Q^{\star}$. Since $Q^{\star}$ satisfy the sufficient conditions A, B, C in Theorem IV.1 for identifiability, by Theorem IV.1 we conclude that $\theta_{j,(0,\boldsymbol{\alpha}')} = \bar{\theta}_{j,(0,\boldsymbol{\alpha}')}$ for any $j \in \{3,\ldots,J\}$ and any $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$. This indicates $\bar{c}_k = c_k$ for $k = 3,\ldots,K+1$ and $\bar{g}_j = g_j$ for $j = K+2,\ldots,J$.

Then an important observation is that, fix any particular pair of $(r_1, r_2) \in \{0,1\}^2$, quantities in (C.32) can be viewed parameters associated with the $(J-2)\times(K-1)$ matrix $Q^{\star}$, just similar to the argument in the previous paragraph. Specifically, $\theta_{j,(0,\boldsymbol{\alpha}')}$ and $\bar{\theta}_{j,(0,\boldsymbol{\alpha}')}$ for $j = 3,\ldots,J$ are item parameters (slipping or guessing) associated with the $Q^{\star}$, and $\mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid Q, \boldsymbol{\Theta}, \boldsymbol{p})$ and $\mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ for each $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$ can be viewed as the "proportion parameters" associated with $Q^{\star}$. Now because the submatrix $Q^{\star}$ satisfy the identifiability conditions A, B, C; and $\bar{Q}_{3:J,\boldsymbol{\cdot}} = Q_{3:J,\boldsymbol{\cdot}} = Q^{\star}$ and $\bar{c}_j = c_j$, $\bar{g}_j = g_j$ for $j = 3,\ldots,J$, we must have

$$\begin{aligned} \forall \boldsymbol{\alpha}' \in \{0,1\}^{K-1}, \qquad & \mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid Q, \boldsymbol{\Theta}, \boldsymbol{p}) \\ & = \mathbb{P}(R_1 \geq r_1,\, R_2 \geq r_2,\, \mathbf{A}_{2:K} = \boldsymbol{\alpha}' \mid \bar{Q}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}). \end{aligned} \tag{C.33}$$

Now take (r_1, r_2) to be $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ in the above (C.33) respectively, we obtain

$$
\begin{cases}
p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} = \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}; \\
\theta_{1,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{1,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} = \bar{\theta}_{1,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}; \\
\theta_{2,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{2,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} = \bar{\theta}_{2,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{2,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}; \\
\theta_{1,(0,\boldsymbol{\alpha}')}\theta_{2,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{1,(1,\boldsymbol{\alpha}')}\theta_{2,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} \\
\qquad = \bar{\theta}_{1,(0,\boldsymbol{\alpha}')}\bar{\theta}_{2,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,(1,\boldsymbol{\alpha}')}\bar{\theta}_{2,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}.
\end{cases}
\tag{C.34}
$$

Next we show $\boldsymbol{v} = \bar{\boldsymbol{v}}$. (C.34) implies that,

$$
\forall \boldsymbol{\alpha}' \geq \boldsymbol{v},\ \boldsymbol{\alpha}' \ngeq \bar{\boldsymbol{v}}, \quad
\begin{cases}
p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} = \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')} \\
g_1 p_{(0,\boldsymbol{\alpha}')} + c_1 p_{(1,\boldsymbol{\alpha}')} = \bar{g}_1 \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{c}_1 \bar{p}_{(1,\boldsymbol{\alpha}')} \\
g_2 p_{(0,\boldsymbol{\alpha}')} + c_2 p_{(1,\boldsymbol{\alpha}')} = \bar{g}_2 [\bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}] \\
g_1 g_2 p_{(0,\boldsymbol{\alpha}')} + c_1 c_2 p_{(1,\boldsymbol{\alpha}')} = \bar{g}_2 [\bar{g}_1 \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{c}_1 \bar{p}_{(1,\boldsymbol{\alpha}')}]
\end{cases}
$$

If $\bar{\boldsymbol{v}} \ngeq \boldsymbol{v}$, then taking $\boldsymbol{\alpha}' = \boldsymbol{v}$ in the above equation and doing some transformation gives

$$
\begin{cases}
(g_2 - \bar{g}_2) p_{(0,\boldsymbol{\alpha}')} + (c_2 - \bar{g}_2) p_{(1,\boldsymbol{\alpha}')} = 0, \\
(g_1 - c_1)(g_2 - \bar{g}_2) p_{(0,\boldsymbol{\alpha}')} = 0.
\end{cases}
$$

Since $g_1 \neq c_1$, we have $g_2 - \bar{g}_2 = 0$, which further gives $c_2 - \bar{g}_2 = 0$. This contradicts $c_h > \bar{g}_h$ for any item h, so $\bar{\boldsymbol{v}} \ngeq \boldsymbol{v}$ can not happen. Similarly $\bar{\boldsymbol{v}} \nleq \boldsymbol{v}$ also can not happen, so $\bar{\boldsymbol{v}} = \boldsymbol{v}$.

Step 4. In the final step we show c_1, c_2, g_1, g_2 and $\boldsymbol{p}$ are generically identifiable if

$\boldsymbol{v} \neq \mathbf{1}$. First we show that if there exist $\boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2 \in \{0,1\}^{K-1}$, $\boldsymbol{\alpha}'_1 \neq \boldsymbol{\alpha}'_2$ such that

$$p_{(1,\boldsymbol{\alpha}'_1)}p_{(0,\boldsymbol{\alpha}'_2)} \neq p_{(1,\boldsymbol{\alpha}'_2)}p_{(0,\boldsymbol{\alpha}'_1)}, \tag{C.35}$$

then one must have

$$c_i = \bar{c}_i,\ g_i = \bar{g}_i, \quad i = 1, 2. \tag{C.36}$$

After some transformations, the system of equations (C.40) yields

$$\begin{cases} (g_1 - c_1) \cdot (g_2 - \bar{c}_2) \cdot p_{(0,\boldsymbol{\alpha}')} = (\bar{g}_1 - c_1) \cdot (\bar{g}_2 - \bar{c}_2) \cdot \bar{p}_{(0,\boldsymbol{\alpha}')}, \\ (g_2 - \bar{c}_2) \cdot p_{(0,\boldsymbol{\alpha}')} + (c_2 - \bar{c}_2) \cdot \bar{p}_{(1,\boldsymbol{\alpha}')} = (\bar{g}_2 - \bar{c}_2) \cdot \bar{p}_{(0,\boldsymbol{\alpha}')}. \end{cases}$$

Since we have $\bar{g}_1 \neq c_1$, the left hand side of the first equation above is nonzero. And obviously the right hand side of the second equation above is nonzero. Taking the ratio of the above two equations gives

$$\frac{(g_1 - c_1) \cdot (g_2 - \bar{c}_2)}{(g_2 - \bar{c}_2) + (c_2 - \bar{c}_2) \cdot p_{(1,\boldsymbol{\alpha}')}/p_{(0,\boldsymbol{\alpha}')}} = (\bar{g}_1 - c_1) \equiv f(\boldsymbol{\alpha}').$$

The right hand side of the above display does not involve any proportion parameter $\boldsymbol{p}$ or $\bar{\boldsymbol{p}}$. So for $\boldsymbol{\alpha}'_1$, $\boldsymbol{\alpha}'_2$ satisfying (C.35), $f(\boldsymbol{\alpha}'_1) = f(\boldsymbol{\alpha}'_2)$. Note that the left hand side of the above equation involves a ratio $p_{(1,\boldsymbol{\alpha}')}/p_{(0,\boldsymbol{\alpha}')}$ depending on $\boldsymbol{\alpha}'$. Equality $f(\boldsymbol{\alpha}'_1) = f(\boldsymbol{\alpha}'_2)$ along with (C.35) imply

$$(c_2 - \bar{c}_2) \cdot \frac{p_{(1,\boldsymbol{\alpha}'_1)}}{p_{(0,\boldsymbol{\alpha}'_1)}} = (c_2 - \bar{c}_2) \cdot \frac{p_{(1,\boldsymbol{\alpha}'_2)}}{p_{(0,\boldsymbol{\alpha}'_2)}}$$

$$(c_2 - \bar{c}_2) \cdot \left(\frac{p_{(1,\boldsymbol{\alpha}'_1)}}{p_{(0,\boldsymbol{\alpha}'_1)}} - \frac{p_{(1,\boldsymbol{\alpha}'_2)}}{p_{(0,\boldsymbol{\alpha}'_2)}} \right) = 0$$

then since $p_{(1,\boldsymbol{\alpha}'_1)}p_{(0,\boldsymbol{\alpha}'_2)} \neq p_{(1,\boldsymbol{\alpha}'_2)}p_{(0,\boldsymbol{\alpha}'_1)}$ by assumption (C.35), one must have $c_2 = \bar{c}_2$. By symmetry of the four item parameters g_1, c_1, g_2 and c_2 in (C.40), equalities (C.36) hold as claimed following similar arguments. Now that all the item parameters are identified, $\boldsymbol{p} = \bar{\boldsymbol{p}}$. This completes the proof of part (b.1) and part (c) of the theorem. The proof of Theorem IV.2 is now complete.

C.4 Proof of Theorem IV.3

When Condition C fails and some attribute is required by less than three items, there are two possible scenarios: some attribute is required by only one item, or only two items. We consider them separately, and in both cases prove that $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ are not generically identifiable.

(a) If some attribute is required by only one item. Then Q must take the following form in (C.37) up to column and row permutations, where $\boldsymbol{v}_1$ is a binary vector of length $K-1$.

$$Q = \begin{pmatrix} 1 & \boldsymbol{v}_1^\top \\ \mathbf{0} & Q^\star \end{pmatrix}; \qquad \bar{Q} = \begin{pmatrix} 1 & \mathbf{1}^\top \\ \mathbf{0} & Q^\star \end{pmatrix}. \tag{C.37}$$

Now for arbitrary model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ associated with Q, we also construct $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ associated with the $\bar{Q}$ in (C.37), such that (C.1) holds. Firstly, for any item $j \geq 2$, set $\bar{\theta}_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$, then following a similar argument as in Step 3 of the proof of Theorem IV.2 (b.1) and (c), we have that (C.1) hold as long as the following constraints are satisfied: for any $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$,

$$\begin{cases} p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} = \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}; \\ \theta_{1,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{1,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} = \bar{\theta}_{1,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}. \end{cases} \tag{C.38}$$

For each $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$, we now still arbitrarily set the value of $\bar{\theta}_{1,(0,\boldsymbol{\alpha}')}$ and

$\bar{\theta}_{1,(1,\boldsymbol{\alpha}')}$, and set the proportions parameters to be

$$\bar{p}_{(1,\boldsymbol{\alpha}')} = \frac{(\theta_{1,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{1,(0,\boldsymbol{\alpha}')})p_{(0,\boldsymbol{\alpha}')} + (\theta_{1,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{1,(0,\boldsymbol{\alpha}')})p_{(1,\boldsymbol{\alpha}')}}{\bar{\theta}_{1,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{1,(0,\boldsymbol{\alpha}')}}$$

$$\bar{p}_{(0,\boldsymbol{\alpha}')} = p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} - \bar{p}_{(1,\boldsymbol{\alpha}')},$$

for each $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$. Then (C.38) holds and further (C.1) holds. Since the choice of the 2^K item parameters $\{\theta_{1,(0,\boldsymbol{\alpha}')}, \theta_{1,(1,\boldsymbol{\alpha}')} : \boldsymbol{\alpha}' \in \{0,1\}^{K-1}\}$ are arbitrary, the original Q and associated parameters are not generically identifiable.

(b) If some attribute is required by only two items, then Q takes the form in (C.39) up to column/row permutations, where $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are vectors of length $K-1$ and $Q^\star$ is a submatrix of size $(J-2)\times(K-1)$.

$$Q = \begin{pmatrix} 1 & \boldsymbol{v}_1^\top \\ 1 & \boldsymbol{v}_2^\top \\ \mathbf{0} & Q^\star \end{pmatrix}; \qquad \bar{Q} = \begin{pmatrix} 1 & \mathbf{1}^\top \\ 1 & \mathbf{1}^\top \\ \mathbf{0} & Q^\star \end{pmatrix}, \tag{C.39}$$

Then for arbitrary model parameters $(\boldsymbol{\Theta}, \boldsymbol{p})$ associated with Q, we next carefully construct $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ associated with the $\bar{Q}$ in (C.39), such that (C.1) holds. This would prove the conclusion that joint generic identifiability fails. Firstly, for any item $j \geq 3$, set $\bar{\theta}_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}}$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$, then following the same argument as in Step 3 of the proof of Theorem IV.2 (b.1) and (c), we have that (C.1) hold as long as the following constraints are satisfied for every $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$,

$$
\begin{cases}
p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} = \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{p}_{(1,\boldsymbol{\alpha}')}; \\
\theta_{1,\,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{1,\,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} = \bar{\theta}_{1,\,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,\,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}; \\
\theta_{2,\,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{2,\,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} = \bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{2,\,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}; \\
\theta_{1,\,(0,\boldsymbol{\alpha}')}\theta_{2,\,(0,\boldsymbol{\alpha}')} \cdot p_{(0,\boldsymbol{\alpha}')} + \theta_{1,\,(1,\boldsymbol{\alpha}')}\theta_{2,\,(1,\boldsymbol{\alpha}')} \cdot p_{(1,\boldsymbol{\alpha}')} \\
\qquad = \bar{\theta}_{1,\,(0,\boldsymbol{\alpha}')}\bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')} \cdot \bar{p}_{(0,\boldsymbol{\alpha}')} + \bar{\theta}_{1,\,(1,\boldsymbol{\alpha}')}\bar{\theta}_{2,\,(1,\boldsymbol{\alpha}')} \cdot \bar{p}_{(1,\boldsymbol{\alpha}')}.
\end{cases}
\tag{C.40}
$$

For each $\boldsymbol{\alpha}' \in \{0,1\}^{K-1}$, arbitrarily choose $\bar{\theta}_{1,\,(0,\boldsymbol{\alpha}')}$ and $\bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')}$ from the neighborhood of the true parameter values $\theta_{1,(0,\boldsymbol{\alpha}')}$ and $\theta_{2,(1,\boldsymbol{\alpha}')}$ respectively. Then set

$$
\begin{cases}
\bar{\theta}_{1,\,(1,\boldsymbol{\alpha}')} = \theta_{1,(0,\boldsymbol{\alpha}')} + \frac{([\theta_{1,\,(1,\boldsymbol{\alpha}')} - \theta_{1,\,(0,\boldsymbol{\alpha}')}][\theta_{2,\,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')}]p_{(1,\boldsymbol{\alpha}')}}{[\theta_{2,\,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')}]p_{(0,\boldsymbol{\alpha}')} + [\theta_{2,\,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')}]p_{(1,\boldsymbol{\alpha}')}}, \\
\bar{\theta}_{2,\,(1,\boldsymbol{\alpha}')} = \theta_{2,(0,\boldsymbol{\alpha}')} + \frac{[\theta_{2,\,(1,\boldsymbol{\alpha}')} - \theta_{2,\,(0,\boldsymbol{\alpha}')}][\theta_{1,\,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{1,\,(0,\boldsymbol{\alpha}')}]p_{(1,\boldsymbol{\alpha}')}}{[\theta_{1,\,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{1,\,(0,\boldsymbol{\alpha}')}]p_{(0,\boldsymbol{\alpha}')} + [\theta_{1,\,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{1,\,(0,\boldsymbol{\alpha}')}]p_{(1,\boldsymbol{\alpha}')}}, \\
\bar{p}_{(1,\boldsymbol{\alpha}')} = \frac{[\theta_{2,\,(0,\boldsymbol{\alpha}')} - \bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')}]p_{(0,\boldsymbol{\alpha}')} + [\theta_{2,\,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')}]p_{(1,\boldsymbol{\alpha}')}}{\bar{\theta}_{2,\,(1,\boldsymbol{\alpha}')} - \bar{\theta}_{2,\,(0,\boldsymbol{\alpha}')}}, \\
\bar{p}_{(0,\boldsymbol{\alpha}')} = p_{(0,\boldsymbol{\alpha}')} + p_{(1,\boldsymbol{\alpha}')} - \bar{p}_{(1,\boldsymbol{\alpha}')}.
\end{cases}
\tag{C.41}
$$

Then one can check that (C.40) holds and further (C.1) holds. Since in the above construction the choice of the 2^K item parameters $\{\theta_{1,\,(0,\boldsymbol{\alpha}')}, \theta_{2,\,(0,\boldsymbol{\alpha}')} : \boldsymbol{\alpha}' \in \{0,1\}^{K-1}\}$ are arbitrary, we have proved that the Q and associated model parameters are not generically identifiable.

C.5 Proof of Theorem IV.4

We prove this theorem following a similar argument as the proof of Theorem 7 in Gu and Xu (2020a). Assume Q takes the form $Q = (Q_1^\top, Q_2^\top, (Q^\star)^\top)^\top$, where Q_1 and

Q_2 have all diagonal elements being 1. Assume

$$\begin{aligned}\theta_{j,\boldsymbol{\alpha}} =& f\Big(\beta_{j,0} + \sum_{k=1}^{K}\beta_{j,k}q_{j,k}\alpha_k + \sum_{k'=k+1}^{K}\sum_{k=1}^{K-1}\beta_{j,kk'}(q_{j,k}\alpha_k)(q_{j,k'}\alpha_{k'}) + \cdots \\ &+ \beta_{j,12\cdots K}\prod_k (q_{j,k}\alpha_k)\Big),\end{aligned}$$

where $f(\boldsymbol{\cdot})$ is some link function and when $f(\boldsymbol{\cdot})$ is the identify function, the model is the GDINA model. We first show that under Condition D, the $2^K \times 2^K$ matrices $T(Q_1, \boldsymbol{\Theta}_{Q_1})$ and $T(Q_2, \boldsymbol{\Theta}_{Q_2})$ both have full rank 2^K generically. It suffices to find some valid $\boldsymbol{\Theta}$ (i.e., $\boldsymbol{\Theta}_Q$) that gives

$$\det(T(Q_1, \boldsymbol{\Theta}_{Q_1})) \neq 0, \quad \det(T(Q_2, \boldsymbol{\Theta}_{Q_2})) \neq 0. \tag{C.42}$$

The reason is as follows. (C.42) would imply the polynomials defining the two matrix determinants are not zero polynomials in the Q-restricted parameter space. Therefore for almost all parameters, $T(Q_1, \boldsymbol{\Theta}_{Q_1})$ and $T(Q_2, \boldsymbol{\Theta}_{Q_2})$ would have full rank. Next we only focus on $T(Q_1, \boldsymbol{\Theta}_{Q_1})$. For every item $k = 1, \ldots, K$, we set $\beta_{k,k} = 1$, $\beta_{k,k'} = 0$ for any $k' \neq k$, and set all the interaction effects to zero. Then $T(Q_1, \boldsymbol{\Theta}_{Q_1})$ becomes identical to $T(I_K, \widehat{\boldsymbol{\Theta}}_{I_K})$ under a Q-matrix I_K with associated item parameters $\widehat{\boldsymbol{\Theta}}_{I_K}$ defined as follows: $\widehat{\theta}_{\boldsymbol{e}_k,\mathbf{0}} = \beta_{k,0}$, and $\widehat{\theta}_{\boldsymbol{e}_k,\boldsymbol{e}_k} = \widehat{\theta}_{\boldsymbol{e}_k,\mathbf{1}} = \beta_{k,0}+\beta_{k,k}$ for $k \in \{1, \ldots, K\}$. It is not hard to see that $T(I_K, \widehat{\boldsymbol{\Theta}}_{I_K})$ can be viewed as a T-matrix under the DINA model with the Q-matrix equal to I_K, and guessing parameters $\beta_{k,0}$, slipping parameters $1 - \beta_{k,0} - \beta_{k,k}$ for $k = 1, \ldots, K$. Therefore $T(I_K, \widehat{\boldsymbol{\Theta}}_{I_K})$ has full rank as argued in Step 1 of the proof of Theorem IV.1. So $T(Q_1, \boldsymbol{\Theta}_{Q_1})$ has full rank generically.

We next prove that if Condition E holds in addition, then any two different columns of $T(Q^\star, \boldsymbol{\Theta}_{Q^\star})$ are distinct generically. For $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}' \in \{0,1\}^K$ and $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$, they at least differ in one element. Assume without loss of generality that $\alpha_k = 1 > 0 = \alpha'_k$. Then Condition E ensures that there is some item $j > 2K$ with $q_{j,k} = 1$. Under the general RLCM, this implies $\theta_{j,\boldsymbol{\alpha}} \neq \theta_{j,\boldsymbol{\alpha}'}$ generically. By Kruskal (1977), a matrix's Kruskal rank is the largest number I such that every set of I columns of the matrix are independent. When a matrix has full rank, its Kruskal rank equals its rank. By this definition, $T(Q^{\star}, \boldsymbol{\Theta}_{Q^{\star}})$ has Kruskal rank at least 2 generically, and $T(Q_1, \boldsymbol{\Theta}_{Q_1})$, $T(Q_2, \boldsymbol{\Theta}_{Q_2})$ have Kruskal rank 2^K generically. Then for generic $\boldsymbol{\Theta}_Q$, we have

$$\operatorname{rank}_K\{T(Q_1, \boldsymbol{\Theta}_{Q_1})\} + \operatorname{rank}_K\{T(Q_2, \boldsymbol{\Theta}_{Q_2})\} + \operatorname{rank}_K\{T(Q^{\star}, \boldsymbol{\Theta}_{Q^{\star}})\} \geq 2 \times 2^K + 2. \tag{C.43}$$

Applying Corollary 2 of Rhodes (2010) to this 2^K-class latent class model, we get $T(Q, \boldsymbol{\Theta}) = T(Q, \bar{\boldsymbol{\Theta}})$ and $\boldsymbol{p} = \bar{\boldsymbol{p}}$ up to column permutation. This proves generic identifiability of $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ in the model. Moreover, we also have the following form of the identifiable set

$$\begin{aligned} \boldsymbol{\vartheta}_Q \setminus \boldsymbol{\vartheta}_{non} = \{&(\boldsymbol{\Theta}_Q, \boldsymbol{p}) : \det(T(Q_1, \boldsymbol{\Theta}_{Q_1})) \neq 0, \det(T(Q_2, \boldsymbol{\Theta}_{Q_2})) \neq 0, \\ &T(Q^{\star}, \boldsymbol{\Theta}_{Q^{\star}}) \cdot \operatorname{Diag}(\boldsymbol{p}) \text{ has column vectors different from each other}\}. \end{aligned}$$

This is because when $(\boldsymbol{\Theta}_Q, \boldsymbol{p}) \in \boldsymbol{\vartheta}_Q \setminus \boldsymbol{\vartheta}_{non}$, the rank condition (C.43) is satisfied and joint identifiability of $(Q, \boldsymbol{\Theta}_Q, \boldsymbol{p})$ follows.

C.6 Proof of Theorem IV.5.

We prove the theorem in two steps. In the first step, we show that if Q is not generically complete, than it must take the following form (up to column/row permutations) for some $k > m$,

$$Q = \left(\begin{array}{ccc|ccc} q_{1,1} & \cdots & q_{1,k} & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{m,1} & \cdots & q_{m,k} & * & \cdots & * \\ \hline 0 & \cdots & 0 & * & \cdots & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & * & \cdots & * \end{array}\right) = \left(\begin{array}{c|c} Q_{11} & Q_{12} \\ \hline Q_{21} & Q_{22} \end{array}\right) = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}. \tag{C.44}$$

The bottom-left submatrix $Q_{21} = \mathbf{0}_{(J-m)\times k}$. Any entry not in Q_{21} can be either 0 or 1. We introduce some definitions first. Given a Q-matrix Q, define a family S_Q of K finite sets $S_Q = \{\mathcal{A}_1, \mathcal{A}_1, \dots, \mathcal{A}_K\}$, where $\mathcal{A}_k = \{1 \le j \le J : \ q_{j,k} = 1\}$ for each k. Then $\mathcal{A}_k$ denotes the set of items that require attribute k. For the family S_Q, a *transversal* is a system of distinct representatives from each of its elements $\mathcal{A}_1, \dots, \mathcal{A}_K$. For example, for

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

we have $S_Q = \{\mathcal{A}_1 = \{1,3\},\ \mathcal{A}_2 = \{1,2\},\ \mathcal{A}_3 = \{2,3\}\}$. Then $(1,2,3)$ is a valid transversal of S_Q, and so as $(3,1,2)$; but $(1,1,2)$ is not a transversal. Now it is not hard to see that, the assumption that Q is not generically complete is equivalent to the following statement $H^\star$,

$H^\star$. Given Q, the family S_Q does not have a valid transversal.

Then by Hall's Marriage Theorem (Hall, 1967), the nonexistence of a transversal indicates the failure of the marriage condition. So there must exist a subfamily $W \subseteq S_Q$ such that $|W| > |\bigcup_{\mathcal{A}\in W} \mathcal{A}|$. More specifically, this means there exist some

$l_1, l_2, \ldots, l_k \in \{1, \ldots, K\}$ and $W = \{\mathcal{A}_{l_1}, \ldots, \mathcal{A}_{l_k}\}$ such that

$$|W| = k > |\mathcal{A}_{l_1} \cup \cdots \cup \mathcal{A}_{l_k}| \stackrel{\text{def}}{=} m.$$

In other words, we have shown that there exist some attributes, the number of which (e.g., k) exceeds the number of items that require any of these attributes (e.g., m). This is exactly saying that Q has to take the form of (C.44) with $k > m$ after some column/row permutation.

In the second step, we show that if Q takes the form of (C.44) with $k > m$, then $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ under general RLCMs are not generically identifiable. Now we define another potentially different $\bar{Q}$ as

$$\bar{Q} = \left(\begin{array}{c|c} Q_{11} & \bar{Q}_{12} \\ \hline Q_{21} & Q_{22} \end{array}\right) = \begin{pmatrix} \bar{Q}_1 \\ \\ Q_2 \end{pmatrix}, \quad \text{where } \bar{Q}_{12} = \mathbf{1}_{m\times(K-k)}.$$

Then given arbitrary $(\boldsymbol{\Theta}, \boldsymbol{p})$ associated with Q, we set $\bar{\theta}_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}}$ for every $j = m+1, \ldots, J$ and every $\boldsymbol{\alpha} \in \{0,1\}^K$. Because Q_{21} is a $(J-m) \times k$ zero matrix, we claim that under the current construction, the original 2^J constraints in (C.1) are satisfied as long as the following constraints are satisfied

$$\forall \boldsymbol{\alpha'} = (\alpha_{k+1}, \ldots, \alpha_K) \in \{0,1\}^{K-k}, \quad \forall \boldsymbol{r'} = (r_1, \ldots, r_m) \in \{0,1\}^m,$$
$$\sum_{\boldsymbol{\alpha}^\star \in \{0,1\}^k} T_{\boldsymbol{r'},(\boldsymbol{\alpha}^\star,\boldsymbol{\alpha'})}(Q_1, \boldsymbol{\Theta}_{Q_1}) \cdot p_{(\boldsymbol{\alpha}^\star,\boldsymbol{\alpha'})} = \sum_{\boldsymbol{\alpha}^\star \in \{0,1\}^k} T_{\boldsymbol{r'},(\boldsymbol{\alpha}^\star,\boldsymbol{\alpha'})}(\bar{Q}_1, \bar{\boldsymbol{\Theta}}_{\bar{Q}_1}) \cdot \bar{p}_{(\boldsymbol{\alpha}^\star,\boldsymbol{\alpha'})}.$$

This claim can be shown following a similar argument as that in Step 3 of the proof of Theorem IV.2 (b.1) and (c). Then the above system of equations contain $2^{K-k} \times 2^m$ constraints, while under the general RLCMs the number of free variables in $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$

involved is

$$\begin{aligned}&\left|\{\bar{p}_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0,1\}^K\} \bigcup \{\bar{\theta}_{j,\boldsymbol{\alpha}} : j \in \{1,\dots,m\}, \boldsymbol{\alpha} \in \{0,1\}^K\}\right| \\ &= 2^K + 2^{K-k} \times \left(\sum_{j=1}^m 2^{q_{j,1}+\dots+q_{j,k}}\right) \geq 2^K + 2^{K-k} \times m.\end{aligned}$$

Under the assumption $m < k$, we have that the number of constraints $2^{K-k} \times 2^m$ is smaller than the number of variables to solve (which is lower bounded by $2^{K-k} \times (2^k + m)$), because $2^m < 2^k + m$. So there exist infinitely many different sets of solutions of $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}})$ associated with $\bar{Q}$ such that $T(Q, \boldsymbol{\Theta})\boldsymbol{p} = T(\bar{Q}, \bar{\boldsymbol{\Theta}})\bar{\boldsymbol{p}}$. Therefore $(Q, \boldsymbol{\Theta}, \boldsymbol{p})$ are not generically identifiable and the proof of the theorem is complete.

C.7 Proof of Proposition IV.4

We show the conclusion following a similar argument as the proof of Proposition 1 in Xu and Shang (2018). To establish the bound (4.7) in the proposition, we check the technical conditions in Theorem 1 in Shen et al. (2012b). We first define some notations. For a family of probability mass functions $\mathcal{F}$, define $H(\cdot, \mathcal{F})$ to be the bracketing Hellinger metric entropy of $\mathcal{F}$. We call a finite set of function pairs $S(\epsilon, n) = \{(f_1^l, f_1^u), \dots, (f_n^l, f_n^u)\}$ a *Hellinger ϵ-bracketing* of $\mathcal{F}$ if the L_2 norm $\left\|\sqrt{f_i^l} - \sqrt{f_i^u}\right\| \leq \epsilon$ for all $i = 1, \dots, n$; and further fur any $f \in \mathcal{F}$, there is an i such that $f_i^l \leq f \leq f_i^u$. The *bracketing Hellinger metric entropy* is defined to be the logarithm of the cardinality of the ϵ-bracketing with the smallest size, namely $H(\cdot, \mathcal{F}) = \log \min\{n : S(\epsilon, n)\}$. We next argue that the size of the parameter space of $(\boldsymbol{\Theta}, \boldsymbol{p})$ is well controlled under the Hellinger metric. Recall S is defined in the main text before Proposition IV.4, and we define $\mathcal{B}_S = \mathcal{F}_S \cap \{h(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \leq 2\epsilon\}$ as the local parameter space with $\boldsymbol{\eta} = (\boldsymbol{B}, \boldsymbol{p})$ denoting general model parameters and $\boldsymbol{\eta}^0 = (\boldsymbol{B}^0, \boldsymbol{p}^0)$ denoting the true model parameters. According to the argument in the proof of Proposition 1 in Xu and Shang (2018), in the considered scenario with fixed J and K, for any $\epsilon < 1$ and any $t \in (\epsilon/2^4, \epsilon)$, there is $H(t, \mathcal{B}_S) \leq c \log(J2^K)|S| \log(2\epsilon/t)$; indeed, there is $H(t, \mathcal{B}_S) = O(\log(2\epsilon/t))$ uniformly for any S, ϵ and t.

With this upper bound on the Hellinger bracketing entropy, we can apply Theorem 1 in Shen et al. (2012b) to obtain

$$\mathbb{P}(\widehat{Q} \neq Q^0) \leq \mathbb{P}(\widehat{\boldsymbol{\eta}} \neq \widehat{\boldsymbol{\eta}}^0) \leq c_2 \exp\{-c_1 N C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)\},$$

where $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) := \inf_{\boldsymbol{\eta}\colon |S| \leq m, S \neq S_0} h^2(\boldsymbol{\eta}, \boldsymbol{\eta}^0)$. The above display is the desired (4.7) in the proposition.

Next we show that when the proposed sufficient conditions for joint strict identifiability hold, the $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0)$ in (4.7) is bounded away from zero by some positive constant. When the proposed conditions for joint strict identifiability (such as Conditions A, B and C under DINA model are satisfied), the $(\boldsymbol{B}^0, \boldsymbol{p}^0)$ here are strictly identifiable. The consequence is that there exists a constant $\delta > 0$ such that $h^2(\boldsymbol{\eta}, \boldsymbol{\eta}^0) \geq \delta$, where the m denotes the number of free parameters under the Q^0 and the RLCM specification. Therefore,

$$C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) \geq \inf_{\boldsymbol{\eta}\colon |S| \leq m, S \neq S_0} \frac{h^2(\boldsymbol{\eta}, \boldsymbol{\eta}^0)}{2m} \geq \frac{\delta}{2m} > 0,$$

so $C_{\min}(\boldsymbol{\Theta}^0, \boldsymbol{p}^0) \geq c_0$ for some positive constant c_0 holds. This proves the conclusion that under the proposed strict identifiability conditions, the finite sample error bound $\mathbb{P}(\widehat{Q} \neq Q^0)$ has an exponential rate. This completes the proof of the proposition.

C.8 Simulation Studies for Chapter IV

In this section, we provide more simulation results to verify the developed identifiability theory. In Section C.8.1, we perform simulation studies to verify Theorems

1 and 2 for the DINA model. In Section C.8.2, we perform simulation studies to verify Theorems 3 and 4 for the GDINA model. The Matlab code for performing the simulation studies are available at `https://github.com/yuqigu/Identify_Q`.

To better illustrate the identifiability or non-identifiability phenomena of Q-matrix, in some of the following simulation studies, we conduct exhaustive search of all possible Q-matrices of a certain size 5×2. Specifically, consider the set of all the 5×2 binary Q-matrices other than those containing some all-zero row vectors. If treating two Q-matrices that are identical up to permuting the two columns as equivalent (because they are indeed equivalent in terms of model identifiability), then there are in total 121 types of Q-matrices. We denote such a set of Q-matrices by $\text{Exhaus}(Q_{5\times 2})$, and denote its elements by $Q^1, Q^2, \ldots, Q^{121}$. For example, the first three and the last three Q-matrices in $\text{Exhaus}(Q_{5\times 2})$ are

$$Q^1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}; \quad Q^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}; \quad Q^3 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}; \quad \cdots\cdots$$

$$Q^{119} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}; \quad Q^{120} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}; \quad Q^{121} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The complete list of the 121 Q-matrices in the set $\text{Exhaus}(Q_{5\times 2})$ is available in the Matlab file `Q_aa.mat` at `https://github.com/yuqigu/Identify_Q`.

In the exhaustive-search scenario, to illustrate the identifiability/non-identifiability phenomenon, we will generate data using some particular Q-matrix, and fit the dataset using all the 121 candidate Q-matrices in $\text{Exhaus}(Q_{5\times 2})$ and plot the log-likelihood values corresponding to all these 121 Q-matrices. Investigating whether the true data-generating Q-matrix achieves the maximum of the likelihood would help gain insight into whether this true Q-matrix is identifiable in the considered practical setting. We will see from these simulations how the developed identifiability theory matches the practice.

C.8.1 Two-Parameter SLAM: DINA Model

In this section, we carry out four simulation studies.

Study I: When Q-matrix satisfies the necessary and sufficient conditions A, B and C for strict identifiability.

In this simulation study, we choose those Q-matrices from $\text{Exhaus}(Q_{5\times 2})$ that satisfies the proposed necessary and sufficient identifiability conditions A, B and C in Theorem IV.1 in Chapter IV. In particular, after rearranging rows, there are exactly two forms the 5×2 Q-matrix that satisfies A, B and C. Their representatives are Q^{18} and Q^{15} as follows,

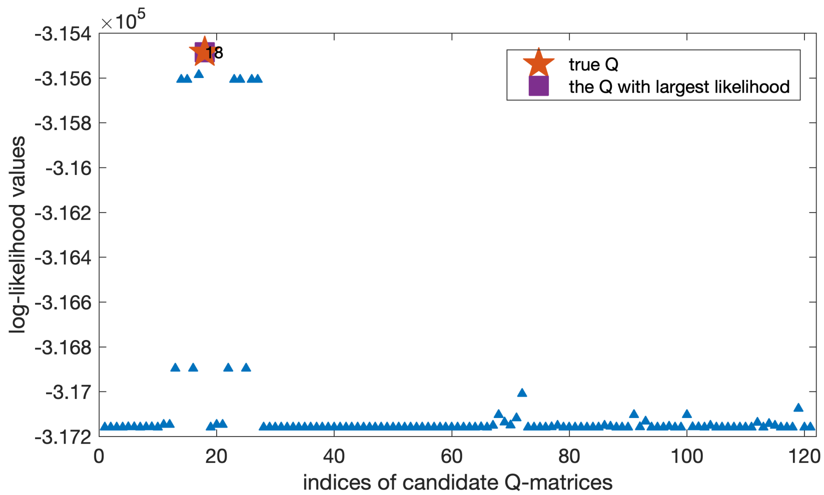
$$Q^{18} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}; \qquad Q^{15} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that Q^{18} contains only on identity submatrix I_2, while Q^{15} contains two copies of submatrix I_2. As introduced prior to this section C.8.1, we generate datasets with sample size $N = 10^5$ with true Q-matrix being Q^{18} and Q^{15}, respectively; and for each case, we run EM algorithm with several random initializations to fit the dataset with all the 121 Q-matrixes in $\text{Exhaus}(Q_{5\times 2})$ and obtain their log-likelihood values.

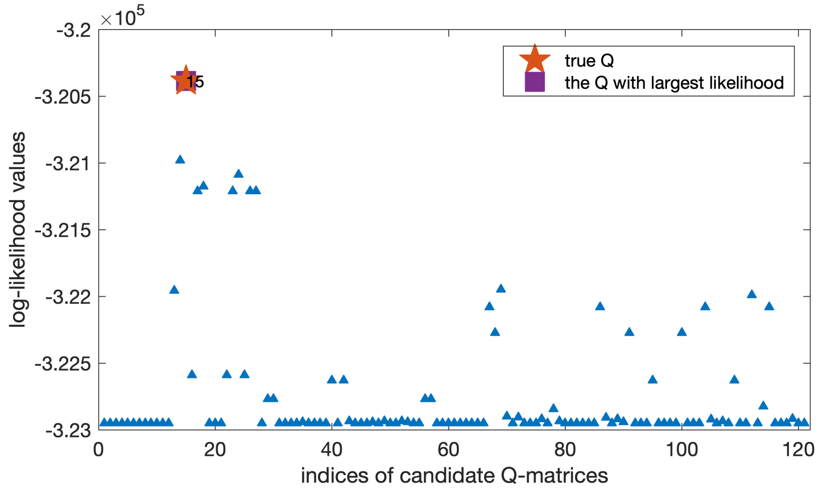
Figure C.1a and C.1b present the log-likelihood plots, with x-axis denoting the indices of the 121 candidate Q-matrices in $\text{Exhaus}(Q_{5\times 2})$, and y-axis denoting the log-likelihood values. Each blue triangle denotes a candidate Q-matrix; the red star denotes the true data-generating Q-matrix, and the purple square denotes the Q-matrix that achieves the largest likelihood.

We can see from these two plots in Figure C.1 that when the true data-generating Q-matrix (Q^{15} and Q^{18}) satisfies our proposed conditions A, B and C, it indeed achieves the largest likelihood compared to all other possible candidate Q-matrices. Therefore for any algorithm seeking the maximum likelihood estimator of $(Q, \boldsymbol{\theta}^{+}, \boldsymbol{g}, \boldsymbol{p})$, the true Q-matrix can be identified and any other Q-matrix will not be confused with the true Q. Another observation from Figure C.1a and C.1b is that, for Q^{15} that contains one more identity submatrix I_2 than Q^{18}, the true Q can be relatively better distinguished from the other Q's due to the larger gap in the likelihood values. This phenomenon might imply that the more identity submatrices the true data-generating Q-matrix contain, the easier the estimation for the true structure would be.

Figure C.1: DINA: exhaustive search in the set of 5×2 Q-matrices with a true Q-matrix satisfying Conditions A, B and C in Theorem 1.



(a) true Q containing one I_2: $Q^{18} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}^{\top}$



(b) true Q containing two I_2's: $Q^{15} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix}^\top$

Study II: When Q-matrix does not satisfy all of Conditions A, B, C but satisfies conditions in Theorem 2 for generic identifiability.

In this simulation study, we take the data-generating Q-matrix from Exhaus($Q_{5\times 2}$) that do NOT satisfy some of Conditions A, B and C, but satisfy the conditions in Theorem 2 for joint generic identifiability of $(Q, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$. In particular, for the considered case of $K = 2$, the only possibility for (global) generic identifiability is scenario (b.2) described in Theorem 2, where Condition C is violated and some column of Q contains only two entries of "1". After rearranging the rows of Q, it is not hard to see that there is only one possible case of the form of Q leading to generic identifiability, and the following Q^5 is a representative,

$$Q^5 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}. \tag{C.45}$$

The log-likelihood value plot is presented in Figure C.2. One can see in this generically identifiable scenario, with randomly generated true parameters, the true Q-matrix Q^5 achieves the largest likelihood and hence can be identified from data. We point out that although only the result of one simulated dataset is presented here, the generically identifiable Q-matrix (as the true Q-matrix) generally can achieve the largest likelihood among all the candidate Q-matrices, based on our experience in various simulations.

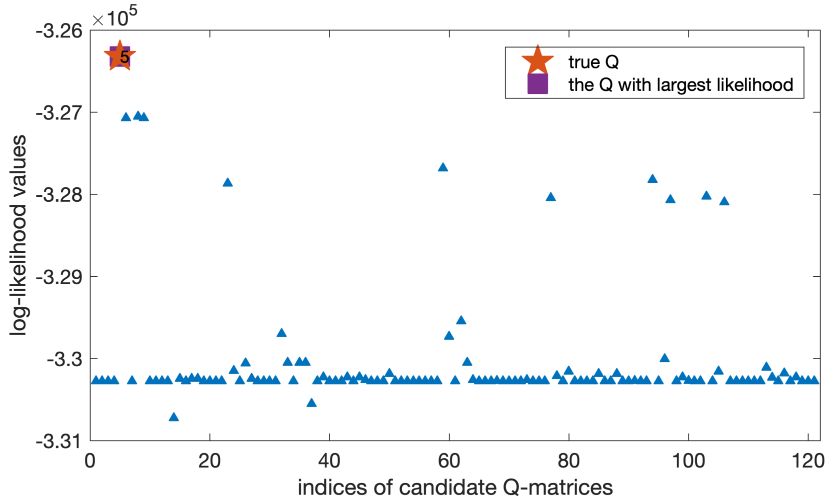


Figure C.2: DINA: exhaustive search in 5×2 Q-matrices with a true Q-matrix Q^5 in (C.45) generically identifiable, corresponding to scenario (b.2) in Theorem 2.

Study III: When Q-matrix does not even lead to local identifiability.

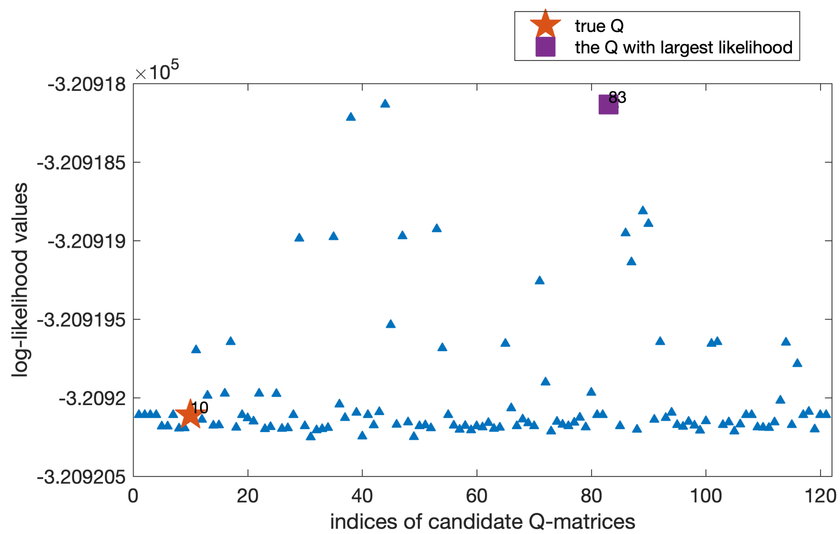
In this simulation study, we take the data-generating Q-matrix from Exhaus$(Q_{5\times 2})$ that *do not even lead to local identifiability.* That is, under such true Q-matrix, even in a small neighborhood of the true parameters, there exist infinitely many different alternative parameters that are not distinguishable from the true one.

Consider the following three different forms of Q-matrices from the set Exhaus$(Q_{5\times 2})$,

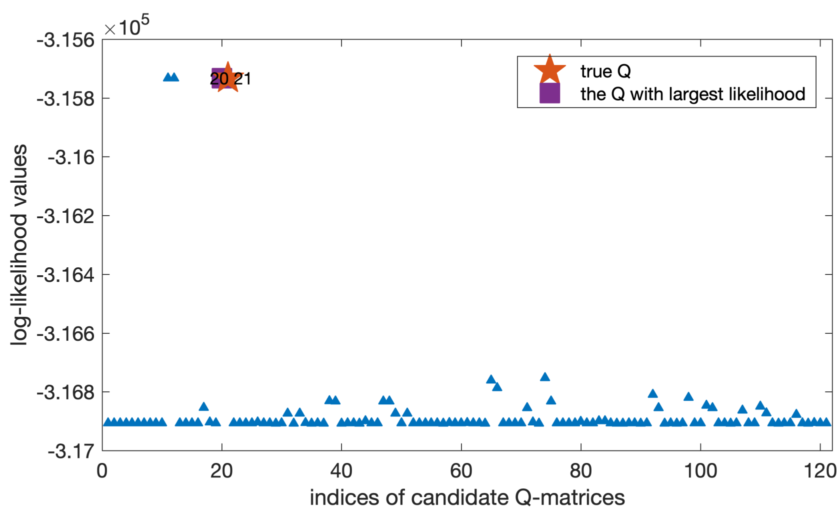
$$Q^{10} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad Q^{21} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}; \quad Q^{55} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix},$$

where Q^{10} contains only one entry of "1" in one column, Q^{21} is incomplete (i.e., lacks I_2), and Q^{55} contains an all-zero column. Their corresponding log-likelihood plots in the exhaustive-search scenario are presented in Figure C.3a, C.3b and C.3c. One can see from these plots that in these no even locally identifiable settings, the true data-generating Q-matrix does not achieve the maximum of the likelihood. Instead, many other alternative Q-matrices would have larger likelihood, and a wrong Q-matrix will be selected as the maximum likelihood estimator.

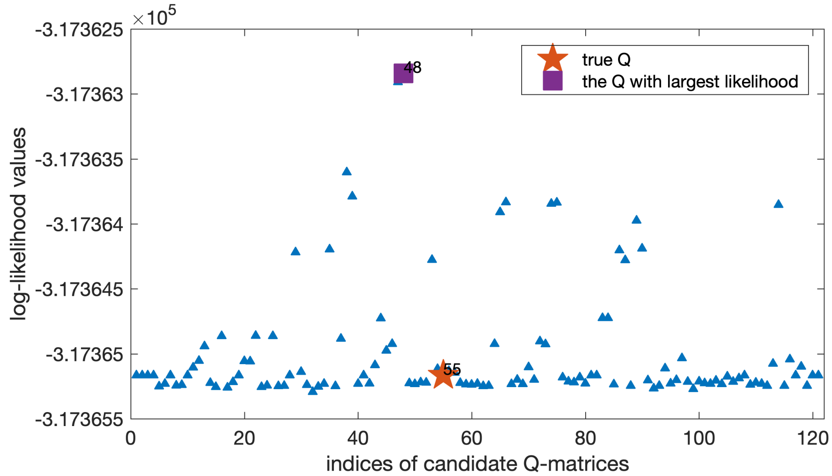
Figure C.3: DINA: exhaustive search in 5×2 Q-matrices with a true Q-matrix *not even locally* identifiable, corresponding to scenario (b.1) in Theorem 2.



(a) true Q not even locally identifiable: $Q^{10} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}^\top$



(b) true Q not even locally identifiable: $Q^{21} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}^\top$



(c) true Q not even locally identifiable: $Q^{55} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}^\top$

Study IV: Verifying necessity of Condition A "completeness".

We verify the necessity of Condition A "*completeness*" of the Q-matrix for identifiability. Consider two settings of incomplete Q-matrices, Q_1 with $(K, J) = (3, 20)$ and Q_2 with $(K, J) = (5, 20)$. For $i = 1, 2$, for the matrix $Q = Q_i$ and arbitrary DINA model parameters $(\boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$, we follow our theoretical derivations to construct two alternative Q-matrices $Q' = Q_i'$ and $Q'' = Q_i''$ and corresponding parameters $(\boldsymbol{\theta}^{+'}, \boldsymbol{g}', \boldsymbol{p}')$ and $(\boldsymbol{\theta}^{+''}, \boldsymbol{g}'', \boldsymbol{p}'')$. Then we compute the marginal probabilities for all the possible $2^{20} \approx 10^6$ response patterns under each of the Q, Q' and Q'', which characterize the distribution of the 20-dimensional binary vector $\boldsymbol{R}$. We give visualization plots to show how these different Q-matrices and different model parameters lead to exactly the same distribution of the observed responses $\boldsymbol{R}$.

$$
Q_1 = \begin{pmatrix} 1&0&0\\0&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&1&1\\1&1&1 \end{pmatrix}_{20\times 3} \quad
Q_1' = \begin{pmatrix} 1&0&0\\0&1&0\\\mathbf{0}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{1}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{1}&\mathbf{1}\\1&1&1\\1&1&1 \end{pmatrix}_{20\times 3} \quad
Q_1'' = \begin{pmatrix} 1&0&0\\0&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&0&0\\1&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&1&1\\1&1&1 \end{pmatrix}_{20\times 3} \tag{C.46}
$$

$$
Q_2 = \begin{pmatrix} 1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\\mathbf{1}&\mathbf{1}&\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\1&1&1&1&1\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\1&1&1&1&1 \end{pmatrix}_{20\times 5} \quad
Q_2' = \begin{pmatrix} 1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{1}&\mathbf{1}&\mathbf{1}\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\1&1&1&1&1\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\1&1&1&1&1 \end{pmatrix}_{20\times 5} \quad
Q_2'' = \begin{pmatrix} 1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{0}&\mathbf{1}\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\1&1&1&1&1\\1&0&0&0&0\\1&1&0&0&0\\1&1&1&0&0\\1&1&1&1&0\\1&1&1&1&1 \end{pmatrix}_{20\times 5} \tag{C.47}
$$

First, consider the following Q_1 with $(K, J) = (3, 20)$ in (C.46), which is incomplete because its row vectors does not contain the unit vector $(0, 0, 1)$. For arbitrarily generated parameters $(\boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$, we set $\boldsymbol{\theta}^{+''} = \boldsymbol{\theta}^{+'} = \boldsymbol{\theta}^+$ and $\boldsymbol{g}^{''} = \boldsymbol{g}^{'} = \boldsymbol{g}$ and set the proportion parameters as follows,

$$\begin{cases} p'_{(011)} = 0, \\ p'_{(010)} = p_{(010)} + p_{(011)}, \\ p'_{\boldsymbol{\alpha}} = p_{\boldsymbol{\alpha}},\ \forall \boldsymbol{\alpha} \neq (011), (010); \end{cases} \qquad \begin{cases} p''_{(001)} = p''_{(011)} = p''_{(111)} = 0, \\ p''_{(000)} = p_{(000)} + p_{(001)}, \\ p''_{(010)} = p_{(010)} + p_{(011)}, \\ p''_{(110)} = p_{(110)} + p_{(111)}, \\ p''_{\boldsymbol{\alpha}} = p_{\boldsymbol{\alpha}},\ \forall \boldsymbol{\alpha} = (100), (101). \end{cases} \tag{C.48}$$

We define a notation $\Gamma(Q)$ to briefly explain the rationale behind the above constructions. The $\Gamma(Q)$ is a $J \times 2^K$ binary matrix defined based on Q. The columns and rows of $\Gamma(Q)$ are indexed by the J items and the 2^K possible attribute patterns, respectively; and the $(j, \boldsymbol{\alpha})$th entry of it is defined to be $\Gamma_{j,\boldsymbol{\alpha}}(Q) = I(\boldsymbol{\alpha} \succeq \boldsymbol{q}_j)$. An important observation is that, due to the forms of Q, Q' and Q'', the unique column vectors in $\Gamma(Q)$ form a subset of those of $\Gamma(Q')$; and further the unique column vectors of $\Gamma(Q')$ form a subset of those of $\Gamma(Q'')$. Therefore, to construct $\boldsymbol{p}'$ such that $(Q, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$ and $(Q', \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}')$ that are non-distinguishable, we only need to set $p'_{\boldsymbol{\alpha}} = 0$ for those $\boldsymbol{\alpha}$ whose corresponding column vector in $\Gamma(Q')$ does not appear as the column vector of $\Gamma(Q)$; and let the proportions (in vector $\boldsymbol{p}'$) of other attribute patterns to absorb the proportions of these $\boldsymbol{\alpha}$'s in the vector $\boldsymbol{p}'$. The proportions $\boldsymbol{p}''$ under Q'' are constructed similarly. This is exactly how Equation (C.48) are derived. For the Q_2, Q'_2 and Q''_2 defined in (C.47), we construct the proportion parameters $\boldsymbol{p}'$ under Q'_2 and $\boldsymbol{p}''$ under Q''_2 following the same rationale; the details of defining them are omitted but their values are later revealed in Figure C.5(c).

In Figure C.4, we visualize the non-identifiability phenomenon of Q_1. In Figure C.4(a), we plot the differences of proportions parameters under the alternative models and the true model with Q_1. The red dotted line with "×" plots the values $\boldsymbol{p}' - \boldsymbol{p} = (p'_{000} - p_{000},\ p'_{001} - p_{001},\ p'_{010} - p_{010},\ p'_{011} - p_{011},\ p'_{100} - p_{100},\ p'_{101} - p_{101},\ p'_{110} -$

$p_{110},\ p'_{111} - p_{111}$) correspondent to the 8 attribute patterns; and the green dotted line with "+" plots $\boldsymbol{p}'' - \boldsymbol{p}$. Despite these three sets of parameters are quite different, the 2^{20}-dimensional vector of marginal probabilities of $\boldsymbol{R}$ are exactly the same, as shown in plots (b) and (c) of Figure C.4. In particular, in plot (b), the x-axis presents the indices of the response patterns in $\boldsymbol{r} \in \{0,1\}^J$, the y-axis presents the values of $\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p})$, where the blue circles denote those under $(Q_1, \boldsymbol{p})$, red "$\times$" for $(Q'_1, \boldsymbol{p}')$, and green "+" for $(Q''_1, \boldsymbol{p}'')$. Plot (c) of Figure C.4 is a zoomed-in version of plot (b), by only showing those marginal probabilities in $[0.2 \times 10^{-4}, 2 \times 10^{-4}]$, which contains around 7×10^3 response patterns. One can roughly see from both plots (b) and (c) that the three underlying parameters yield identical distribution of the response vector. Indeed, the computation carried out using Matlab yields

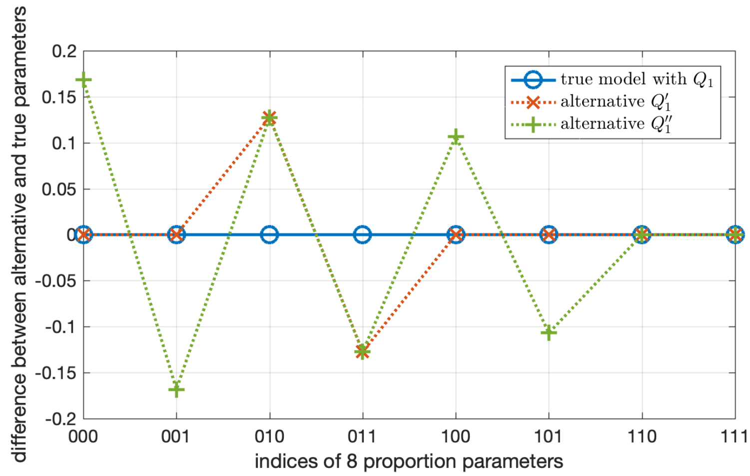
$$\max_{\boldsymbol{r} \in \{0,1\}^{20}} \left| \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q_1, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}) - \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q'_1, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}') \right| = 2.17 \times 10^{-19},$$
$$\max_{\boldsymbol{r} \in \{0,1\}^{20}} \left| \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q_1, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}) - \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q''_1, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}'') \right| = 4.34 \times 10^{-19},$$

which are both smaller than the machine epsilon (machine error) of Matlab 2.22×10^{-16}. This confirms that Q_1 defined in (C.46) is not identifiable.

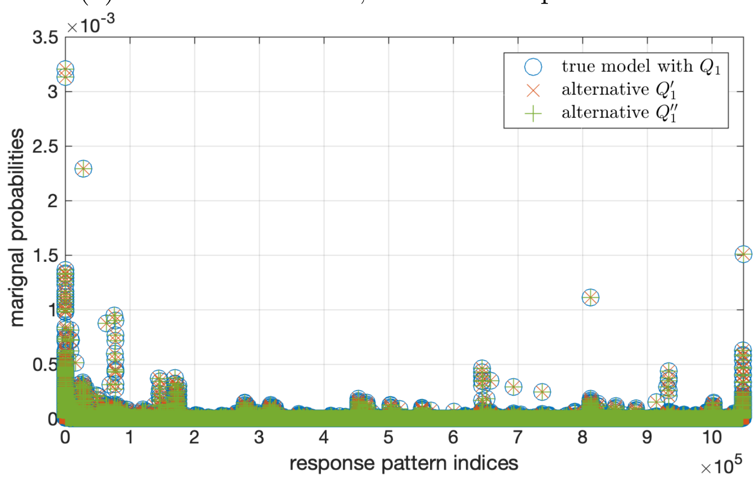
Figure C.5 shows the analogous results for Q_2 of size 20×5. Plot (a) in Figure C.5 shows the difference of the $2^5 = 32$-dimensional proportion parameters under alternative and true Q-matrices, and plots (b) and (c) give marginal probabilities of $\boldsymbol{R}$. The computation using Matlab gives

$$\max_{\boldsymbol{r} \in \{0,1\}^{20}} \left| \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q_2, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}) - \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q'_2, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}') \right| = 2.17 \times 10^{-19},$$
$$\max_{\boldsymbol{r} \in \{0,1\}^{20}} \left| \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q_2, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}) - \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q''_2, \boldsymbol{\theta}^+, \boldsymbol{g}, \boldsymbol{p}'') \right| = 6.51 \times 10^{-19},$$

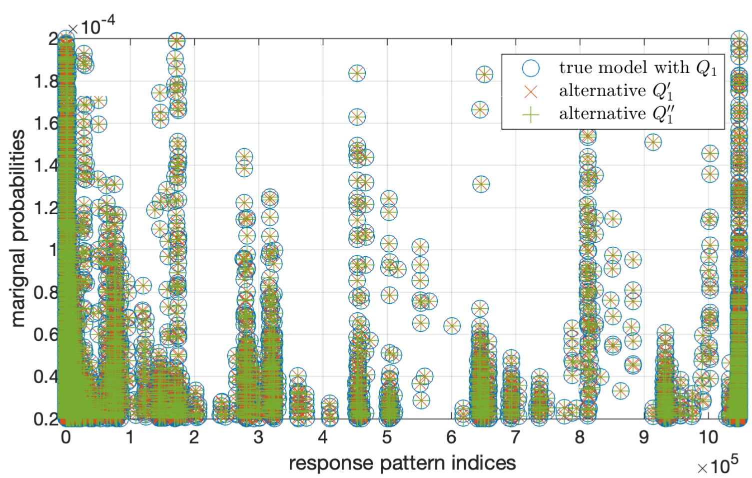
which are also both smaller than the machine error 2.22×10^{-16} of Matlab. This verifies the non-identifiability of Q_2 defined in (C.47).



(a) $K = 3$ and $J = 20$, three sets of parameters

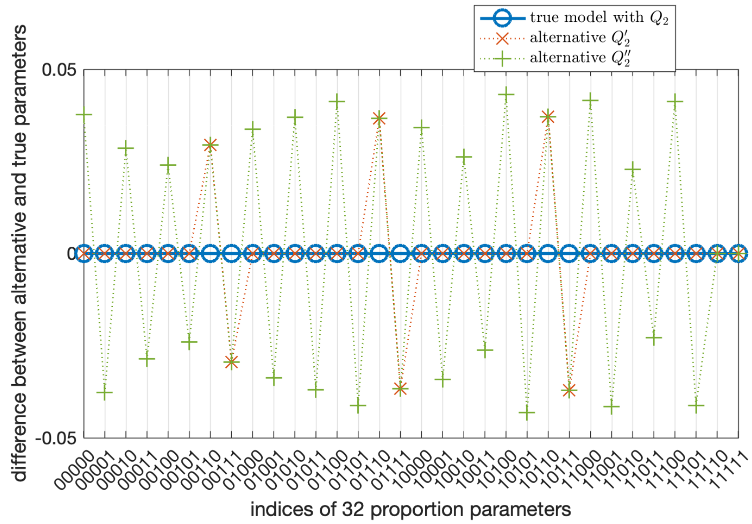


(b) $K = 3$ and $J = 20$, $|\{0,1\}^{20}| = 2^{20}$ response probabilities

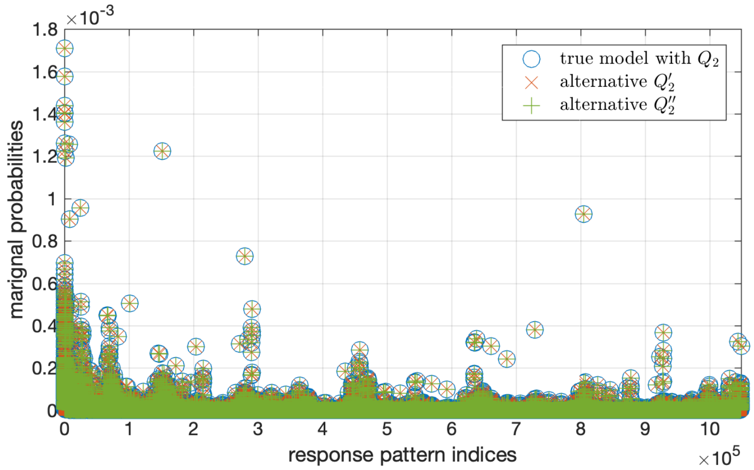


(c) $K = 3$ and $J = 20$, response probabilities zoomed in

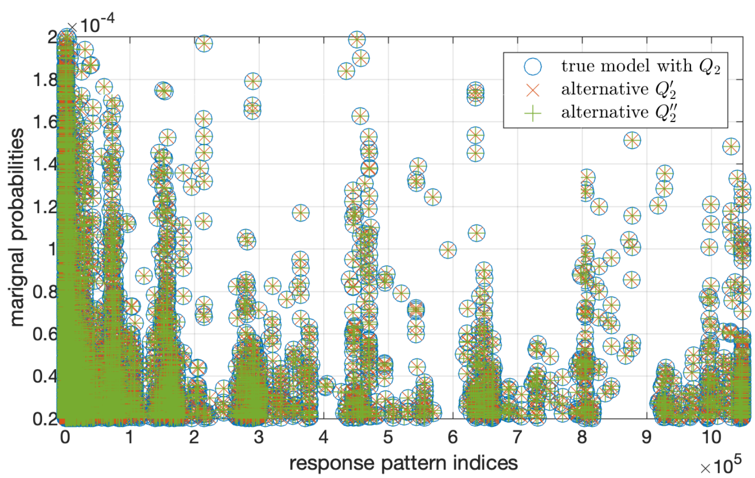
Figure C.4: DINA: true Q-matrix of size 20×3 is not complete and hence not identifiable.



(a) $K = 5$ and $J = 20$, three sets of parameters



(b) $K = 5$ and $J = 20$, $|\{0,1\}^{20}| = 2^{20}$ response probabilities



(c) $K = 5$ and $J = 20$, response probabilities zoomed in

Figure C.5: DINA: true Q-matrix of size 20×5 is not complete and hence not identifiable.

C.8.2 General SLAM: GDINA Model

In this section, we design simulation studies to verify the proposed identifiability conditions under the GDINA model introduced in Example I.3. In Study V, we use exhaustive search within 5×2 Q-matrices to verify the sufficient conditions in Theorem IV.4. In Study VI and Study VII, we verify the necessary conditions in Theorem IV.3.

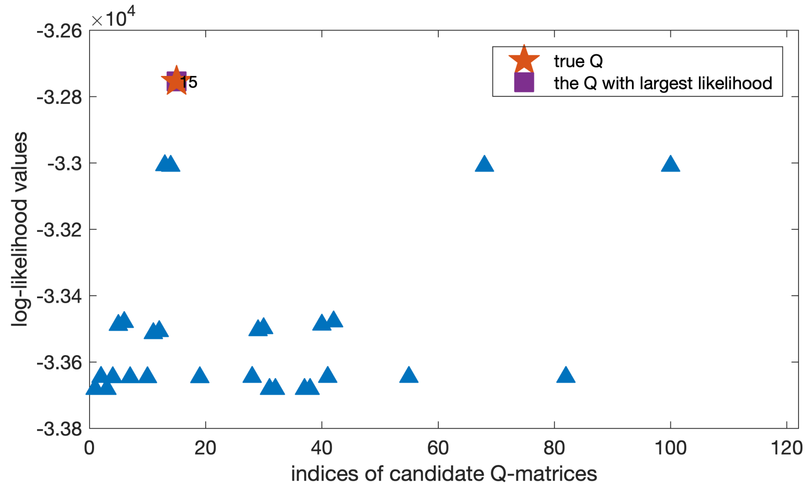
Study V: When Q-matrix satisfies Conditions D, E for generic identifiability.

Within the set of 5×2 Q-matrices Exhaus($Q_{5\times 2}$), if Q satisfies the sufficient conditions D and E for generic identifiability under the GDINA model, then other than the all-one Q-matrix Q^{121} which corresponds to the unrestricted latent class model, Q can take the forms of Q^{15}, Q^{18}, Q^{27}, Q^{54}, and Q^{81} (up to rearrangement of rows and columns). When using some Q-matrix to generate data, we also set the sample size to $N = 10^5$ and randomly set the true parameters which satisfy the monotonicity constraint (1.3) in the main text. In plots (a), (b), (c), (d) and (e) in Figure C.6, we present the exhaustive search results when the true data-generating Q-matrix is Q^{15}, Q^{18}, Q^{27}, Q^{54}, or Q^{81}. We point out that for GDINA model, in each scenario, we did not plot all the 121 Q-matrices' log-likelihood values, although we fit all the 121 ones to the simulated data. Instead, we only plot those Q-matrices under which the estimated parameters satisfies the stringent monotonicity constraint

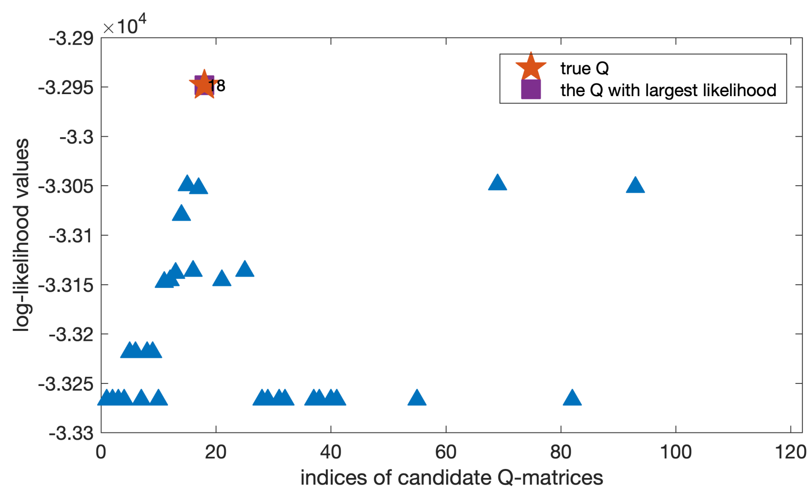
$$\theta_{j,\boldsymbol{\alpha}} > \theta_{j,\boldsymbol{\alpha}'} \text{ if } \boldsymbol{\alpha} \odot \boldsymbol{q}_j \succ \boldsymbol{\alpha}' \odot \boldsymbol{q}_j. \tag{C.49}$$

This constraint is stronger than requiring merely (1.3), and it is often imposed in practice when fitting the general RLCM that models the main and interaction effects of the latent attributes; for example, see the LCDM proposed in Henson et al. (2009). So each blue triangle in each plot of Figure C.6 corresponds to a Q-matrix with estimated $\mathbf{\Theta}$ satisfying (C.49). We can see from the five plots in Figure C.6 that when the generic identifiability conditions D and E are satisfied, the true data-generating Q-matrix achieves the maximum of the data likelihood compared to all the candidate Q-matrices of the same size.

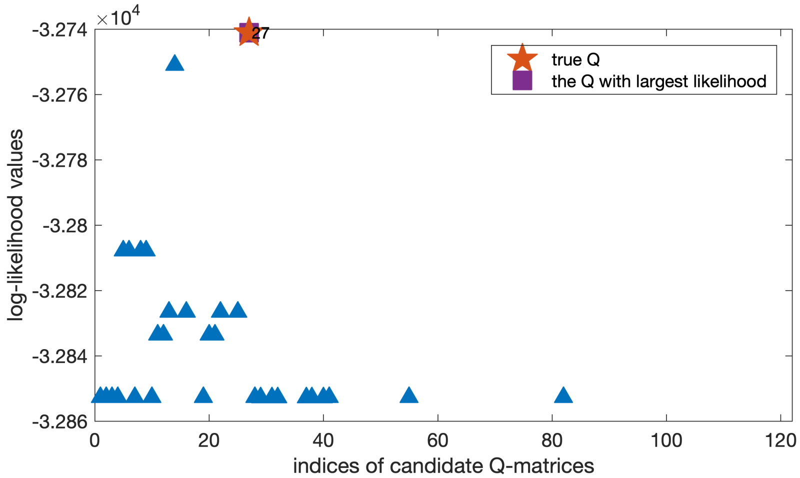
Figure C.6: GDINA: exhaustive search in 5×2 Q-matrices with a true Q satisfying Conditions D and E.



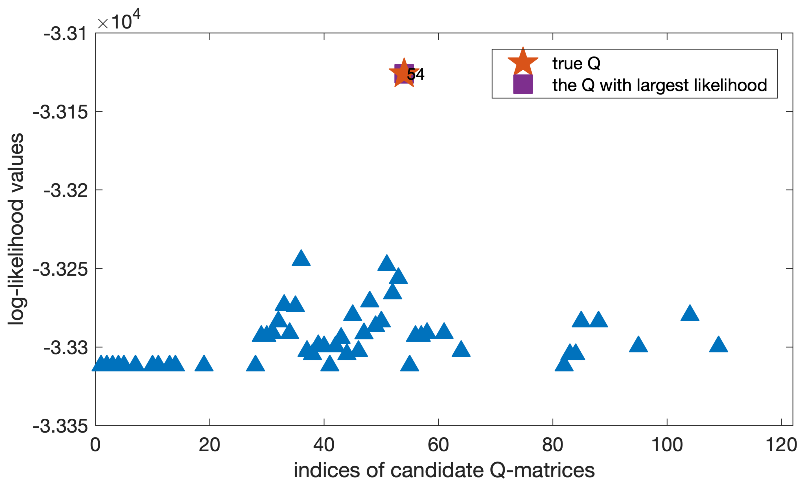
(a) GDINA: generically identifiable: $Q^{15} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix}^\top$



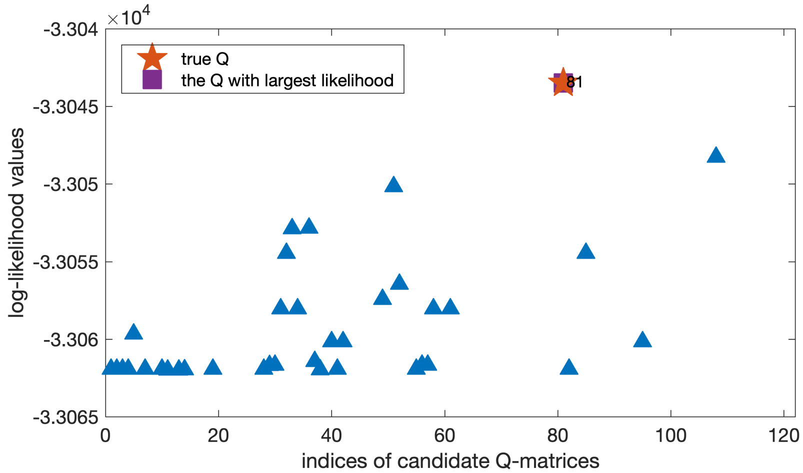
(b) GDINA: generically identifiable: $Q^{18} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}^\top$



(c) GDINA: generically identifiable: $Q^{27} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}^\top$



(d) GDINA: generically identifiable: $Q^{54} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}^\top$



(e) GDINA: generically identifiable: $Q^{81} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}^\top$

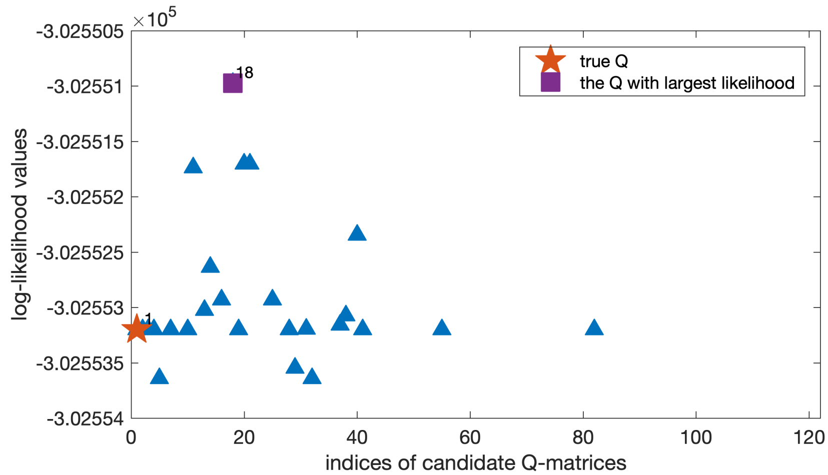
Study VI: When Q-matrix does not even lead to local generic identifiability.

We now use the not even locally generically identifiable Q-matrices Q^1, Q^2, or Q^3 to generate the data, and perform the exhaustive search among Exhaus($Q_{5\times 2}$). The log-likelihood plots along with the forms of the data generating matrices Q^1, Q^2, Q^3 are presented in Figure C.7. Similar to the previous Study V, here in each scenario we only plot those Q-matrix whose estimated $\boldsymbol{\Theta}$ parameters satisfy the stringent monotonicity constraint (C.49). One can see from the plots in Figure C.7 that these Q^1, Q^2, Q^3 do not maximize the data likelihood, implying severe non-identifiability. Note that for Figure C.7(b) corresponding to Q^2, there are only two Q-matrices satisfying the constraint (C.49) among the 121 Q-matrices fitted to the data; these two Q-matrices are the true Q-matrix Q^2 and another Q-matrix Q^{56},

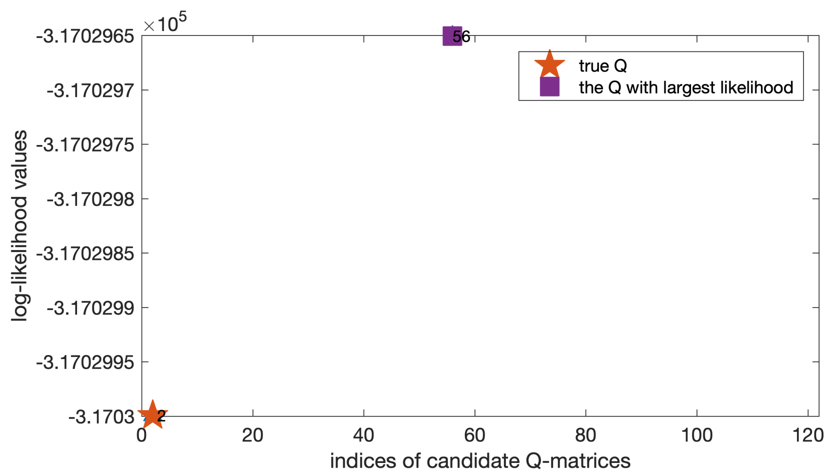
$$Q^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \qquad Q^{56} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Note that even there are only two Q-matrices satisfying the monotonicity constraint (C.49), the true Q^2 used to generate the data is not the one that has the larger likelihood, according to Figure C.7(b). This illustrates the non-identifiability of Q_2.

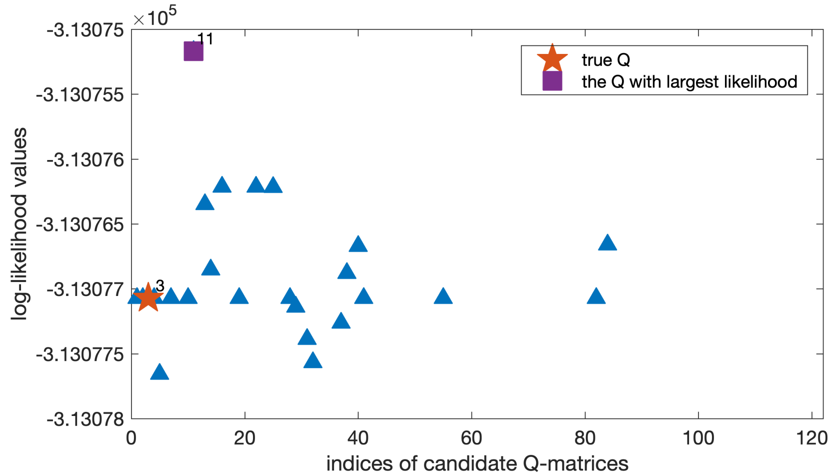
Figure C.7: GDINA: exhaustive search in 5×2 Q-matrices with a true Q-matrix which leads to a not even locally generically identifiable model.



(a) GDINA: true Q not even locally identifiable: $Q^1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}^\top$



(b) GDINA: true Q not even locally identifiable: $Q^2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \end{pmatrix}^\top$



(c) GDINA: true Q not even locally identifiable: $Q^3 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}^\top$

Study VII: Construction of many alternative sets of parameters when true Q-matrix violates the necessary condition for generic identifiability.

In this study, we verify Theorem IV.3, i.e., verify the necessity of Condition C that *each attribute is required by at least two items in the Q-matrix* for joint generic identifiability. We consider two cases with $(K, J) = (3, 20)$ and $(K, J) = (5, 20)$.

First, for $(K, J) = (3, 20)$, consider the following Q-matrix Q_3 and an alternative

$\bar{Q}_3$.

$$Q_3 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}_{20\times 3} \qquad \bar{Q}_3 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}_{20\times 3} \tag{C.50}$$

We first construct true parameters $(\mathbf{\Theta}, \boldsymbol{p})$ under Q_3. For each attribute pattern $\boldsymbol{\alpha}$, we set its population proportion $p_{\boldsymbol{\alpha}}$ to be $1/2^K$. For each item, set the baseline probability, the positive response probability of the all-zero attribute profile $\boldsymbol{\alpha} = \mathbf{0}^\top$, to be 0.2 and the positive response probability of $\boldsymbol{\alpha} = \mathbf{1}^\top$ to be 0.8. And we take all the main effects and interaction effects parameters to be equal.

For the defined true parameters $(\mathbf{\Theta}, \boldsymbol{p})$ under Q_3, we next construct 70 alternative sets of parameters $(\bar{\mathbf{\Theta}}^\ell, \bar{\boldsymbol{p}}^\ell)$ for $\ell = 1, 2, \ldots, 70$, all under the alternative Q-matrix $\bar{Q}_3$, that are non-distinguishable from the true parameters. Following the proof of Theorem IV.3, we first set $\bar{\theta}_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}}$ for any $j > 2$ and any $\boldsymbol{\alpha}$. Then we *randomly generate* the values of the $\bar{\mathbf{\Theta}}_{1:2,\,1:4}$ (the first four elements of the first two rows of $\bar{\mathbf{\Theta}}$) from the neighborhood of their true values, and enforce the monotonicity constraint (1.3). Specifically, for each alternative set (the ℓ-th set) of parameters, there is

$$\bar{\mathbf{\Theta}}_{i,j}^\ell = \mathbf{\Theta}_{i,j} + \mathcal{U}(-0.1, 0.1), \quad i = 1, 2;\ j = 1, 2, 3, 4;\ \ell = 1, 2, \ldots, 70.$$

where $\mathcal{U}(-0.1, 0.1)$ denotes a uniformly distributed random variable between -0.1

and 0.1. Next we just use Equation (C.41) to get the remaining item parameters $\bar{\mathbf{\Theta}}^{\ell}_{1:2,\,5:8}$ and $\bar{\boldsymbol{p}}^{\ell}$.

Figure C.8 presents the constructed 70 other parameters sets $(\bar{\mathbf{\Theta}}^{\ell}, \bar{\boldsymbol{p}}^{\ell})$ under the alternative $\bar{Q}_3$, by plotting the values of difference between the alternative parameters and the true parameters. In particular, In Figure C.8(a), the black solid line with dots is the reference line at zero, and each of the 70 colored dotted line with "+"'s represents one particular set of alternative parameters. For each colored line corresponding to the ℓth set of parameters, the following 16-dimensional vector of parameter difference is plotted,

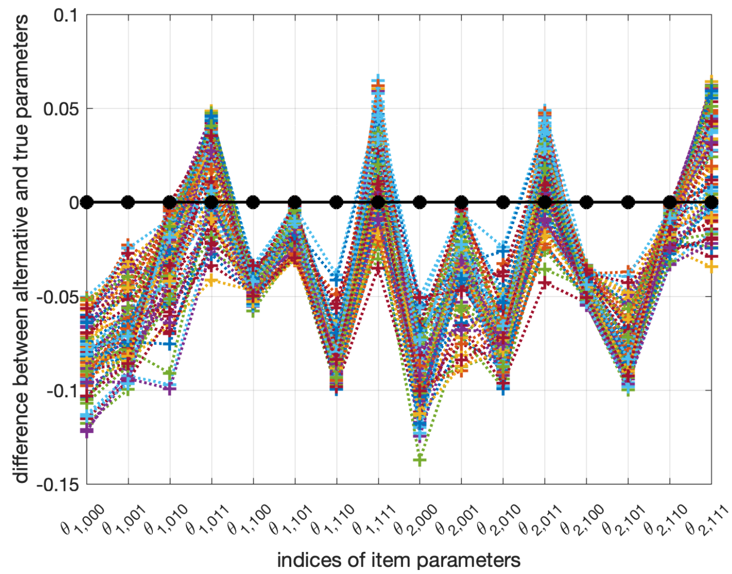
$$\begin{aligned}
(&\bar{\theta}^{\ell}_{1,000} - \theta_{1,000},\ \bar{\theta}^{\ell}_{1,001} - \theta_{1,010},\ \bar{\theta}^{\ell}_{1,010} - \theta_{1,010},\ \bar{\theta}^{\ell}_{1,011} - \theta_{1,011},\\
&\bar{\theta}^{\ell}_{1,100} - \theta_{1,100},\ \bar{\theta}^{\ell}_{1,101} - \theta_{1,110},\ \bar{\theta}^{\ell}_{1,110} - \theta_{1,110},\ \bar{\theta}^{\ell}_{1,111} - \theta_{1,111},\\
&\bar{\theta}^{\ell}_{2,000} - \theta_{2,000},\ \bar{\theta}^{\ell}_{2,001} - \theta_{2,010},\ \bar{\theta}^{\ell}_{2,010} - \theta_{2,010},\ \bar{\theta}^{\ell}_{2,011} - \theta_{2,011},\\
&\bar{\theta}^{\ell}_{2,100} - \theta_{2,100},\ \bar{\theta}^{\ell}_{2,101} - \theta_{2,110},\ \bar{\theta}^{\ell}_{2,110} - \theta_{2,110},\ \bar{\theta}^{\ell}_{2,111} - \theta_{2,111}).
\end{aligned}$$

Similarly, in Figure C.8(b), for each colored line corresponding to the ℓth set of parameters, the following 8-dimensional vector of parameter difference is plotted, $(\bar{p}^{\ell}_{000} - p_{000},\ \bar{p}^{\ell}_{001} - p_{010},\ \bar{p}^{\ell}_{010} - p_{010},\ \bar{p}^{\ell}_{011} - p_{011},\ \bar{p}^{\ell}_{100} - p_{100},\ \bar{p}^{\ell}_{101} - p_{110},\ \bar{p}^{\ell}_{110} - p_{110},\ \bar{p}^{\ell}_{111} - p_{111})$. In summary, a total number of 70 colored lines corresponding to 70 alternative sets of parameters are plotted in Figure C.8.

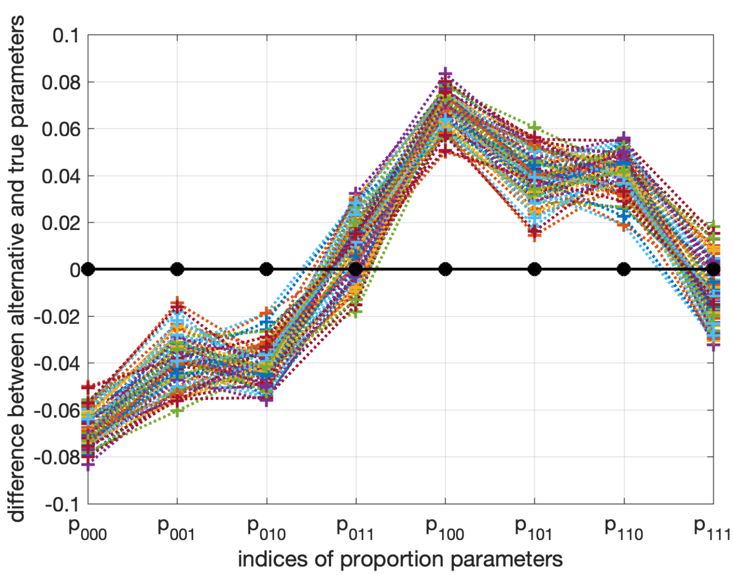
The $(\mathbf{\Theta}, \boldsymbol{p})$ and all the $(\bar{\mathbf{\Theta}}^{\ell}, \bar{\boldsymbol{p}}^{\ell})$, $\ell = 1, \ldots, 70$ give the identical distribution of $\boldsymbol{R}$. Specifically, from the computation in Matlab, we have

$$\max_{1\leq\ell\leq 70} \max_{\boldsymbol{r}\in\{0,1\}^{20}} \left|\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q_3, \mathbf{\Theta}, \boldsymbol{p}) - \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \bar{Q}_3, \mathbf{\Theta}^{\ell}, \boldsymbol{p}^{\ell})\right| = 1.30 \times 10^{-18},$$

which is smaller than the Matlab machine error 2.22×10^{-16}. This verifies that despite the underlying parameters are different from the truth, they all lead to the identical distribution of responses. So $(Q_3, \boldsymbol{\Theta}, \boldsymbol{p})$ are not identifiable. We emphasize that under this Q_3, for any true parameters, one can similarly construct arbitrarily many such alternative parameter sets under $\bar{Q}_3$.



(a) $K = 3$ and $J = 20$, 70 alternative sets of parameters



(b) $K = 3$ and $J = 20$, 70 sets of parameters

Figure C.8: GDINA: true Q is Q_3 with $(K, J) = (3, 20)$; each of the 70 colored line corresponds to one set of alternative parameters under $\bar{Q}_3$; all sets non-distinguishable.

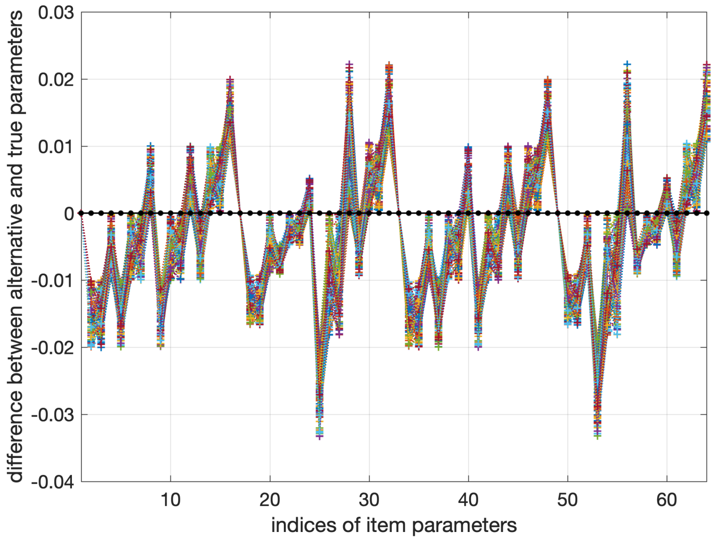
For $(K, J) = (5, 20)$, consider the following Q_4 and an alternative $\bar{Q}_4$,

$$Q_4 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}_{20\times 5} \qquad \bar{Q}_4 = \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}_{20\times 5}. \tag{C.51}$$

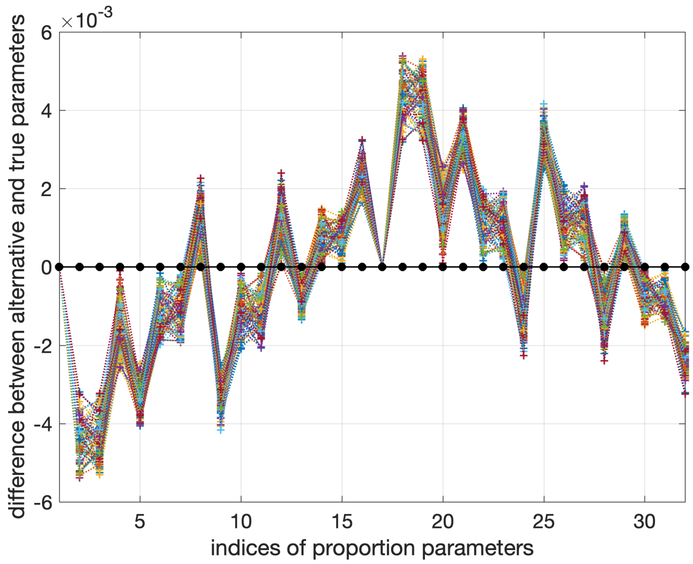
We set the true parameters under Q_4 similarly as those under Q_3, and also use (C.41) in the proof of Theorem IV.3 to randomly construct 70 sets of parameters under the $\bar{Q}_4$. Figure C.9 (a) and (b) plot the values of difference between alternative and true item parameters (of the first two items), and that between alternative and true proportion parameters, respectively. Despite the differences in parameter values, our computation in Matlab shows the maximum difference between marginal response probabilities is

$$\max_{1\leq \ell \leq 70} \max_{\boldsymbol{r}\in\{0,1\}^{20}} \left| \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid Q_4, \boldsymbol{\Theta}, \boldsymbol{p}) - \mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \bar{Q}_4, \boldsymbol{\Theta}^\ell, \boldsymbol{p}^\ell) \right| = 5.42 \times 10^{-19},$$

also smaller than the Matlab machine error 2.22×10^{-16}. This illustrates the non-identifiability of Q_4.



(a) $K = 5$ and $J = 20$, 70 alternative sets of parameters



(b) $K = 5$ and $J = 20$, 70 alternative sets of parameters

Figure C.9: GDINA: true Q is Q_4 with $(K, J) = (5, 20)$; each of the 70 colored line corresponds to one set of alternative parameters under $\bar{Q}_4$; all sets non-distinguishable.

APPENDIX D

Appendix of Chapter V

This is the appendix to Chapter V and it is organized as follows. Appendix D.1 presents the proof of Theorem V.1 and Corollary V.1. Appendix D.2 presents the proof of Theorem V.2 and Corollary V.2. Appendix D.3 gives the proof of Corollary V.3. Appendix D.4 presents the proof of Theorem V.3 and Proposition V.1. Appendix D.5 presents the proof of Theorem V.4. Appendix D.6 presents some additional numerical results.

D.1 Proof of Theorem V.1 and Corollary V.1.

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

$$T(\Gamma, \mathbf{\Theta})\boldsymbol{p} = T(\bar{\Gamma}, \bar{\mathbf{\Theta}})\bar{\boldsymbol{p}} \tag{D.1}$$

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

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

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

We first show $T(\bar{\Gamma}^1, \bar{\mathbf{\Theta}}^1)$ and $T(\bar{\Gamma}^2, \bar{\mathbf{\Theta}}^2)$ both have full column rank L_0, and that $\bar{p}_l > 0$ for all $l \in \{1, \dots, L_0\}$. By Proposition 3 in Gu and Xu (2020a), Condition A ensures that $T(\Gamma^1, \mathbf{\Theta}^1)$ of size 2^{J_1} and $T(\Gamma^2, \mathbf{\Theta}^2)$ of size 2^{J_2} both have full column rank L_0, since Γ^1 and Γ^2 are both separable. Moreover, in the proof of that conclusion, an invertible square matrix W_1 of size $2^{J_1} \times 2^{J_1}$ as well as L_0 response patterns $\boldsymbol{r}_1, \dots, \boldsymbol{r}_{L_0} \in \{0, 1\}^L$ were constructed such that the row vectors in the transformed $W_1 \cdot T(\Gamma^1, \mathbf{\Theta}^1)$, which are indexed by the chosen $\boldsymbol{r}_1, \ \ldots, \ \boldsymbol{r}_{L_0}$, form a $L_0 \times L_0$ lower triangular matrix with nonzero diagonal elements. In other words, in the $2^{J_1} \times L_0$ rectangular matrix $W_1 T(\Gamma^1, \mathbf{\Theta}^1)$, there is a $L_0 \times L_0$ submatrix that is lower triangular and full-rank. For notational simplicity, we denote this submatrix by $\{W_1 T(\Gamma^1, \mathbf{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}$. Similarly, there exists W_2 and $\boldsymbol{r}'_1, \ \ldots, \ \boldsymbol{r}'_{L_0} \in \{0,1\}^{L_0}$ such that there is a $L \times L$ full-rank submatrix of $W_2 T(\Gamma^2, \mathbf{\Theta}^2)$ with rows indexed by $\boldsymbol{r}'_1, \ \ldots, \ \boldsymbol{r}'_{L_0}$, which we denote by $\{W_2 T(\Gamma^2, \mathbf{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}$.

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

$$\boldsymbol{u}_l^\top \cdot \{W_1 T(\Gamma^1, \mathbf{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}} = (\mathbf{0}, \underbrace{\ 1\ }_{\text{column } l}, \mathbf{0}). \tag{D.2}$$

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

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

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

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

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

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

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

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

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

which contradicts (D.1). This shows that $\bar{p}_l > 0$ must hold for all $l \in [L_0]$.

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

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

This is because otherwise there would exist $l \in [L_0]$ such that $\{U_1 \cdot T(\bar{\Gamma}^1, \bar{\mathbf{\Theta}}^1)\}_{\cdot,l}$ equals the zero vector, which contradicts the fact that both U_1 and $\{W_1 T(\bar{\Gamma}^1, \bar{\mathbf{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}$ are invertible matrices. Given the permutation σ, there exists a $L_0 \times L_0$ invertible matrix

V with row vectors denoted by $\{\boldsymbol{v}_l,\ l \in [L_0]\}$ such that for each $\boldsymbol{\alpha} \in \mathcal{A}$,

$$\boldsymbol{v}_l^\top \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}} = (\mathbf{0}, \underbrace{1}_{\text{column } \sigma(l)}, \mathbf{0}). \tag{D.3}$$

Then we have

$$[\boldsymbol{u}_l^\top \cdot \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \boldsymbol{p} = f_l p_l, \tag{D.4}$$

$$[\boldsymbol{u}_l^\top \cdot \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}] \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\bar{\Gamma}^2, \bar{\boldsymbol{\Theta}}^2)\}_{\boldsymbol{r}'_{1:L_0}}] \cdot \bar{\boldsymbol{p}} = \bar{f}_{\sigma(l)} \bar{p}_{\sigma(l)} \neq 0, \tag{D.5}$$

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

$$\theta_{j,l} = \frac{T_{\boldsymbol{e}_j,\cdot}(\Gamma, \boldsymbol{\Theta}) \odot \text{(D.4)}}{\text{(D.4)}} = \frac{T_{\boldsymbol{e}_j,\cdot}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}}) \odot \text{(D.5)}}{\text{(D.5)}} = \bar{\theta}_{j,\sigma(l)}.$$

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

Lemma D.1. *Under the assumptions of Theorem V.1, the vectors* $\{\boldsymbol{v}_l,\ l \in \mathcal{A}_0\}$ *constructed in* (D.3) *satisfy that*

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

Proof of Lemma D.1 If $\{\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}\}_{l'} = z_{l'} \neq 0$, then similar to (D.4) and (D.5) we have

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

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

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

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

$$\begin{aligned} b_l &:= T_{\boldsymbol{r}^*,l}(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \mathbf{1}^\top) \neq 0; \\ T_{\boldsymbol{r}^*,l'}&(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \mathbf{1}^\top) = 0, \quad \forall \boldsymbol{\alpha}_{l'} \not\succeq_{S_1} \boldsymbol{\alpha}_l \text{ under } \Gamma^{\mathcal{A}_0}. \end{aligned} \tag{D.6}$$

The proof of Step 2 as well as Lemma D.1 ensures

$$\begin{aligned} f_l =& [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}]_l \neq 0; \\ & [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}_{\boldsymbol{r}'_{1:L_0}}]_{l'} = 0, \quad \forall \boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l \text{ under } \Gamma^{\mathcal{A}_0}. \end{aligned} \tag{D.7}$$

Consider any $j \in S_1 \cup S_2$ such that $\Gamma_{j,l} = 1$, then obviously $\boldsymbol{e}_j$ is not included in the sum in the previously defined response pattern $\boldsymbol{r}^*$, because $\boldsymbol{r}^*$ only contains those items that $\boldsymbol{\alpha}_l$ is not capable of, i.e., those j s.t. $\Gamma^{\mathcal{A}_0}_{j,l} = 0$. The above two equations (D.6) and (D.7) indicate

$$T_{\boldsymbol{r}^*,\cdot}(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}] = \Big(\mathbf{0}^\top, \underbrace{b_l \cdot f_l}_{\text{column } l}, \mathbf{0}^\top\Big), \tag{D.8}$$

$$T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\Gamma^1, \boldsymbol{\Theta}^1 - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot [\boldsymbol{v}_l^\top \cdot \{W_2 T(\Gamma^2, \boldsymbol{\Theta}^2)\}] = \Big(\mathbf{0}^\top, \underbrace{\theta_{j,H} \cdot b_l \cdot f_l}_{\text{column } l}, \mathbf{0}^\top\Big). \tag{D.9}$$

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

$$T_{\boldsymbol{r}^*,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^* \mathbf{1}^\top) \odot \{\boldsymbol{v}_l^\top \cdot T(\bar{\boldsymbol{\Theta}}^2)\} = \Big(\mathbf{0}^\top, \underbrace{\prod_{h \in S_1 : \Gamma_{h,l}=0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H}), \mathbf{0}^\top}_{\text{column } \sigma(l)}\Big), \tag{D.10}$$

$$T_{\boldsymbol{r}^*+\boldsymbol{e}_j,\cdot}(\bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top) \odot \{\boldsymbol{v}_l^\top \cdot T(\bar{\boldsymbol{\Theta}}^2)\} = \Big(\mathbf{0}^\top, \underbrace{\bar{\theta}_{j,H} \cdot \prod_{h\in S_1:\Gamma_{h,l}=0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H})}_{\text{column } \sigma(l)}, \mathbf{0}^\top\Big). \tag{D.11}$$

Equation (D.1) implies (D.8) $\cdot\, \boldsymbol{p} =$ (D.10) $\cdot\, \bar{\boldsymbol{p}}$. By (D.1), the above four equations give that

$$\theta_{j,H} = \theta_{j,l} = \frac{\text{(D.9)} \cdot \boldsymbol{p}}{\text{(D.8)} \cdot \boldsymbol{p}} = \frac{\text{(D.11)} \cdot \bar{\boldsymbol{p}}}{\text{(D.10)} \cdot \bar{\boldsymbol{p}}} = \bar{\theta}_{j,\sigma(l)} = \bar{\theta}_{j,H}, \quad \forall j \in S_2.$$

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

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

$$\forall j \in S_1 \cup S_2, \quad \theta_{j,l} = \bar{\theta}_{j,\sigma(l)}, \quad p_l = \bar{p}_{\sigma(l)}. \tag{D.12}$$

We first introduce the *lexicographic order* between two binary vectors of the same length. For two vectors $\boldsymbol{a} = (a_1, \ldots, a_L)$ and $\boldsymbol{b} = (b_1, \ldots, b_L)$, we say $\boldsymbol{a}$ has smaller lexicographic order than $\boldsymbol{b}$ and denote by $\boldsymbol{a} \prec_{\text{lex}} \boldsymbol{b}$, if either $a_1 < b_1$, or $a_l < b_l$ for some integer $l \le L$ and $a_m = b_m$ for all $m = 1, \ldots, l-1$. By Condition A, $\Gamma^{(S_i,\mathcal{A}_0)}$ has distinct column vectors for $i = 1, 2$, so without loss of generality, we can assume the columns of it are sorted in an increasing lexicographic order, i.e.,

$$\Gamma^{(S_1,\mathcal{A}_0)}_{\bullet,1} \prec_{\text{lex}} \cdots \prec_{\text{lex}} \Gamma^{(S_1,\mathcal{A}_0)}_{\bullet,L_0}. \tag{D.13}$$

Firstly, we prove (D.12) hold for $l = 1$, where from (D.13) we have $\Gamma^{(S_1,\mathcal{A}_0)}_{\bullet,1}$ has the smallest lexicographical order among the column vectors of $\Gamma^{(S_1,\mathcal{A}_0)}$. We claim that $\Gamma^{(S_2,\mathcal{A}_0)}_{\bullet,1}$ has the smallest lexicographical order among the column vectors of $\Gamma^{(S_2,\mathcal{A}_0)}$, because otherwise " $\succeq_{S_1}=\succeq_{S_2}$" under $\mathcal{A}_0$ will not hold. For $l=1$ we define

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

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

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

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

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

and

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

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

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\Gamma,\boldsymbol{\Theta}-\boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(\prod_{h\in S_1:\Gamma_{h,1}=0}(\theta_{h,1}-\theta_{h,H})\cdot\theta_{j,1},\ 0,\dots,0\Big), \tag{D.14}$$

and

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\bar{\Gamma},\bar{\boldsymbol{\Theta}}-\boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(0,\dots,0,\underbrace{\prod_{h\in S_1:\Gamma_{h,1}=0}(\bar{\theta}_{h,\sigma(1)}-\theta_{h,H})\cdot\bar{\theta}_{j,\sigma(1)},}_{\text{column }\sigma(1)}\ 0,\dots,0\Big), \tag{D.15}$$

respectively. The only potentially nonzero term in the first column of (D.14) is indeed nonzero, because we have $\theta_{h,1} < \theta_{h,H}$ for $h \in S_1$, $\Gamma_{h,1} = 0$. Now Equation (C.1) implies that

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

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

After obtaining these $\bar{\theta}_{j,\sigma(1)} = \theta_{j,1}$ for $j \in (S_1 \cup S_2)$ and $\Gamma_{j,1} = 0$, the previous equations (D.14) and (D.15) just become the following,

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(\prod_{h \in S_1:\Gamma_{h,1}=0} (\theta_{h,1} - \theta_{h,H}) \cdot \theta_{j,1},\ 0, \dots, 0\Big), \tag{D.16}$$

$$T_{\boldsymbol{r}+\boldsymbol{e}_j,\cdot}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top) = \Big(0, \dots, 0, \underbrace{\prod_{h \in S_1:\Gamma_{h,1}=0} (\theta_{h,\sigma(1)} - \theta_{h,H}) \cdot \theta_{j,\sigma(1)},}_{\text{column } \sigma(1)}\ 0, \dots, 0\Big). \tag{D.17}$$

Therefore (D.16) $\cdot\, \boldsymbol{p} =$ (D.17) $\cdot\, \bar{\boldsymbol{p}}$ just gives $p_1 = \bar{p}_{\sigma(1)}$.

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

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

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

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

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

$$T_{\boldsymbol{r}^*,\cdot}(\Gamma, \boldsymbol{\Theta} - \boldsymbol{\theta}^*\mathbf{1}^\top)\boldsymbol{p} = \sum_{\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l} t_{\boldsymbol{r}^*,l'} \cdot p_{l'} + \prod_{h\in S_1:\Gamma_{h,l}=0} (\theta_{h,l} - \theta_{h,H}) \prod_{h\in S_1:\Gamma_{h,l}=1} (\theta_{h,l} - \theta_{h,1}) \cdot p_l, \tag{D.18}$$

$$T_{\boldsymbol{r}^*,\cdot}(\bar{\Gamma}, \bar{\boldsymbol{\Theta}} - \boldsymbol{\theta}^*\mathbf{1}^\top)\bar{\boldsymbol{p}} = \sum_{\boldsymbol{\alpha}_{l'} \preceq_{S_1} \boldsymbol{\alpha}_l} \bar{t}_{\boldsymbol{r}^*,\sigma(l')} \cdot \bar{p}_{\sigma(l')} + \prod_{h\in S_1:\Gamma_{h,l}=0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H}) \prod_{h\in S_1:\Gamma_{h,l}=1} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,1}) \cdot \bar{p}_{\sigma(l)}, \tag{D.19}$$

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

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

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

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

So (D.18) = (D.19) gives

$$\prod_{h\in S_1:\Gamma_{h,l}=0} (\theta_{h,l} - \theta_{h,H}) \prod_{h\in S_1:\Gamma_{h,l}=1} (\theta_{h,l} - \theta_{h,1}) \cdot p_l = \prod_{h\in S_1:\Gamma_{h,l}=0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H}) \prod_{h\in S_1:\Gamma_{h,l}=1} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,1}) \cdot \bar{p}_{\sigma(l)}, \tag{D.20}$$

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

yields

$$\begin{aligned}
&\theta_{j,l} \cdot \prod_{h\in S_1:\Gamma_{h,l}=0} (\theta_{h,l} - \theta_{h,H}) \prod_{h\in S_1:\Gamma_{h,l}=1} (\theta_{h,\boldsymbol{\alpha}} - \theta_{h,1}) \cdot p_l \qquad \text{(D.21)}\\
&= \bar{\theta}_{j,\sigma(l)} \cdot \prod_{h\in S_1:\Gamma_{h,l}=0} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,H}) \prod_{h\in S_1:\Gamma_{h,l}=1} (\bar{\theta}_{h,\sigma(l)} - \theta_{h,1}) \cdot \bar{p}_{\sigma(l)}.
\end{aligned}$$

Taking the ratio of the above two equations (D.21) and (D.20) gives $\theta_{j,l} = \bar{\theta}_{j,\sigma(l)}, \quad \forall j \notin S_1$. Redefining $\boldsymbol{r}^* := \sum_{h\in S_2} \boldsymbol{e}_h$ similarly as above we have $\theta_{j,l} = \bar{\theta}_{j,\sigma(l)}$ for any $j \in S_1$. Plug $\theta_{j,l} = \bar{\theta}_{j,\sigma(l)}$ for all $j \in S_1$ into (D.20), then we have $p_l = \bar{p}_{\sigma(l)}$. Now we have shown (D.12) hold for this particular l. Then the induction argument gives

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

Now we have shown for any item j and latent class index l, $\theta_{j,l} = \bar{\theta}_{j,\sigma(l)}$, which we denote by $\bar{\boldsymbol{\Theta}} = \sigma(\boldsymbol{\Theta})$. We claim that this result also indicates that the permutation σ is unique. This is because $U_1 \bullet \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}} = I_L$ implies that

$$U_1 \bullet \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}} = U_1 \bullet \{W_1 T(\Gamma^1, \boldsymbol{\Theta}^1)\}_{\boldsymbol{r}_{1:L_0}} \bullet \sigma(I_L) = \sigma(I_L),$$

which means given U_1 constructed from $(\Gamma, \boldsymbol{\Theta})$, the form of $U_1 \bullet \{W_1 T(\bar{\Gamma}^1, \bar{\boldsymbol{\Theta}}^1)\}_{\boldsymbol{r}_{1:L_0}}$ explicitly and uniquely determines σ. Now we have shown $\bar{\Gamma} = \Gamma = \Gamma^{\mathcal{A}_0}$ and $(\bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{p}}) = (\boldsymbol{\Theta}, \boldsymbol{p})$ must hold up to the column permutation σ.

As stated in the beginning of the proof, combining Condition C that any column in $\Gamma^{\mathcal{A}_0}$ is different from any column in $\Gamma^{\mathcal{A}_0^c}$, the identification of $\Gamma^{\mathcal{A}_0}$ uniquely identifies the set of true patterns $\mathcal{A}_0$. The proof of both Theorem V.1 and Corollary V.1 is complete. $\square$

D.2 Proof of Theorem V.2 and Corollary V.2.

The following proofs of Theorem V.2 and Corollary V.2 use a similar proof idea as that of Allman et al. (2009); see also proofs of Theorems 4.2 and 4.3 in Gu and Xu (2020a).

Proof of Theorem V.2. We need to introduce the definition of *algebraic variety*, a concept in algebraic geometry. An algebraic variety $\mathcal{V}$ is defined as the simulateneous zero-set of a finite collection of multivariate polynomials $\{f_i\}_{i=1}^n \subseteq \mathbb{R}[x_1, x_2, \dots, x_d]$, $\mathcal{V} = \mathcal{V}(f_1, \dots, f_n) = \{\boldsymbol{x} \in \mathbb{R}^d \mid f_i(\boldsymbol{x}) = 0,\ 1 \leq i \leq n.\}$ An algebraic variety $\mathcal{V}$ is all of $\mathbb{R}^d$ only when all the polynomials defining it are zero polynomials; otherwise, $\mathcal{V}$ is called a *proper subvariety* and is of dimension less than d, hence necessarily of Lebesgue measure zero in $\mathbb{R}^d$. The same argument holds when $\mathbb{R}^d$ is replaced by the parameter space $\Omega \subseteq \mathbb{R}^d$ that has full dimension in $\mathbb{R}^d$. For the structured latent attribute model, we consider the following parameter space,

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

On Ω, altering some entries of zero to one in the Γ-matrix is equivalent to impose more affine constraints on the parameters and force them to be in a subset Ω^* of Ω. Condition $A^\star$ guarantees that, there exists a Ω^* such that Condition A holds for model parameters belonging to this Ω^*, the proof of Theorem V.1 gives that the matrix $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ has full column rank C for $i = 1, 2$ for $(\boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)}, \boldsymbol{p}^{\mathcal{A}_0}) \in \Omega^*$. Note that the statement that $2^{|S_i|} \times C$ matrix $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ has full column rank is equivalent to the statement that the map sending $T(\Gamma^{(S_i,\mathcal{A}_0)}, \boldsymbol{\Theta}^{(S_i,\mathcal{A}_0)})$ to all its $\binom{2^{|S_i|}}{C}$ possible $C \times C$ minors $A_1^i, A_2^i, \dots, A_{2^{|S_i|}}^i$ yields at least one nonzero minor, where

$A_1^i, A_2^i, \dots, A_{2^{|S_i|}}^i$ are all polynomials of the item parameters $\mathbf{\Theta}_{S_i}$. Define

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

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

Proof of Corollary V.2. Consider a Q-matrix in the form of (5.5). We denote $S_1 = \{1, \dots, K\}$, $S_2 = \{K+1, \dots, 2K\}$ and $S_3 = \{2K+1, \dots, J\}$, which are item sets corresponding to Q_1, Q_2 and Q', respectively. According to the proof of Theorem 4.3 in Gu and Xu (2020a), since the two submatrices Q_1 and Q_2 have all the diagonal elements equal to one, the $2^K \times 2^K$ T-matrices $T(\Gamma^{(S_1,\text{all})}, \mathbf{\Theta}^{(S_1,\text{all})})$ and $T(\Gamma^{(S_2,\text{all})}, \mathbf{\Theta}^{(S_2,\text{all})})$ are generically full-rank. Furthermore, the matrix $T(\Gamma^{(S_3,\text{all})}, \mathbf{\Theta}^{(S_3,\text{all})}) \cdot \text{Diag}(\boldsymbol{p}^{\text{all}})$ has Kruskal rank at least two. This means generically, any two columns of $T(\Gamma^{(S_3,\text{all})}, \mathbf{\Theta}^{(S_3,\text{all})}) \cdot \text{Diag}(\boldsymbol{p}^{\text{all}})$ are linearly independent.

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

D.3 Proof of Corollary V.3.

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

D.4 Proof of Theorem V.3 and Proposition V.1.

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

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

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

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{\sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}}}\widehat{p}_{\boldsymbol{\alpha}}\prod_j\widehat{\theta}_{j,\boldsymbol{\alpha}}^{R_{i,j}}(1-\widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}}}{\sum_{\boldsymbol{\alpha}\in\mathcal{A}_0}\widehat{p}^0_{\boldsymbol{\alpha}}\prod_j(\widehat{\theta}^0_{j,\boldsymbol{\alpha}})^{R_{i,j}}(1-\widehat{\theta}^0_{j,\boldsymbol{\alpha}})^{1-R_{i,j}}}\right]\\ >&\frac{|\lambda_N|}{N}\Big\{\log_{\rho_N}(\widehat{\boldsymbol{p}}_{\text{full}}^{\widehat{\mathcal{A}}}) - \log_{\rho_N}(\widehat{\boldsymbol{p}}_{\text{full}}^{\mathcal{A}_0})\Big\}.\end{aligned} \tag{D.22}$$

In the case of $|\widehat{\mathcal{A}}| > |\mathcal{A}_0|$ (which we call the overfitted case), the right hand side (RHS) of (D.22) regarding the difference between the penalty terms has order $O(N^{-1}|\lambda_N|\cdot|\mathcal{A}_0|\cdot|\log\rho_N|)$. In this overfitted case, we now consider the left hand side (LHS) of (D.22),

$$\begin{aligned}\text{LHS of (D.22)} =& \frac{1}{N}\sum_{i=1}^{N}\log\left[\sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}}}\widehat{p}_{\boldsymbol{\alpha}}\prod_j\widehat{\theta}_{j,\boldsymbol{\alpha}}^{R_{i,j}}(1-\widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}}\right]\\ &-\frac{1}{N}\sum_{i=1}^{N}\log\left[\sum_{\boldsymbol{\alpha}\in\mathcal{A}_0}\widehat{p}^0_{\boldsymbol{\alpha}}\prod_j(\widehat{\theta}^0_{j,\boldsymbol{\alpha}})^{R_{i,j}}(1-\widehat{\theta}^0_{j,\boldsymbol{\alpha}})^{1-R_{i,j}}\right] \equiv I_1 - I_0,\end{aligned}$$

where the I_1 part can be written as

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

$$= \frac{1}{N}\sum_{i=1}^{N}\log\Big[\sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}_{\text{input}},\\ \widehat{p}_{\boldsymbol{\alpha}}>\rho_N}}\widehat{p}_{\boldsymbol{\alpha}}\prod_j \widehat{\theta}_{j,\boldsymbol{\alpha}}^{R_{i,j}}(1-\widehat{\theta}_{j,\boldsymbol{\alpha}})^{1-R_{i,j}}\Big] + O(N^{-\delta}), \tag{D.23}$$

where the last equality follows from the assumption $|\mathcal{A}_{\text{input}}|\cdot\rho_N = O(N^{-\delta})$ in the theorem. So we further have the LHS of (D.22) equal to

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

Note that other than the last term $O(N^{-\delta})$ in the above display, the difference of the first two terms also has order $O_p(N^{-\delta})$ from assumption (5.11), so LHS of (D.22)= $I_1 - I_0 = O_p(N^{-\delta})$. In order to have selection consistency in the overfitted case, we need the event described in (D.22) to happen with probability tending to zero, so the $|\lambda_N|$ needs to be sufficiently large such that

$$N^{-\delta} \lesssim O\Big(N^{-1}|\lambda_N|\cdot|\log\rho_N|\Big). \tag{D.24}$$

Note that by (5.10), we have $\rho_N \asymp N^{-d}$ for some $d>0$. So if $\delta<1$, i.e., if the convergence rate is slower than the $\sqrt{N}$ rate, then λ_N must go to negative infinity as N goes to infinity since $\delta<1$. Specifically, we obtain the following lower bound of the magnitude of the penalty parameter λ_N,

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

would suffice for (D.24) to hold.

We now prove the conclusion of Proposition V.1. A further implication of the above discussion is that, with $\rho_N \asymp N^{-d}$ as assumed in (5.10), just imposing a proper Dirichlet prior with a positive hyperparameter would fail to select the true model consistently. In particular, with a proper Dirichlet prior density with hyperparameter $\beta = \lambda_N + 1 \in (0, 1)$, Equation (D.24) instead becomes $N^{1-\delta} = o(\log N)$. However, when $0 < \delta < 1$, $N^{1-\delta}/\log N \to \infty$. So (D.24) fails to hold, and one can not have consistent selection in the overfitted case. So if we denote the set of attribute patterns estimated by maximizing (5.9) by $\widehat{\mathcal{A}}^{\lambda}$. Then for any $\{\lambda_N\} \subseteq [-1, 0)$, $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda} = \mathcal{A}_0) \not\to 1$ as $N \to \infty$. This proves Proposition V.1.

Now we consider the random set $\{\boldsymbol{\alpha} \in \mathcal{A}_{\text{input}} : \widehat{p}_{\boldsymbol{\alpha}} > \rho_N\} =: \widehat{\mathcal{A}}$ appearing in I_1 in (D.23). With probability tending to one, the cardinality of this set is smaller than $|\mathcal{A}_0|$. This is because if $|\widehat{\mathcal{A}}| > |\mathcal{A}_0|$, the log-penalty term corresponding to $\widehat{\mathcal{A}}$ would be smaller than that corresponding to $\mathcal{A}_0$ by $N^{-1}|\lambda_N| \cdot |\log \rho_N|$ which has order at least $N^{-\delta}$. Recall that the right hand side of (D.22) has order $O_P(N^{-\delta})$, which means when $|\widehat{\mathcal{A}}| > |\mathcal{A}_0|$ the extent that the log-penalty part favors the a smaller model $\mathcal{A}_0$ would dominate the extent that the likelihood part favors a larger model $\widehat{\mathcal{A}}$ in the proposed penalized likelihood. Therefore any larger model $\widehat{\mathcal{A}}$ with $|\widehat{\mathcal{A}}| \geq |\mathcal{A}_0|$ would be favored over $\mathcal{A}_0$ with probability tending to zero. Therefore we have the conclusion that $\mathbb{P}(\widehat{\mathcal{A}} \neq \mathcal{A}_0) \not\to 0$ could only happen for $|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|$. So in the following discussion we will focus on the case where $|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|$ and prove consistency in this case. Namely, we aim to bound

$$\mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|,\, \widehat{\mathcal{A}} \neq \mathcal{A}_0} [\ell^{\lambda_N}(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell^{\lambda_N}(\boldsymbol{\eta}^{\mathcal{A}_0})] > 0 \Big). \tag{D.25}$$

Next, we consider the upper bound of the magnitude of the penalty term. In order to have selection consistency in the case of $|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|$ and $\widehat{\mathcal{A}} \neq \mathcal{A}_0$, the log-penalty term can not be too large such that the extent that the penalty part favors a smaller model does not dominate the extent that the likelihood part favors the true model. We follow a similar argument to Shen et al. (2012a). Specifically, considering the term $-\epsilon_N^2 \to 0$ in the large deviation inequality (D.28) below; for a small constant $t > \epsilon_N$, we need that the difference of the penalty part of the true and any alternative smaller model to be less than t^2, i.e.,

$$|\lambda_N| \cdot |\log \rho_N|/N \lesssim t^2, \tag{D.26}$$

Equation (D.26) would hold if

$$|\lambda_N| = o(N/|\log \rho_N|). \tag{D.27}$$

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

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

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

Consider the local parameter space $\mathcal{B}_{\mathcal{A}} = \{\boldsymbol{\eta}^{\mathcal{A}} = (\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) : |\mathcal{A}| \leq |\mathcal{A}_0|, h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0}) \leq 2\epsilon_N^2\}$, the $H(t, \mathcal{B}_{\mathcal{A}})$ is defined as the logarithm of the cardinality of the t-bracketing of $\mathcal{B}_{\mathcal{A}}$ of the smallest size. More specifically, following the definition in Shen et al. (2012a), consider a bracket covering $S(t,m) = \{f_1^l, f_1^u, \dots, f_m^l, f_m^u\}$ satisfying that $\max_{1\leq j\leq m} \left\| f_j^u - f_j^l \right\|_2 \leq t$ and for any $f \in \mathcal{B}_{\mathcal{A}}$ there is some j such that $f_j^l \leq f \leq f_j^u$

almost surely. Then $H(t, \mathcal{B}_{\mathcal{A}})$ is $\log(\min\{m : S(t,m)\})$. The $H(t, \mathcal{B}_{\mathcal{A}})$ measures the complexity of the local parameter space. The next lemma gives an upper bound for the bracketing Hellinger metric entropy $H(t, \mathcal{B}_{\mathcal{A}})$ for $|\mathcal{A}| \leq |\mathcal{A}_0|$.

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

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

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

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

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

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

$$
\begin{aligned}
&\mathbb{P}\Big(\sup_{h^2(\widehat{\boldsymbol\eta}^{\widehat{\mathcal{A}}},\boldsymbol\eta^{\mathcal{A}_0})\geq\epsilon_N^2} \Big[\frac{1}{N}\ell(\widehat{\boldsymbol\eta}^{\widehat{\mathcal{A}}}) - \frac{1}{N}\ell(\widehat{\boldsymbol\eta}^{\mathcal{A}_0})\Big] > -\epsilon_N^2\Big) \\
&\leq \mathbb{P}\Big(\sup_{h^2(\widehat{\boldsymbol\eta}^{\widehat{\mathcal{A}}},\boldsymbol\eta^{\mathcal{A}_0})\geq\epsilon_N^2} \Big[\frac{1}{N}\ell(\widehat{\boldsymbol\eta}^{\widehat{\mathcal{A}}}) - \frac{1}{N}\ell(\boldsymbol\eta^{\mathcal{A}_0})\Big] > -\epsilon_N^2\Big) \leq \exp(-N\epsilon_N^2).
\end{aligned} \tag{D.28}
$$

where $\boldsymbol\eta^{\mathcal{A}_0} = (\boldsymbol\Theta^{\mathcal{A}_0}, \boldsymbol p^{\mathcal{A}_0})$ denote the true parameters. Indeed, Theorem 1 in Wong and Shen (1995) guarantees the inequality (D.28) holds with ϵ_N replaced by any $t > \epsilon_N = \sqrt{|\mathcal{A}_0| \log |\mathcal{A}_{\text{input}}|/N}$. This large deviation inequality will be used later to bound the mis-selection probability in the case of $|\mathcal{A}| \leq |\mathcal{A}_0|$.

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

$$
\begin{aligned}
&\frac{h^2(\boldsymbol\eta^{\mathcal{A}}, \boldsymbol\eta^{\mathcal{A}_0})}{\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)} \\
&\asymp [\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)]^{-1} \sum_{\boldsymbol r\in\{0,1\}^J} \Big[\Big(\sum_{\boldsymbol\alpha\in\mathcal{A}} \mathbb{P}(\boldsymbol R = \boldsymbol r \mid \boldsymbol\Theta^{\mathcal{A}}, \mathcal{A} = \boldsymbol\alpha) p_{\boldsymbol\alpha}^{\mathcal{A}}\Big)^{1/2} - \\
&\qquad\qquad \Big(\sum_{\boldsymbol\alpha\in\mathcal{A}_0} \mathbb{P}(\boldsymbol R = \boldsymbol r \mid \boldsymbol\Theta^{\mathcal{A}_0}, \mathcal{A} = \boldsymbol\alpha) p_{\boldsymbol\alpha}^{\mathcal{A}_0}\Big)^{1/2}\Big]^2 \\
&\asymp [\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)]^{-1} \sum_{\boldsymbol r\in\{0,1\}^J} \Big(\sum_{\boldsymbol\alpha\in\mathcal{A}} T_{\boldsymbol r,\boldsymbol\alpha}(\boldsymbol\Theta^{\mathcal{A}})\, p_{\boldsymbol\alpha}^{\mathcal{A}} - \sum_{\boldsymbol\alpha\in\mathcal{A}_0} T_{\boldsymbol r,\boldsymbol\alpha}(\boldsymbol\Theta^{\mathcal{A}_0})\, p_{\boldsymbol\alpha}^{\mathcal{A}_0}\Big)^2 \\
&= [\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)]^{-1} \big\|T(\Gamma^{\mathcal{A}}, \boldsymbol\Theta^{\mathcal{A}})\boldsymbol p^{\mathcal{A}} - T(\Gamma^{\mathcal{A}_0}, \boldsymbol\Theta^{\mathcal{A}_0})\boldsymbol p^{\mathcal{A}_0}\big\|_2^2
\end{aligned}
$$

To proceed with the proof, we need to use Theorem V.1 to establish an identifiability argument. Theorem V.1 and Corollary V.1 state that if the true constraint matrix $\Gamma^{\mathcal{A}_0}$ satisfies conditions A, B and C, then $(\Gamma^{\mathcal{A}_0},\ \boldsymbol\Theta^{\mathcal{A}_0},\ \boldsymbol p^{\mathcal{A}_0})$ are jointly identifiable. This implies that given the set of true attribute patterns $\mathcal{A}_0$, for any other set $\mathcal{A} \neq \mathcal{A}_0$, $|\mathcal{A}| \leq |\mathcal{A}_0|$, and model parameters defined by $\mathcal{A}$ must lead to different $T(\boldsymbol\Theta^{\mathcal{A}})\boldsymbol p^{\mathcal{A}}$ that is different from $T(\boldsymbol\Theta^{\mathcal{A}_0})\boldsymbol p^{\mathcal{A}_0}$. Moreover, consider the parameter space $\mathcal{B} =$

$\{(\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) : \ |\mathcal{A}| \leq |\mathcal{A}_0|, \ p_{\boldsymbol{\alpha}} > \rho_N \ \forall \boldsymbol{\alpha} \in \mathcal{A}\}$. Then $(\boldsymbol{\Theta}^{\mathcal{A}_0}, \boldsymbol{p}^{\mathcal{A}_0}) \in \mathcal{B}$ and for any $(\boldsymbol{\Theta}^{\mathcal{A}}, \boldsymbol{p}^{\mathcal{A}}) \in \mathcal{B}$ with $\mathcal{A} \neq \mathcal{A}_0$, **either** some elements in $\boldsymbol{\Theta}^{\mathcal{A}}$ differs from those in $\boldsymbol{\Theta}^{\mathcal{A}_0}$ by a nonzero constant, **or** some elements in $\boldsymbol{p}^{\mathcal{A}}$ differs from those in $\boldsymbol{p}^{\mathcal{A}_0}$ by a nonzero constant. Since $T^{\mathcal{A}}(\boldsymbol{\Theta})\boldsymbol{p}^{\mathcal{A}}$ is a continuous vector-valued function of the model parameters, we must have $[\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)]^{-1} \left\| T^{\mathcal{A}}(\boldsymbol{\Theta})\boldsymbol{p}^{\mathcal{A}} - T^{\mathcal{A}_0}(\boldsymbol{\Theta})\boldsymbol{p}^{\mathcal{A}_0} \right\|_2^2 \geq C_0$ for some $C_0 > 0$. By the conditions of the theorem $\epsilon_N^2 = o(1)$, so we have obtained for some small constant $t > \epsilon_N$,

$$C_{\min}(\boldsymbol{\eta}^0) \equiv \inf_{\boldsymbol{\eta}^{\mathcal{A}}:\, \mathcal{A} \neq \mathcal{A}_0, |\mathcal{A}| \leq |\mathcal{A}_0|} \left\{ \frac{h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0})}{\max(|\mathcal{A}_0 \setminus \mathcal{A}|, 1)} \right\} \geq C_0 \gtrsim t^2 > \epsilon_N^2. \tag{D.29}$$

Finally, with the λ_N of the previously specified order, we use the large deviation inequality (D.28) and also the (D.29) to bound the false selection probability (D.25). The following argument uses a similar proof idea as that of Theorem 1 in Shen et al. (2012a) which establishes finite sample mis-selection error bound of the L_0-constrained maximum likelihood estimation. Consider $|\widehat{\mathcal{A}} \cap \mathcal{A}_0| = m \leq |\mathcal{A}_0| - 1$, by (D.29) we have $h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0}) \geq (|\mathcal{A}_0| - m) C_{\min}(\boldsymbol{\eta}^0)$. So

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}}| \leq |\mathcal{A}_0|, \widehat{\mathcal{A}} \neq \mathcal{A}_0} \Big[\frac{1}{N} \ell^{\lambda_N}(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \frac{1}{N} \ell^{\lambda_N}(\boldsymbol{\eta}^0) \Big] > 0 \Big) \\
&\leq \sum_{m=0}^{|\mathcal{A}_0|-1} \sum_{j=1}^{|\mathcal{A}_0|-m} \mathbb{P}\Big(\sup_{\substack{h^2(\boldsymbol{\eta}^{\mathcal{A}}, \boldsymbol{\eta}^{\mathcal{A}_0}) \geq \\ (|\mathcal{A}_0|-m) C_{\min}(\boldsymbol{\eta}^0)}} \frac{1}{N} \Big[\ell(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell(\boldsymbol{\eta}^0) \Big] > - \frac{|\lambda_N| \cdot |\mathcal{A}_0| \cdot |\log \rho_N|}{N} \Big) \\
&\leq \sum_{m=0}^{|\mathcal{A}_0|-1} \sum_{j=1}^{|\mathcal{A}_0|-m} \mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}} \cap \mathcal{A}_0| = m} \frac{1}{N} \Big[\ell(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell(\boldsymbol{\eta}^0) \Big] > -t^2 \Big) \quad \text{(by (D.27))} \\
&\leq \sum_{m=0}^{|\mathcal{A}_0|-1} \sum_{j=1}^{|\mathcal{A}_0|-m} \mathbb{P}\Big(\sup_{|\widehat{\mathcal{A}} \cap \mathcal{A}_0| = m} \frac{1}{N} \Big[\ell(\boldsymbol{\eta}^{\widehat{\mathcal{A}}}) - \ell(\boldsymbol{\eta}^0) \Big] > -(|\mathcal{A}_0| - m) C_{\min}(\boldsymbol{\eta}^0) \Big) \quad \text{(by (D.29))} \\
&\leq \sum_{m=0}^{|\mathcal{A}_0|-1} \binom{|\mathcal{A}_0|}{m} \exp\Big(-c_2 N (|\mathcal{A}_0| - m) C_{\min}(\boldsymbol{\eta}^0) \Big) \sum_{j=1}^{|\mathcal{A}_0|-m} \binom{|\mathcal{A}_{\text{input}}| - |\mathcal{A}_0|}{j} \quad \text{(by (D.28))} \\
&\leq c_3 \exp\Big(-c_2 N C_{\min}(\boldsymbol{\eta}^0) + 2 \log(|\mathcal{A}_{\text{input}}| + 1) \Big),
\end{aligned}$$

where the last but one line above uses the large deviation inequality in (D.28), and c_2, c_3 are some constants. And the last line follows from the calculations in the proof of Theorem 1 in Shen et al. (2012a) using some basic inequalities about binomial coefficients. Since $C_{\min}(\boldsymbol{\eta}^0) \geq C_0$, and $\log|\mathcal{A}_{\text{input}}| = o(N)$ by the assumption of the theorem, the right hand side of the above display goes to zero as $N \to \infty$. Therefore $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} \neq \mathcal{A}_0,\ |\widehat{\mathcal{A}}^{\lambda_N}| \leq |\mathcal{A}_0|) \to 0$ as $N \to \infty$. Combined with the previously shown result $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} \neq \mathcal{A}_0) \not\to 0$ could potentially happen only for $|\widehat{\mathcal{A}}^{\lambda_N}| \leq |\mathcal{A}_0|$, we have the conclusion $\mathbb{P}(\widehat{\mathcal{A}}^{\lambda_N} \neq \mathcal{A}_0) \to 0$ as $N \to \infty$. The proof of the theorem is complete. $\square$

D.5 Proof of Theorem V.4.

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

$$\mathbb{P}(\boldsymbol{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathcal{A} = \boldsymbol{\alpha} \mid \boldsymbol{\Theta}) > \mathbb{P}(\boldsymbol{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathcal{A} = \widetilde{\boldsymbol{\alpha}} \mid \boldsymbol{\Theta}), \quad \forall \widetilde{\boldsymbol{\alpha}} \neq \boldsymbol{\alpha}. \tag{D.30}$$

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

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

Therefore

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

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

$$\begin{aligned}&\log\mathbb{P}(\boldsymbol{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathcal{A} = \boldsymbol{\alpha} \mid \boldsymbol{\Theta}) - \log\mathbb{P}(\boldsymbol{R} = \boldsymbol{r}^{\boldsymbol{\alpha}},\ \mathcal{A} = \widetilde{\boldsymbol{\alpha}} \mid \boldsymbol{\Theta}) \\ \geq& \min_{j=1,\dots,J}\{\log\theta_j^+ - \log\theta_{j,\widetilde{\boldsymbol{\alpha}}}^-,\ \log(1-\theta_{j,\boldsymbol{\alpha}}^+) - \log(1-\theta_j^+)\} \geq d > 0.\end{aligned} \tag{D.31}$$

That the above probability is bounded away from zero follows from the second part of assumption (5.12). So the claim (D.30) is proved. We next bound the probability of failure of including all the true patterns in the screening stage. First, since $\mathcal{A}_1,\dots,\mathcal{A}_N \overset{i.i.d.}{\sim}$ Multinomial$(N,\ (p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}_0))$, then $|\{i \in [N] : \mathcal{A}_i = \boldsymbol{\alpha}\}|$ denotes the number of subjects in the random sample whose attribute pattern is $\boldsymbol{\alpha}$. By the concentration inequality of the multinomial distribution, for any $\boldsymbol{\alpha} \in \mathcal{A}_0$,

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

Because of (5.12), we have $Np_{\boldsymbol{\alpha}} \geq Nc_0 \to \infty$ for all $\boldsymbol{\alpha} \in \mathcal{A}_0$. Assume that $\widehat{\theta}_j^+ - \widehat{\theta}_j^- > \delta > 0$ for each $j \in [J]$. This constraint can be incorporated into the screening procedure or checked *a posteiriori* after screening. So with probability at least $1-$

$2^{|\mathcal{A}_0|}\exp(-2t^2)$ for a suitable t,

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

as $N \to \infty$. Here $\widehat{\mathbb{P}}$ refers to the probability measure of $\boldsymbol{R}$ and $\mathcal{A}$ given the estimated item parameters $\widehat{\boldsymbol{\theta}}^{+}$ and $\widehat{\boldsymbol{\theta}}^{-}$. This is because the probability inside the bracket in the above expression is strictly less than 1 due to (D.31); we denote this quantity by C_δ since it depends on δ. Therefore there is

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

where $\beta_{\min}$ is a positive constant which can be taken as $c_0/2\log(1/C_\delta)$. The last inequality above results from $p_{\boldsymbol{\alpha}} \geq c_0$ for $\boldsymbol{\alpha} \in \mathcal{A}_0$ in (5.12) and that $C_\delta < 1$. Now we have obtained $\mathbb{P}(\widehat{\mathcal{A}}_{\text{screen}} \supseteq \mathcal{A}_0) \geq 1 - |\mathcal{A}_0|\exp(-N\beta_{\min})$, so the sure screening property holds and the proof is complete. $\square$

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

possibilities of choosing $|\mathcal{A}|$ components from $|\mathcal{A}_{\text{input}}|$ ones, we have

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

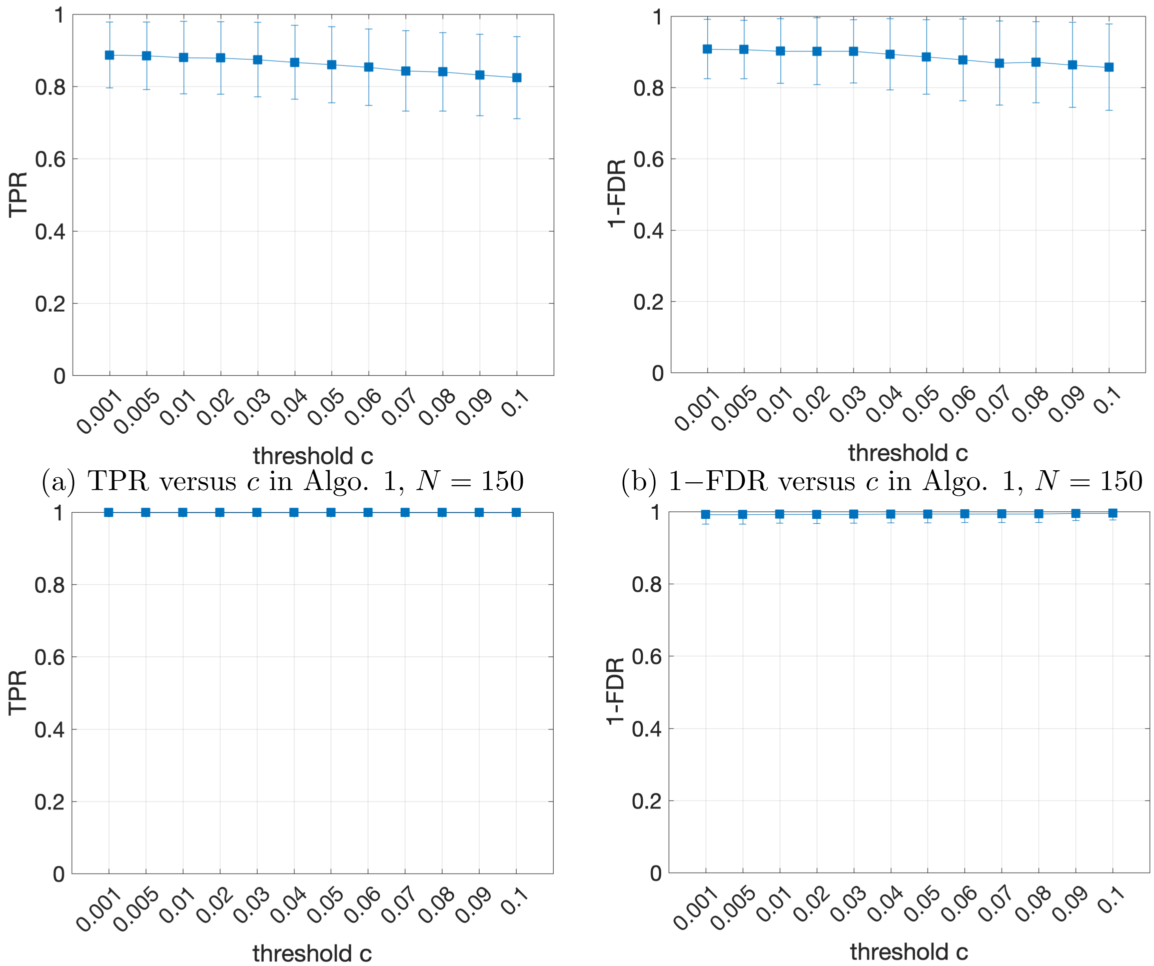
Next, Lemma 2 in Genovese and Wasserman (2000) gives the following bracketing entropy bound for the simplex, $N_{[]}(t, \mathcal{T}^{|\mathcal{A}|-1}) \leq |\mathcal{A}|(2\pi e)^{|\mathcal{A}|/2}/t^{|\mathcal{A}|-1}$. Since we consider the local parameter space around the true parameters (with squared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

$$\begin{aligned} H(t, \mathcal{B}_{\mathcal{A}}) &\leq \log\left\{\binom{|\mathcal{A}_{\text{input}}|}{|\mathcal{A}|} \frac{|\mathcal{A}|(2\pi e)^{|\mathcal{A}|/2}(\epsilon)^{|\mathcal{A}|-1}}{t^{|\mathcal{A}|-1}} \left(\frac{\epsilon}{t}\right)^{|\mathcal{A}|}\right\} \\ &\lesssim |\mathcal{A}| \log |\mathcal{A}_{\text{input}}| + \log |\mathcal{A}| + |\mathcal{A}| \log(\epsilon/t) \\ &\lesssim |\mathcal{A}_0| \log |\mathcal{A}_{\text{input}}| \log(\epsilon/t). \end{aligned}$$

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

D.6 Additional Experimental Results for Chapter V

Impact of the value of the pre-specified c in Algorithm 1. In Algorithm 1, there is a pre-specified constant $c > 0$ when updating the Δ_l's. This constant c should be small, ideally close to zero. In all of our experiments in Section 6.4, we take $c = 0.01$. Next we examine how the value of c impacts the selection result of Algorithm 1. Since the performance of Algorithm 1 is the focus here, we choose the simulation setting with $K = 10$ such that screening can be omitted. Under sample sizes $N = 150$ and $N = 500$, the plots of the two accuracy measures versus c are presented in Figure D.1. We observe that the results of Algorithm 1 are generally not that sensitive to the choice of c, though smaller c gives slightly better results for both accuracy measures under a small sample size $N = 150$. For N as large as 500, for all the values of $c \in \{0.001, 0.005\} \cup \{0.01 \times i : i = 1, 2, \ldots, 10\}$, the two accuracy measures are very close to one and do not have much variation. In practice, we recommend fixing c to a value no greater than 0.01.



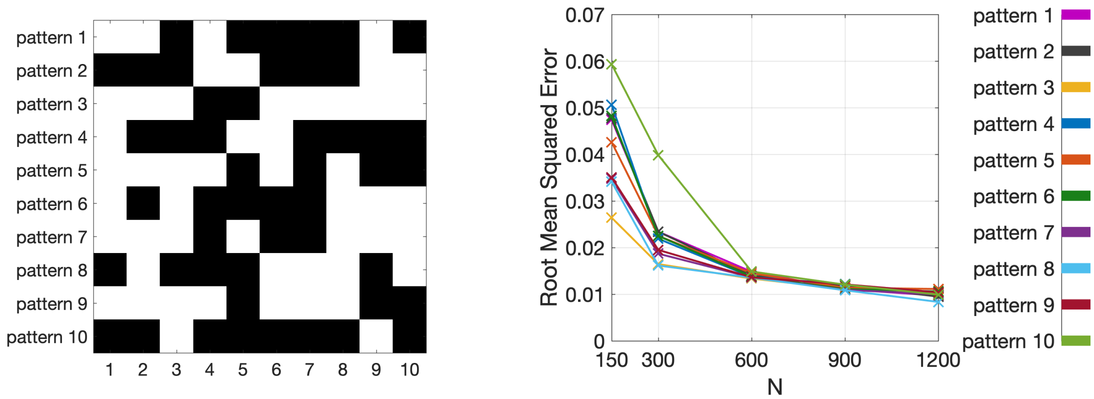
(a) TPR versus c in Algo. 1, $N = 150$ (b) 1−FDR versus c in Algo. 1, $N = 150$

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

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

Algorithm 1's performance on estimating the actual proportions of patterns. Other than the two accuracy measures for pattern selection presented in Table 5.2, we also evaluate how well the algorithms perform on estimating the actual proportions of the latent patterns. We use the simulation setting of the two-parameter SLAM with $K = 10$, $|\mathcal{A}_0| = 10$, $Q = (Q_1^\top, Q_2^\top, Q_3^\top)^\top$, with parameters $p_{\boldsymbol{\alpha}} = 0.1$ for $\boldsymbol{\alpha} \in \mathcal{A}$ and $1 - \theta_j^+ = \theta_j^- = 0.2$. This is the same setting as that of Example V.4. We vary the sample size $N \in \{150, 300, 600, 900, 1200\}$ and compute the Root Mean Square Errors (RMSEs) of estimating the true proportions of latent patterns. The randomly generated 10 true patterns in $\mathcal{A}_0$ are presented in Figure D.2(a), where each row represents a K-dimensional binary pattern. For each N, the RMSE of each proportion $p_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathcal{A}_0$ is computed based on 200 runs; and in each run, we first perform pattern selection by using EBIC to choose $\lambda \in \{-0.2 \times i : i = 1, 2, \ldots, 20\}$ in Algorithm 1 and then estimate the proportions based on the selected set of patterns. The results of RMSEs are presented in Figure D.2(b). As can be seen from the figure, under a small sample size $N = 150$, the RMSEs of patterns are relatively diverse. In particular, the largest RMSE is around 0.06 and corresponds to pattern 10, $\boldsymbol{\alpha}_{10} = (0010000010)$, which is the pattern consisting of most "0"s; while the smallest RMSE is less than half of the largest and corresponds to pattern 3, $\boldsymbol{\alpha}_3 = (1110011111)$, which is the pattern consisting of most "1"s. Interestingly, this observation implies for a very small sample size and a sparse Q-matrix (each row having at most three entries of "1"s), those attribute patterns possessing fewer attributes are harder to estimate while those possessing more attributes are easier to estimate. While as N increases, the RMSEs of all the proportions decrease and their difference become not discernible. For $N = 1200$, all the RMSEs are around 0.01.

Evaluating the screening procedure under the multi-parameter SLAM. In the multi-parameter setting, we also evaluate the performance of the approximate screening procedure that is developed based on the likelihood of the two-parameter model. The results of the coverage probabilities are presented in Figure D.3. The figure shows that despite being an approximate procedure, the screening Algorithm 4 has

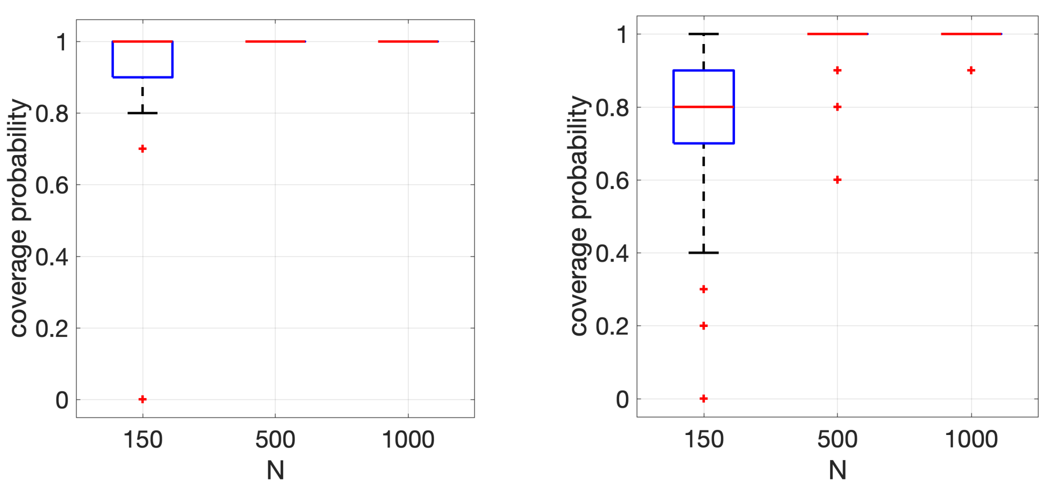


(a) patterns in $\mathcal{A}_0$: white for 1, black for 0

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

Figure D.2: Root Mean Square Errors (RMSEs) for estimating the true proportions of patterns decrease as sample size N increases. Results are based on 200 runs for each N.

excellent performance for the multi-parameter SLAM that covers the two-parameter model as a submodel. Specifically, Figure D.3 shows that for both $K = 15$ and $K = 20$, the approximate screening procedure almost always has a 100% coverage probability for $N = 500$ and $N = 1000$.

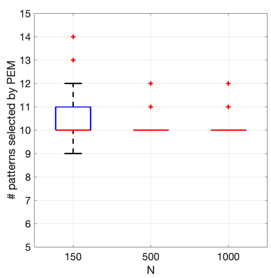


(a) $K = 15$, multi-parameter SLAM

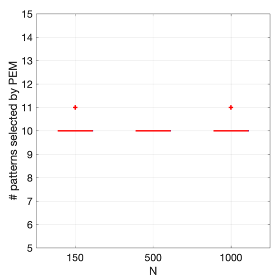
(b) $K = 20$, multi-parameter SLAM

Figure D.3: Coverage probabilities of the true patterns, from the approximate screening procedure under the multi-parameter SLAM. Boxplots are from 200 runs in each scenario.

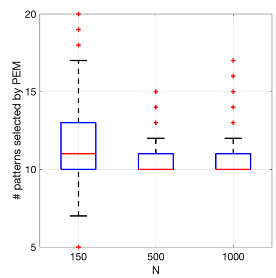
Sizes of the set of finally selected patterns under scenarios in Table 5.2. We present the results of the number of patterns that are finally selected by the proposed methods, corresponding to simulation scenarios in Table 5.2. Denote the set of patterns selected by the PEM algorithm and that selected by the FP-VEM algorithm by $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$, respectively. As shown in Figure D.4, in the relatively strong signal setting with $1-\theta_j^+=\theta_j^-=10\%$, the sizes of $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$ almost always equal 10, the number of true patterns. Combined with the accuracy measures presented in Table 5.2 in the main text, in most cases these selected 10 patterns are indeed exactly the true ones in $\mathcal{A}_0$. And in the relatively weak signal setting with $1-\theta_j^+=\theta_j^-=20\%$, the sizes of $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$ can be slightly larger than $|\mathcal{A}_0|$ but still close to it.



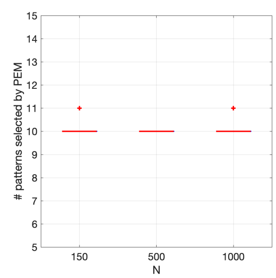
(a) $K=15$, noise 20%; PEM



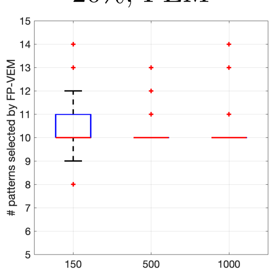
(b) $K=15$, noise 10%; PEM



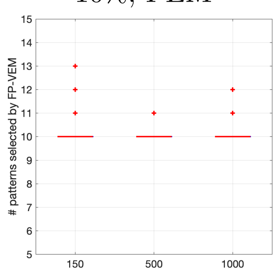
(c) $K=20$, noise 20%; PEM



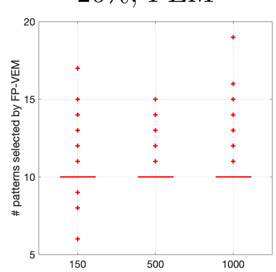
(d) $K=20$, noise 10%; PEM



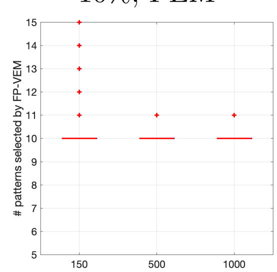
(e) $K=15$, noise 20%; FP-VEM



(f) $K=15$, noise 10%; FP-VEM



(g) $K=20$, noise 20%; FP-VEM

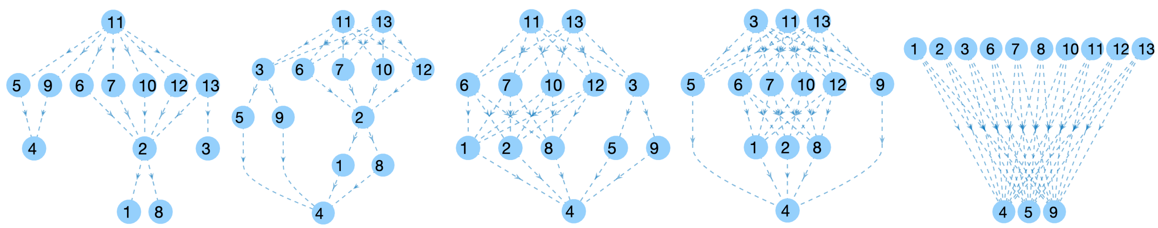


(h) $K=20$, noise 10%; FP-VEM

Figure D.4: Sizes of the finally selected patterns $\widehat{\mathcal{A}}_{\text{PEM}}$ and $\widehat{\mathcal{A}}_{\text{FP-VEM}}$ under the two-parameter SLAM. The "noise" refers to the value of $1-\theta_j^+=\theta_j^-$. The number of true patterns is $|\mathcal{A}_0|=10$.

TIMSS Data: Attribute structures corresponding to different Υ's. For the TIMSS data, we obtain those different attribute structures corresponding to different

Υ's in the FP-VEM algorithm. The results are presented in Figure D.5. Apart from the five structures shown in Figure D.5(a)–(e), the two patterns selected when $\Upsilon \in [0.70, 0.74]$ are the all-zero and the all-one patterns, which do not result in any structure among the 13 attributes. Note that the structure in Figure D.5(d) is equivalent to the structure selected by EBIC in Figure 5.10(b).



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

Figure D.5: Different attribute structures corresponding to various Υ's in Algorithm 2. Plot (d) here is equivalent to Figure 5.10(b), the attribute structure selected by EBIC.

APPENDIX E

Appendix of Chapter VI

This is the appendix to Chapter VI and it is organized as follows. Section E.1 presents the proof of the main theorem, Theorem VI.1. Section E.2 includes some further details on computation, with details of EBIC in Appendix E.2.1, algorithms handling missing data in Appendix E.2.2, and details on the experiments in Section 6.4 of the main text in Appendix E.2.3. Appendix E.3 includes simulation results on large noisy binary matrix factorization/reconstruction and structural matrix estimation. The Matlab codes for implementing the algorithms and reproducing the experimental results are included in another zip archive.

E.1 Proof of Theorem VI.1

There is one basic fact about any attribute hierarchy $\mathcal{E}$ and the resulting $\mathcal{A}$: the all-zero and all-one attribute patterns $\mathbf{0}_K$ and $\mathbf{1}_K$ always belong to $\mathcal{A}$ that is induced by an arbitrary $\mathcal{E}$. This is because any prerequisite relation among attributes would not rule out the existence of the pattern possessing no attributes or the pattern possessing all attributes.

The proofs of part (i) and part (ii) are presented as follows.

Proof of part (i). *We first show the sufficiency of Conditions A, B and C for identifiability of* $(\Gamma(Q,\mathcal{A})$, $\boldsymbol{\theta}^+$, $\boldsymbol{\theta}^-$, $\boldsymbol{p})$. Since Condition A is satisfied, from now on we assume without loss of generality that

$$Q = \begin{pmatrix} Q^0 \\ Q^\star \end{pmatrix}, \quad \Gamma(Q^0,\mathcal{A}) = \Gamma(I_K,\mathcal{A}). \tag{E.1}$$

We next show that if for any $\boldsymbol{r}\in\{0,1\}^J$,

$$T_{\boldsymbol{r},\boldsymbol{\cdot}}(Q,\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)\boldsymbol{p} = T_{\boldsymbol{r},\boldsymbol{\cdot}}(\bar{Q},\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{\theta}}^-)\bar{\boldsymbol{p}}, \tag{E.2}$$

then $\Gamma(\bar{Q},\mathcal{A}) = \Gamma(Q,\mathcal{A})$ and $(\bar{\boldsymbol{\theta}}^+,\bar{\boldsymbol{\theta}}^-,\bar{\boldsymbol{p}}) = (\boldsymbol{\theta}^+,\boldsymbol{\theta}^-,\boldsymbol{p})$. We denote the submatrix of $\bar{Q}$ consisting of its first K row vectors by $\bar{Q}^0$, and the remaining submatrix by $\bar{Q}^\star$, so $\bar{Q} = ((\bar{Q}^0)^\top,(\bar{Q}^\star)^\top)^\top$.

For any item set $S\subseteq\{1,\ldots,J\}$, denote $\boldsymbol{\theta}_S^+ = \sum_{j\in S}\theta_j^+\boldsymbol{e}_j$, and denote $\boldsymbol{\theta}_S^-$, $\bar{\boldsymbol{\theta}}_S^+$, and $\bar{\boldsymbol{\theta}}_S^-$ similarly. Consider the response pattern $\boldsymbol{r}^\star = \sum_{j\in S}\boldsymbol{e}_j$ and any $\boldsymbol{\theta}^\star = \sum_{j\in S}\theta_j^\star\boldsymbol{e}_j$, then

$$T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q,\boldsymbol{\theta}_S^+-\boldsymbol{\theta}^\star,\boldsymbol{\theta}_S^- -\boldsymbol{\theta}^\star)\boldsymbol{p} = T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q},\bar{\boldsymbol{\theta}}_S^+-\boldsymbol{\theta}^\star,\bar{\boldsymbol{\theta}}_S^- -\boldsymbol{\theta}^\star)\bar{\boldsymbol{p}}. \tag{E.3}$$

When there is no ambiguity, we sometimes will denote $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q,\boldsymbol{\theta}_S^+-\boldsymbol{\theta}^\star,\boldsymbol{\theta}_S^- -\boldsymbol{\theta}^\star) = T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}$ for notational simplicity.

We prove the theorem in 6 steps as follows.

Step 1. In this step we show if (E.2) holds, the $\bar{Q}^0$ must also take the following upper-triangular form with all-one diagonal elements, up to column permutation.

$$\bar{Q}^0 = \begin{pmatrix} 1 & * & \ldots & * \\ 0 & 1 & \ldots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}. \tag{E.4}$$

We need the following useful lemmas.

Lemma E.1. *The following statements about Q, $Q^{\star,B}$ and $Q^{\star,C}$ hold.*

(a) If Q satisfies Conditions A with the first K rows forming the Q^0, then for any $k,h \in [K]$ and $k \neq h$, $\boldsymbol{q}_k \succeq \boldsymbol{q}_h$ happens only if $k \to h$.

(b) If Q satisfies Condition B, then any row vector of the modified $Q^{\star,B}$ defined in Condition B represents an attribute pattern that respects the attribute hierarchy. Namely, for any $j \in \{K+1,\dots,J\}$, there is $\boldsymbol{q}_j^B \in \mathcal{A}$. Similarly, if Q satisfies Condition C, any row vector $\boldsymbol{q}_j^C$ in the modified $Q^{\star,C}$ respects the attribute hierarchy.

(c) Suppose Q satisfies Condition B. If $k \to h$ under the attribute hierarchy, then the $Q^{\star,B}$ defined in Condition B must satisfy $Q^{\star,B}_{\bullet,k} \succ Q^{\star,B}_{\bullet,h}$.

Proof of Lemma E.1. For part (a), we call the type of modification of Q described in Condition A by "Operation" A, which sets every $q_{j,k}$ to zero if $q_{j,h} = 1$ and $k \to h$. Denote the resulting matrix by Q^A. If there exists some $\boldsymbol{q}_k \succeq \boldsymbol{q}_h$ for some $k \not\to h$, then Operation A would not set $q_{k,h}$ to zero, and the first rows of Q^A would not be an I_K. So $\boldsymbol{q}_k \succeq \boldsymbol{q}_h$ happens only if $k \to h$. The proof of part (b) is straightforward; it is true by the definition of the attribute hierarchy. For part (c), if $k \to h$, then under Operation B there is $Q^{\star,B}_{\bullet,k} \succeq Q^{\star,B}_{\bullet,h}$. Since Condition B states that $Q^{\star,B}$ has distinct columns, there must be $Q^{\star,B}_{\bullet,k} \succ Q^{\star,B}_{\bullet,h}$.

Lemma E.2. *Suppose the true Q satisfies Conditions A and B under the attribute hierarchy. If there exists an item set $S \subseteq \{K+1,\dots,J\}$ such that*

$$\max_{m\in S} q_{m,h} = 0, \quad \max_{m\in S} q_{m,j} = 1 \ \forall j \in \mathcal{J}$$

for some attribute $h \in [K]$ and a set of attributes $\mathcal{J} \subseteq [K] \setminus \{h\}$, then

$$\vee_{j\in\mathcal{J}} \bar{\boldsymbol{q}}_j \not\succeq \bar{\boldsymbol{q}}_h.$$

Proof of Lemma E.2. We use proof by contradiction. Assume there exist attribute $h \in [K]$ and a set of attributes $\mathcal{J} \subseteq [K]\setminus\{h\}$, such that $\vee_{j\in\mathcal{J}} \bar{\boldsymbol{q}}_j \not\succeq \bar{\boldsymbol{q}}_h$; and that there exists $S \subseteq \{K+1,\dots,J\}$ such that $\max_{m\in S} q_{m,h} = 0$ and $\max_{m\in S} q_{m,j} = 1$. Define

$$\boldsymbol{\theta}^\star = \bar{\theta}_h^+ \boldsymbol{e}_h + \sum_{j\in\mathcal{J}} \bar{\theta}_j^- \boldsymbol{e}_j + \sum_{m=K+1}^{J} \theta_m^- \boldsymbol{e}_m, \quad \boldsymbol{r}^\star = \boldsymbol{e}_h + \sum_{j\in\mathcal{J}} \boldsymbol{e}_j + \sum_{m=K+1}^{J} \boldsymbol{e}_m,$$

and we claim that $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^\star)$ is an all-zero vector. This is because for any $\boldsymbol{\alpha} \in \{0,1\}^K$, the corresponding element in $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^\star)$ contains a factor $F_{\boldsymbol{\alpha}} = (\bar{\theta}_{h,\boldsymbol{\alpha}} - \bar{\theta}_h^+)\prod_{j\in\mathcal{J}}(\bar{\theta}_{j,\boldsymbol{\alpha}} - \bar{\theta}_j^-)$. While this factor $F_{\boldsymbol{\alpha}} \neq 0$ only if $\bar{\theta}_{h,\boldsymbol{\alpha}} = \bar{\theta}_h^-$ and $\bar{\theta}_{j,\boldsymbol{\alpha}} = \bar{\theta}_j^+$ for all $j \in \mathcal{J}$, which happens if and only if $\boldsymbol{\alpha} \not\succeq \bar{\boldsymbol{q}}_h$ and $\boldsymbol{\alpha} \succeq \bar{\boldsymbol{q}}_j$ for all $j \in \mathcal{J}$, which is impossible because $\vee_{j\in\mathcal{J}}\bar{\boldsymbol{q}}_j \succeq \bar{\boldsymbol{q}}_h$ by our assumption. So the claim $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^\star) = \mathbf{0}$ is proved, and further $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = 0$. Equality (E.3) becomes $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^\star)\bar{\boldsymbol{p}} = 0$, which leads to

$$0 = T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star)\boldsymbol{p} = p_{\mathbf{1}}(\theta_h^+ - \bar{\theta}_h^+)\prod_{j\in\mathcal{J}}(\theta_j^+ - \bar{\theta}_j^-)\prod_{m>K}(\theta_m^+ - \theta_m^-),$$

which is because for any $\boldsymbol{\alpha} \neq \mathbf{1}$, we must have $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_m$ for some $m > K$ under Condition C, and hence the element $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star)$ contains a factor $(\theta_m^- - \theta_m^-) = 0$. Since $\theta_m^+ - \theta_m^- > 0$ for $m > K$ and $\theta_j^+ - \bar{\theta}_j^- \neq 0$, we obtain $\theta_h^+ = \bar{\theta}_h^+$.

We remark here that $\theta_h^+ = \bar{\theta}_h^+$ also implies $\bar{\boldsymbol{q}}_h \neq \mathbf{0}$, because otherwise we would have $\bar{\theta}_h = \bar{\theta}_h^+ = \theta_h^+$, which contradicts the $\theta_h^- < \bar{\theta}_h < \theta_h^+$ proved before the current Step 1. This indicates the $\bar{Q}_{1:K,\boldsymbol{\cdot}}$ can not contain any all-zero row vector, because otherwise $\bar{\boldsymbol{q}}_j \succeq \bar{\boldsymbol{q}}_h$ for the all-zero row vector $\bar{\boldsymbol{q}}_h$, which we showed is impossible.

Consider the item set S in the lemma that satisfies $S \subseteq \{K+1, \dots, J\}$ such that $\max_{m \in S} q_{m,h} = 0$ and $\max_{m\in S} q_{m,j} = 1$ for all $j \in \mathcal{J}$. Define

$$\boldsymbol{\theta}^\star = \bar{\theta}_h^+ \boldsymbol{e}_h + \sum_{j \in \mathcal{J}} \bar{\theta}_j^- \boldsymbol{e}_j + \sum_{m \in S} \theta_m^- \boldsymbol{e}_m.$$

Note that $\theta_h^+ = \bar{\theta}_h^+$. The RHS of (E.3) is zero, and so is the LHS of it. The row vector $T_{\boldsymbol{r}^\star, \boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star)$ has the following property

$$\begin{aligned}
&T_{\boldsymbol{r}^\star, \boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star) \\
&= \begin{cases} (\theta_h^- - \bar{\theta}_h^+) \prod_{j \in \mathcal{J}} (\theta_j^+ - \bar{\theta}_j^-) \prod_{m \in S} (\theta_m^+ - \theta_m^-), & \boldsymbol{\alpha} \not\succeq \boldsymbol{q}_h,\, \boldsymbol{\alpha} \succeq \boldsymbol{q}_{\mathcal{J}},\, \boldsymbol{\alpha} \succeq \boldsymbol{q}_S; \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

Note that $\{\boldsymbol{\alpha} \in \{0,1\}^K : \boldsymbol{\alpha} \not\succeq \tilde{\boldsymbol{q}}_h, \boldsymbol{\alpha} \succeq \tilde{\boldsymbol{q}}_{\mathcal{J}}, \boldsymbol{\alpha} \succeq \tilde{\boldsymbol{q}}_S\} = \{\boldsymbol{\alpha} : \boldsymbol{\alpha} \not\succeq \tilde{\boldsymbol{q}}_h, \boldsymbol{\alpha} \succeq \tilde{\boldsymbol{q}}_S\} = \mathcal{A}_1 \neq \varnothing$, because $q_{S,\ell} = 0$ and $q_{S,k} = 1$ hold. Furthermore, we claim that $\sum_{\boldsymbol{\alpha} \in \mathcal{A}_1} p_{\boldsymbol{\alpha}} > 0$ under the specified attribute hierarchy. This is because Lemma E.1 ensures $\tilde{\boldsymbol{q}}_m \in \mathcal{A}$ for the considered $m > K$, and hence the attribute pattern $\boldsymbol{\alpha}^\star = \tilde{\boldsymbol{q}}_m$ belongs to the set $\mathcal{A}_1$ and also belongs to the set $\mathcal{A}$. This ensures $p_{\boldsymbol{\alpha}^\star} > 0$ and $\sum_{\boldsymbol{\alpha} \in \mathcal{A}_1} p_{\boldsymbol{\alpha}} \geq p_{\boldsymbol{\alpha}^\star} > 0$. Therefore we have

$$\begin{aligned}
&T_{\boldsymbol{r}^\star, \boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star) \boldsymbol{p} \\
&= (\theta_\ell^- - \bar{\theta}_\ell^+)(\theta_k^+ - \bar{\theta}_k^-)(\theta_m^+ - \theta_m^-) \Big(\sum_{\boldsymbol{\alpha} \in \mathcal{A}_1} p_{\boldsymbol{\alpha}} \Big) = 0,
\end{aligned}$$

which leads to a contradiction since $\theta_\ell^- - \bar{\theta}_\ell^+ \neq 0$, $\theta_k^+ - \bar{\theta}_k^- \neq 0$, $\theta_m^+ - \theta_m^- \neq 0$ and $\sum_{\boldsymbol{\alpha} \in \mathcal{A}_1} p_{\boldsymbol{\alpha}} > 0$, i.e., every factor in the above product is nonzero. This completes the proof of Lemma E.2. $\square$

We now proceed with the proof of Step 1. We first introduce the *lexicographic order* between two vectors of the same length. For two binary vectors $\boldsymbol{a} = (a_1, \dots, a_L)^\top$

and $\boldsymbol{b} = (b_1, \ldots, b_L)^\top$ both of length L, we say $\boldsymbol{a}$ is of smaller lexicographic order than $\boldsymbol{b}$ and denote $\boldsymbol{a} \prec_{\text{lex}} \boldsymbol{b}$, if either $a_1 < b_1$, or there exists some $l \in \{2, \ldots, L\}$ such that $a_l < b_l$ and $a_m = b_m$ for all $m = 1, \ldots, l-1$. Since $\tilde{Q}^\star$ contains K distinct column vectors, the K columns of $Q^\star$ can be arranged in an increasing lexicographic order. Without loss of generality, we assume that

$$Q^\star_{,1} \prec_{\text{lex}} Q^\star_{,2} \prec_{\text{lex}} \cdots \prec_{\text{lex}} Q^\star_{,K}. \tag{E.5}$$

We use an induction method to prove the conclusion. First consider attribute 1. Since $Q^\star_{,1}$ has the smallest lexicographic order among the columns of $Q^\star$, there must exist an item set $S \subseteq \{K+1, \ldots, J\}$ such that

$$q_{S,1} = 0, \quad q_{S,\ell} = 1 \ \forall \ell = 2, \ldots, K.$$

Based on the above display, we apply Lemma E.2 to obtain

$$\vee_{\ell=2}^{K} \bar{\boldsymbol{q}}_\ell \not\succeq \bar{\boldsymbol{q}}_1.$$

This means there exists $b_1 \in [K]$ such that the b_1-th column vector of $\bar{Q}^0$ must equal the basis vector

$$(\underbrace{1}_{\text{column } 1}, \boldsymbol{0})^\top = \boldsymbol{e}_1,$$

i.e., we have $\bar{Q}^0_{\bullet, b_1} = \boldsymbol{e}_1$.

Now we assume as the inductive hypothesis that for $h \in [K]$ and $h > 1$, we have a distinct set of attributes $\{m_1, \ldots, m_{h-1}\} \subseteq [K]$ such that their corresponding column vectors in $\bar{Q}_{1:K,\bullet}$ satisfy

$$\forall i = 1, \ldots, h-1, \quad \bar{Q}_{1:K,b_i} = (*, \ldots, *, \underbrace{1}_{\text{column } i}, 0, \ldots, 0)^\top. \tag{E.6}$$

Now we consider attribute h. By (E.5), the column vector $Q^\star_{,h}$ has the smallest lexicographic order among the $K-h-1$ columns in $\{Q^\star_{\bullet,h},\, Q^\star_{\bullet,h+1},\, \ldots,\, Q^\star_{\bullet,K}\}$, therefore similar to the argument in the previous paragraph, there must exist an item set $S \subseteq \{K+1, \ldots, J\}$ such that

$$q_{S,h} = 0, \quad q_{S,\ell} = 1 \ \forall \ell = h+1, \ldots, K. \tag{E.7}$$

Therefore Lemma E.2 gives

$$\vee_{\ell=h+1}^{K} \bar{\boldsymbol{q}}_\ell \not\succeq \bar{\boldsymbol{q}}_h,$$

which further implies there exists an attribute b_h such that

$$\max_{\ell \in \{h+1,\ldots,K\}} \bar{q}_{\ell,b_h} = 0, \quad \bar{q}_{h,b_h} = 1. \tag{E.8}$$

We point out that $b_h \notin \{b_1, \ldots, b_{h-1}\}$, because by the induction hypothesis (E.6) we have $\bar{q}_{h,b_i} = 0$ for $i = 1, \ldots, h-1$. So $\{b_1, \ldots, b_{h-1}, b_h\}$ contains h distinct attributes. Furthermore, (E.8) gives that

$$\bar{Q}^0_{\bullet,b_h} = (*, \ldots, *, \underbrace{1}_{\text{column } h}, 0, \ldots, 0)^\top,$$

which generalizes (E.6) by extending $h-1$ there to h. Therefore, we use the induction argument to obtain

$$\forall k \in [K], \quad \bar{Q}^0_{\bullet,b_k} = (*, \ldots, *, \underbrace{1}_{\text{column } k}, 0, \ldots, 0)^\top,$$

which essentially means

$$\bar{Q}^0_{\boldsymbol{\cdot},(b_1,\ldots,b_K)} = \begin{pmatrix} 1 & * & \ldots & * \\ 0 & 1 & \ldots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \tag{E.9}$$

and the conclusion of Step 1 in (E.4) is proved.

Step 2. In this step we prove $\bar{\theta}_j^+ = \theta_j^+$ for all $j \in \{K+1, \ldots, J\}$ in the same way as Step 2 of the proof of Theorem 1 in Gu and Xu (2020b). The fact $p_{\mathbf{1}} > 0$ under any attribute hierarchy is used.

Step 3. In this step we prove $\bar{\theta}_k^- = \theta_k^-$ for all $k \in \{1, \ldots, K\}$ and $\bar{Q}_{1:K,\boldsymbol{\cdot}} \stackrel{\mathcal{E}}{\sim} I_K$. We use an induction method here.

Step 3.1. First consider those attribute k for which there *does not exist* another attribute h such that $\tilde{Q}^\star_{\boldsymbol{\cdot},h} \prec \tilde{Q}^\star_{\boldsymbol{\cdot},k}$; and we first aim to show $\bar{\theta}_k^- = \theta_k^-$ for such k. By part (c) of Lemma E.1, we have that $k \not\to h$ for any attribute $h \neq k$. For this k, define

$$\boldsymbol{\theta}^\star = \sum_{j=1}^K \bar{\theta}_j^- \boldsymbol{e}_j + \sum_{\substack{j>K: \\ q_{j,k}=0}} \theta_j^- \boldsymbol{e}_j + \sum_{\substack{j>K: \\ q_{j,k}=1}} \theta_j^+ \boldsymbol{e}_j, \tag{E.10}$$

then $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(\bar{Q}, \bar{\boldsymbol{\theta}}^+ - \boldsymbol{\theta}^\star, \bar{\boldsymbol{\theta}}^- - \boldsymbol{\theta}^\star) = \mathbf{0}$. Further, we claim $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star)$ would equal zero for any $\boldsymbol{\alpha} \neq (1, \ldots, 1, \underbrace{0}_{\text{column } k}, 1, \ldots, 1) =: \boldsymbol{\alpha}^\star$, so the only potentially nonzero element in $T_{\boldsymbol{r}^\star,\boldsymbol{\cdot}}$ is $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}^\star}$. More specifically,

$$\begin{aligned} & T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}}(Q, \boldsymbol{\theta}^+ - \boldsymbol{\theta}^\star, \boldsymbol{\theta}^- - \boldsymbol{\theta}^\star) \\ = & \begin{cases} (\theta_k^- - \bar{\theta}_k^-) \prod_{\substack{j \leq K: \\ j \neq k}} (\theta_j^+ - \bar{\theta}_j^-) \prod_{\substack{j>K: \\ q_{j,k}=0}} (\theta_j^+ - \theta_j^-) \prod_{\substack{j>K: \\ q_{j,k}=1}} (\theta_j^- - \theta_j^+), & \boldsymbol{\alpha} = \boldsymbol{\alpha}^\star; \\ 0, & \boldsymbol{\alpha} \neq \boldsymbol{\alpha}^\star. \end{cases} \end{aligned} \tag{E.11}$$

The reasoning behind (E.11) is as follows. Consider any other attribute pattern $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^\star$ with $\alpha_h = 0$ for some $h \neq k$. Since for k we have $\tilde{Q}^\star_{\bullet,k} \nsucceq \tilde{Q}^\star_{\bullet,h}$ for any $h \neq k$, there must exist some item $j > K$ s.t. $q_{j,k} = 0$ and $q_{j,h} = 1$. For this particular item j, we have $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}}$ contains a factor of $(\theta_{j,\boldsymbol{\alpha}} - \theta_j^-) = (\theta_j^- - \theta_j^-) = 0$, so $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}} = 0$. This shows that $T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}} \neq 0$ only if $\alpha_h = 1$ for all $h \neq k$. Further more, we claim that $T_{\boldsymbol{r}^\star,\mathbf{1}_K} = 0$ also holds; this is because there exists $j > K$ s.t. $q_{j,k} = 1$, and for this particular item j we have $\theta_{j,\mathbf{1}_K} = \theta_j^+$ so $T_{\boldsymbol{r}^\star,\mathbf{1}_K}$ contains a factor of $(\theta_j^+ - \theta_j^+) = 0$. Now we have shown (E.11) holds. Equation (E.3) leads to

$$\begin{aligned} 0 &= \sum_{\boldsymbol{\alpha}\in\mathcal{A}_0} T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = T_{\boldsymbol{r}^\star,\boldsymbol{\alpha}^\star} p_{\boldsymbol{\alpha}^\star} \\ &= (\theta_k^- - \bar{\theta}_k^-) \prod_{\substack{j\leq K:\\ j\neq k}} (\theta_j^+ - \bar{\theta}_j^-) \prod_{\substack{j>K:\\ q_{j,k}=0}} (\theta_j^+ - \theta_j^-) \prod_{\substack{j>K:\\ q_{j,k}=1}} (\theta_j^- - \theta_j^+) p_{\boldsymbol{\alpha}^\star}. \end{aligned} \tag{E.12}$$

We claim that $\boldsymbol{\alpha}^\star$ respects the attribute hierarchy so $p_{\boldsymbol{\alpha}^\star} > 0$. This is true because we have shown earlier $k \not\to h$ for any attribute $h \neq k$. Therefore in (E.12) the only factor that could potentially be zero is $(\theta_k^- - \bar{\theta}_k^-)$, and we obtain $\bar{\theta}_k^- = \theta_k^-$. This completes the first step of the induction.

Step 3.2. Now as the inductive hypothesis, we consider attribute k and assume that for any other attribute h s.t. $\tilde{Q}^\star_{\bullet,h} \prec \tilde{Q}^\star_{\bullet,k}$, we already have $\bar{\theta}_h^- = \theta_h^-$. Recall $\mathcal{H}_k = \{h \in [K] \setminus \{k\} : k \to h\}$ denotes all the attributes that have higher level in the attribute hierarchy than attribute k. By part (c) of Lemma E.1, this implies for any $h \in \mathcal{H}_k$, we have $\bar{\theta}_h^- = \theta_h^-$. Also, by Condition C in the theorem, there exist two items $j_1, j_2 > K$ s.t. $q_{j_i,k} = 1$ and $q_{j_i,h} = 0$ for all $h \in \mathcal{H}_k$, for $i = 1, 2$.

Before proceeding with the proof of $\bar{\theta}_k^- = \theta_k^-$, we need to introduce a useful lemma.

Lemma E.3. *Under the conditions of theorem, if* $\vee_{h\in\mathcal{K}} \bar{\boldsymbol{q}}_h \succeq \bar{\boldsymbol{q}}_m$ *for some* $\mathcal{K} \subseteq [J]$, *some* $m \in [J] \setminus \mathcal{K}$ *and* $\#[(\mathcal{K} \cup \{m\}) \cap \{K+1, \ldots, J\}] \leq 1$, *then* $\bar{\theta}_m^+ = \theta_m^+$.

Proof of Lemma E.3. Define

$$\boldsymbol{\theta}^* = \sum_{h\in\mathcal{K}} \bar{\theta}_h^- \boldsymbol{e}_h + \bar{\theta}_m^+ \boldsymbol{e}_m + \sum_{\substack{l>K:\\ l\notin\mathcal{K}\cup\{m\}}} \theta_l^- \boldsymbol{e}_l,$$

then $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}}$ contains a factor $\bar{f}_{\boldsymbol{\alpha}} := \prod_{h\in\mathcal{K}}(\bar{\theta}_{h,\boldsymbol{\alpha}} - \bar{\theta}_h^-)(\bar{\theta}_{m,\boldsymbol{\alpha}} - \bar{\theta}_m^+)$ because of the first two terms in the above display. The $\bar{f}_{\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \succeq \vee_{h\in\mathcal{K}} \bar{\boldsymbol{q}}_h$ and $\boldsymbol{\alpha} \nsucceq \bar{\boldsymbol{q}}_m$. However, since $\vee_{h\in\mathcal{K}} \bar{\boldsymbol{q}}_h \succeq \bar{\boldsymbol{q}}_m$, such $\boldsymbol{\alpha}$ does not exist and $\bar{f}_{\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$. Therefore $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}} = \boldsymbol{0}$ and $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} = 0$, so the RHS of (E.3) is zero. Hence the LHS of (E.3) is also zero. Condition C implies $\sum_{j=K+1}^J q_{j,k} \geq 2$ for all attribute k. Under Condition C and the condition $\#[(\mathcal{K}\cup\{m\})\cap\{K+1,\ldots,J\}] \leq 1$, the attributes required by the items in the set $\{l > K : l \notin \mathcal{K}\cup\{m\}\}$ must cover all the K attributes. because of the term $\sum_{\substack{l>K:\\ l\notin\mathcal{K}\cup\{m\}}} \theta_l^- \boldsymbol{e}_l$ in the defined $\boldsymbol{\theta}^*$, we have $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} = \boldsymbol{1}_K$. So

$$\begin{aligned} 0 &= \text{RHS of (E.3)} = \text{LHS of (E.3)} \\ &= \prod_{h\in\mathcal{K}}(\theta_h^+ - \bar{\theta}_h^-)(\theta_m^+ - \bar{\theta}_m^+) \prod_{\substack{l>K:\\ l\notin\mathcal{K}\cup\{m\}}} (\theta_l^+ - \theta_l^-) p_{\boldsymbol{1}_K}, \end{aligned}$$

which implies $\theta_m^+ - \bar{\theta}_m^+ = 0$ since any other factor in the above display is nonzero. This completes the proof of the lemma. $\square$

Note that by Condition C, there exist two different items $j_1, j_2 > K$ s.t. $q_{j_i,k} = 1$ and $q_{j_i,h} = 0$ for all $h \in \mathcal{H}_k$ for $i = 1,2$. We next aim to show that in $\bar{Q}$, we must also have $\bar{\boldsymbol{q}}_{j_i,h} = 0$ for all $h \in \mathcal{H}_k$ for $i = 1,2$. We prove this in two steps.

Step 3.2 Part I. First, we use proof by contradiction to show the $\bar{\boldsymbol{q}}_h$ satisfies that, for any attribute $m \not\to h$ the following holds,

$$\max(\max_{\substack{\ell\in[K],\\ Q^\star_{\boldsymbol{\cdot},\ell}\nsucceq Q^\star_{\boldsymbol{\cdot},m}}} \bar{\boldsymbol{q}}_\ell,\ \bar{\boldsymbol{q}}_h) \nsucceq \bar{\boldsymbol{q}}_m, \tag{E.13}$$

where the max operator applied to vectors of the same length means taking the element-wise maximum of the vectors and obtaining a new vector of that same length. Suppose (E.13) does not hold, then applying Lemma E.3 we obtain $\bar{\theta}_m^+ = \theta_m^+$. Note that we also have $\bar{\theta}_h^- = \theta_h^-$ by the inductive hypothesis. Define

$$\boldsymbol{\theta}^* = \bar{\theta}_h^- \boldsymbol{e}_h + \sum_{\substack{\ell \le K:\\ Q^\star_{\bullet,\ell} \not\succeq Q^\star_{\bullet,m}}} \bar{\theta}_\ell^- \boldsymbol{e}_\ell + \bar{\theta}_m^+ \boldsymbol{e}_m + \sum_{\substack{j>K:\\ q_{j,m}=0}} \theta_j^- \boldsymbol{e}_j, \tag{E.14}$$

then with this $\boldsymbol{\theta}^*$, we claim that the RHS of (E.3) is zero, $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} = 0$. This claim is true because $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}}$ contains a factor $f_{\boldsymbol{\alpha}}$ of the following form

$$f_{\boldsymbol{\alpha}} = (\bar{\theta}_{h,\boldsymbol{\alpha}} - \bar{\theta}_h^-) \prod_{l:\, Q^\star_{\bullet,\ell} \not\succeq Q^\star_{\bullet,m}} (\bar{\theta}_{\ell,\boldsymbol{\alpha}} - \bar{\theta}_\ell^-)(\bar{\theta}_{m,\boldsymbol{\alpha}} - \bar{\theta}_m^+) \neq 0 \quad \text{only if}$$

$$\boldsymbol{\alpha} \succeq \max\Big(\max_{\substack{\ell \in [K],\\ Q^\star_{\bullet,\ell} \not\succeq Q^\star_{\bullet,m}}} \bar{\boldsymbol{q}}_\ell,\ \bar{\boldsymbol{q}}_h\Big) \quad \text{and} \quad \boldsymbol{\alpha} \not\succeq \bar{\boldsymbol{q}}_m,$$

which is impossible because of (E.13), so $f_{\boldsymbol{\alpha}} = 0$ and $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha}$. Therefore by (E.3) we have $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = \overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} = 0$. Note that $\bar{\theta}_h^- = \theta_h^-$ and $\bar{\theta}_m^+ = \theta_m^+$, and now we consider the term $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}$. Then due to the last term in $\boldsymbol{\theta}^*$ defined in (E.14), we have $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$ for all $j > K$ s.t. $q_{j,m} = 0$. We claim that such $\boldsymbol{\alpha}$ must also satisfy $\boldsymbol{\alpha} \succeq \boldsymbol{q}_\ell$ for any $\ell \le K$ s.t. $Q^\star_{\bullet,\ell} \not\succeq Q^\star_{\bullet,m}$. This is because for any $\ell \le K$ s.t. $Q^\star_{\bullet,\ell} \not\succeq Q^\star_{\bullet,m}$, there must exist an item $j > K$ such that $q_{j,m} = 0$ and $q_{j,\ell} = 1$, then the fact that $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$ for this j ensures $\alpha_\ell = 1$ and $\boldsymbol{\alpha} \succeq \boldsymbol{q}_\ell$ (recall $\boldsymbol{q}_\ell \overset{\mathcal{E}}{\sim} \boldsymbol{e}_k$). Therefore we have

$$T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = \begin{cases} (\theta_h^+ - \bar{\theta}_h^-) \prod_{\substack{\ell \le K:\\ Q^\star_{\bullet,\ell} \not\succeq Q^\star_{\bullet,m}}} (\theta_\ell^+ - \bar{\theta}_\ell^-)(\theta_m^- - \bar{\theta}_m^+) \prod_{\substack{j>K:\\ q_{j,m}=0}} (\theta_j^+ - \theta_j^-), & \text{if } \boldsymbol{\alpha} \in \mathcal{A}_1; \\ 0, & \text{otherwise.} \end{cases} \tag{E.15}$$

where

$$\begin{aligned}\mathcal{A}_1 &= \{\boldsymbol{\alpha} \in \mathcal{A} : \boldsymbol{\alpha} \succeq \boldsymbol{q}_j \text{ s.t. } q_{j,m} = 0;\ \boldsymbol{\alpha} \succeq \boldsymbol{q}_h;\ \boldsymbol{\alpha} \not\succeq \boldsymbol{q}_m\} \\ &= \{\boldsymbol{\alpha} \in \mathcal{A} : \alpha_\ell = 1 \text{ for all } \ell \text{ s.t. } Q^\star_{\bullet,\ell} \not\succ Q^\star_{\bullet,m};\ \alpha_h = 1;\ \alpha_m = 0\}.\end{aligned}$$

We claim that there exists some attribute pattern in $\mathcal{A}_1$ that respects the attribute hierarchy, i.e., there exists $\boldsymbol{\alpha}^\star \in \mathcal{A}_1$ with $p_{\boldsymbol{\alpha}^\star} > 0$. This can be seen by noting the following two facts: first, the assumption $m \not\to h$ in the beginning of the current Step 3.2.1 yields that an $\boldsymbol{\alpha}$ with $\alpha_m = 0$ and $\alpha_h = 1$ does not violate the attribute hierarchy; second, an $\boldsymbol{\alpha}$ satisfying $\alpha_\ell = 1$ for all ℓ s.t. $Q^\star_{\bullet,\ell} \not\succ Q^\star_{\bullet,m}$ also does not contradict $\alpha_m = 0$ under the hierarchy, because by part (c) of Lemma E.1, if $Q^\star_{\bullet,\ell} \not\succ Q^\star_{\bullet,m}$ then $m \not\to h$. Now we have proven the claim there exists $\boldsymbol{\alpha}^\star \in \mathcal{A}_1$ with $p_{\boldsymbol{\alpha}^\star} > 0$. Combined with (E.15), we obtain

$$(\theta_h^+ - \bar{\theta}_h^-) \prod_{\substack{\ell \le K: \\ Q^\star_{\bullet,\ell} \not\succ Q^\star_{\bullet,m}}} (\theta_\ell^+ - \bar{\theta}_\ell^-)(\theta_m^- - \bar{\theta}_m^+) \prod_{\substack{j > K: \\ q_{j,m}=0}} (\theta_j^+ - \theta_j^-)\Big(\sum_{\boldsymbol{\alpha} \in \mathcal{A}_1} p_{\boldsymbol{\alpha}}\Big) = 0;$$

and $\sum_{\boldsymbol{\alpha} \in \mathcal{A}_1} p_{\boldsymbol{\alpha}} \succeq p_{\boldsymbol{\alpha}^\star} > 0$. This gives a contradiction because each factor in the above display is nonzero. Now we have reached the goal of Step 3.2.1 of proving (E.13).

We remark here that (E.13) has some nice consequences. Considering the $K \times K$ matrix $\bar{Q}^0_{\bullet,(b_1,\ldots,b_K)}$ in (E.4) shown in Step 1 and the particular attribute h, we actually have obtained that for any $m \not\to h$, the m-th column of $\bar{Q}^0_{\bullet,(b_1,\ldots,b_K)}$ not only has the last $(K-m)$ entries equal to zero, but also has $\bar{Q}^0_{h,b_m} = 0$. Equivalently, considering the columns of $\bar{Q}$ are arranged just in the order $(b_1, \ldots, b_K)$ without loss of generality, we have

$$\bar{q}_{h,m} = 0 \text{ for any attribute } m \not\to h. \tag{E.16}$$

Step 3.2 Part II. In this step we use proof by contradiction to show that for $i = 1$

and 2 there is

$$\max\Big(\max_{\ell\leq K:\,\ell\to h}\bar{\boldsymbol{q}}_\ell,\ \bar{\boldsymbol{q}}_{j_i}\Big)\not\succeq\bar{\boldsymbol{q}}_h. \tag{E.17}$$

Suppose (E.17) does not hold for $i=1$, i.e., $\max(\max_{\ell\leq K:\,\ell\to h}\bar{\boldsymbol{q}}_\ell,\ \bar{\boldsymbol{q}}_{j_1})\succeq\bar{\boldsymbol{q}}_h$. Then by Lemma E.3 we have $\bar{\theta}_h^+=\theta_h^+$. We define

$$\boldsymbol{\theta}^*=\bar{\theta}_h^+\boldsymbol{e}_h+\sum_{\substack{\ell\leq K:\\ \ell\to h}}\bar{\theta}_\ell^-\boldsymbol{e}_\ell+\bar{\theta}_{j_1}^-\boldsymbol{e}_{j_1}+\sum_{\substack{j>K:\,j\neq j_1,\\ q_{j,h}=0}}\theta_j^-\boldsymbol{e}_j, \tag{E.18}$$

and note that the item j_2 is included in the last term of summation above since $q_{j_2,h}=0$. With $\boldsymbol{\theta}^*$ defined as in (E.18), we have $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}}=0$ for all $\boldsymbol{\alpha}$ because of the first three terms in (E.18) and the assumption that $\max(\max_{\ell\leq K:\,\ell\to h}\bar{\boldsymbol{q}}_\ell,\ \bar{\boldsymbol{q}}_{j_1})\succeq\bar{\boldsymbol{q}}_h$. So (E.3) gives $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p}=\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}}=0$. Consider $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}$, then $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}\neq 0$ only if $\boldsymbol{\alpha}\not\succeq\boldsymbol{q}_h$ and $\boldsymbol{\alpha}\succeq\boldsymbol{q}_{j_2}$ because of the terms $\bar{\theta}_h^+\boldsymbol{e}_h$ and $\theta_{j_2}^-\boldsymbol{e}_{j_2}$ included in $\boldsymbol{\theta}^*$ defined in (E.18). Further, because of the last term in $\boldsymbol{\theta}^*$ defined in (E.18), we have $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}\neq 0$ only if $\boldsymbol{\alpha}$ satisfies $\alpha_k=1$, $\alpha_h=0$, and

$$\alpha_m=1\ \forall m \text{ s.t. } \exists j>K,\ j\neq j_1,\ q_{j,h}=0,\ q_{j,m}=1,$$

or equivalently,

$$\alpha_m=1\ \forall m \text{ s.t. } Q^\star_{-j_1,m}\not\prec Q^\star_{-j_1,h}. \tag{E.19}$$

We claim that any such $\boldsymbol{\alpha}$ satisfying $\alpha_k=1$, $\alpha_h=0$, and (E.19) also satisfies $\boldsymbol{\alpha}\succeq\boldsymbol{q}_{j_1}$, because of the reasoning as follows. We next show $\alpha_b\geq q_{j_1,b}$ for all attribute b. Define

$$\boldsymbol{\theta}^{**}=\boldsymbol{\theta}^* \text{ in (E.18)}+\sum_{\substack{b\leq K:\,k\not\to b,\\ Q^\star_{-j_1,b}\prec Q^\star_{-j_1,h}}}\theta_b^-\boldsymbol{e}_b, \tag{E.20}$$

and with this $\boldsymbol{\theta}^{**}$ and its corresponding response pattern $\boldsymbol{r}^{**}$, we still have $\overline{T}_{\boldsymbol{r}^{**},\boldsymbol{\cdot}}\bar{\boldsymbol{p}}=0$

and hence $T_{\boldsymbol{r}^{**},\boldsymbol{\cdot}}\boldsymbol{p}=0$. The $T_{\boldsymbol{r}^{**},\boldsymbol{\alpha}}\neq 0$ only if $\boldsymbol{\alpha}$ satisfies

$$\begin{cases}\alpha_k=1,\ \alpha_h=0,\\ \alpha_m=1\ \forall m \text{ s.t. } Q^\star_{-j_1,m}\not\prec Q^\star_{-j_1,h},\\ \alpha_b=1\ \forall b \text{ s.t. } Q^\star_{-j_1,b}\prec Q^\star_{-j_1,h} \text{ and } k\not\to b.\end{cases} \tag{E.21}$$

We denote the set of attribute patterns having the above properties by $\mathcal{A}_2=\{\boldsymbol{\alpha}\in\{0,1\}^K:\boldsymbol{\alpha} \text{ satisfies (E.21)}\}$. Note the following two things: (i) first, $Q^\star_{-j_1,m}\not\prec Q^\star_{-j_1,h}$ implies $m\not\to h$, because otherwise by Lemma E.1 there is $Q^\star_{\boldsymbol{\cdot},m}\prec Q^\star_{\boldsymbol{\cdot},h}$ and hence $Q^\star_{-j_1,m}\prec Q^\star_{-j_1,h}$; (ii) second, $k\not\to b$ implies $h\not\to b$, since otherwise $h\to b$ and $k\to h$ would imply $k\to b$. And we have the conclusion that there exists some $\boldsymbol{\alpha}^\star\in\mathcal{A}_2$ that respects the attribute hierarchy with $p_{\boldsymbol{\alpha}^\star}>0$, because $\alpha_h=0$ does not contradict any $\alpha_\ell=1$ as specified in (E.21) according to (i) and (ii). We next show that for $\boldsymbol{\alpha}\in\mathcal{A}_2$, $\alpha_b\geq q_{j_1,b}$ for any b must hold. To show this we only need to consider those b such that $q_{j_1,b}=1$ and show any $\boldsymbol{\alpha}\in\mathcal{A}_2$ must have $\alpha_b=1$ for such b. By Condition C, $q_{j_1,b}=1$ implies $b\notin\mathcal{H}_k$ (i.e., $k\not\to b$). Then for such b, if $Q^\star_{-j_1,b}\not\prec Q^\star_{-j_1,h}$, then by (E.21) we have $\alpha_b=1$; and if $Q^\star_{-j_1,b}\prec Q^\star_{-j_1,h}$, combining the fact that $k\not\to h$, by (E.21) we also have $\alpha_b=1$. So the conclusion that $\boldsymbol{\alpha}\in\mathcal{A}_2$, $\alpha_b\geq q_{j_1,b}$ for any b is reached.

Now we have obtained for $\boldsymbol{\alpha}\in\mathcal{A}_2$ there is $\boldsymbol{\alpha}\succeq\boldsymbol{q}_{j_1}$. This results in $\boldsymbol{\alpha}\succeq\boldsymbol{q}_j$ for any $j>K$ s.t. $q_{j,h}=0$, i.e, $\boldsymbol{\alpha}\succeq\max_{j>K:\,q_{j,h}=0}\boldsymbol{q}_j$. We further claim that for any $\boldsymbol{\alpha}\in\mathcal{A}_2$, the $\boldsymbol{\alpha}\succeq\boldsymbol{q}_\ell$ for all $\ell\to h$ must hold. This is because by Condition C, for any $\ell\to h$ there exists $j>K$ such that $q_{j,h}=0$ and $q_{j,\ell}=1$. And combining with the previously obtained $\boldsymbol{\alpha}\succeq\max_{j>K:q_{j,h}=0}\boldsymbol{q}_j$, we have the conclusion that $\alpha_\ell=1$ and $\boldsymbol{\alpha}\succeq\boldsymbol{q}_\ell$. Therefore $\boldsymbol{\alpha}\succeq\boldsymbol{q}_\ell$ for all $\ell\to h$. Considering the $T_{\boldsymbol{r}^{**},\boldsymbol{\cdot}}\boldsymbol{p}=0$ with $\boldsymbol{\theta}^{**}$

defined in (E.20), we have

$$(\theta_h^- - \bar{\theta}_h^+) \prod_{\substack{\ell \leq K: \\ \ell \to h}} (\theta_\ell^+ - \bar{\theta}_\ell^-)(\theta_{j_1}^+ - \bar{\theta}_{j_1}^-) \prod_{\substack{j > K:\, j \neq j_1, \\ q_{j,h}=0}} (\theta_j^+ - \theta_j^-) \Big(\sum_{\boldsymbol{\alpha} \in \mathcal{A}_2} p_{\boldsymbol{\alpha}} \Big) = 0.$$

This leads to a contradiction, since every factor in the above display is nonzero. Now we have reached the goal of Step 3.2.2 of proving (E.17) for $i = 1$, and using the exactly same argument gives (E.17) for $i = 2$.

Combining the results of Step 3.2.1 (in (E.13)) and Step 3.2.2 (in (E.17)), we obtain an important observation that

$$\bar{q}_{j_i,h} = 0 \;\forall\; h \in \mathcal{H}_k,\; i = 1, 2. \tag{E.22}$$

This is true because Step 3.2.1 reveals $\bar{q}_{h,\ell}$ can potentially equal one only for those ℓ that is the prerequisite of attribute h (i.e., $\bar{q}_{h,\ell} = 1$ only if $\ell \to h$); and further, Step 3.2.2 establishes that taking the element-wise maximum of the vector $\max_{l \to h} \bar{\boldsymbol{q}}_\ell$ and the vector $\bar{\boldsymbol{q}}_{j_i}$ still does not give a vector that requires all the attributes covered by $\bar{\boldsymbol{q}}_h$. Therefore $\bar{q}_{j_i,h}$ must equal zero. Precisely, (E.13) in Step 3.2.1 implies $\bar{\boldsymbol{q}}_h - \max_{\ell \leq K:\, \ell \to h} \bar{\boldsymbol{q}}_\ell = (0, \ldots, 0, \underbrace{1}_{\text{column } h}, 0, \ldots, 0)$. And Step 3.2.2 further implies $\bar{q}_{j_i,h} = 0$, since otherwise $\max\Big(\max_{\ell \leq K:\, \ell \to h} \bar{\boldsymbol{q}}_\ell,\; \bar{\boldsymbol{q}}_{j_i}\Big) \succeq \bar{\boldsymbol{q}}_h$ would happen, contradicting (E.17).

Step 3.2 Part III. In this step we prove $\bar{\theta}_k^- = \theta_k^-$ based on (E.22). Define

$$\boldsymbol{\theta}^* = \bar{\theta}_k^- + \sum_{\substack{m \leq K:\, m \neq k \\ m \notin \mathcal{H}_k}} \bar{\theta}_m^- \boldsymbol{e}_m + \sum_{j > K:\, q_{j,k}=1} \theta_j^+ \boldsymbol{e}_j + \sum_{j > K:\, q_{j,k}=0} \theta_j^- \boldsymbol{e}_j, \tag{E.23}$$

and we claim that $\overline{T}_{\boldsymbol{r}^*, \boldsymbol{\cdot}} \bar{\boldsymbol{p}} = 0$ with this $\boldsymbol{\theta}^*$ defined above, because of the following reasoning. First, due to the first two terms in (E.23), $\overline{T}_{\boldsymbol{r}^*, \boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha}$ satisfies $\alpha_k = 1$ and $\alpha_m = 1$ for any attribute $m \notin \{k\} \cup \mathcal{H}_k$. Note that in Step 2 we obtained $\bar{\theta}_j^+ = \theta_j^+$ for all $j > K$, then $\overline{T}_{\boldsymbol{r}^*, \boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \in \{\boldsymbol{\alpha} :\, \boldsymbol{\alpha} \not\succeq \bar{\boldsymbol{q}}_j \;\forall\; j > K$ s.t. $q_{j,k} =$

$1\} =: \mathcal{A}_3$. However considering the item j_1 with the property $q_{j_1,k} = 1$ and $q_{j_1,h} = 0$ for all $h \in \mathcal{H}_k$, then such item j_1 must be included in the third term in (E.23) (i.e., $\sum_{j>K:\, q_{j,k}=1} \theta_j^+ \boldsymbol{e}_j$), and we have shown (E.22) in Step 3.2.1 and 3.2.2 that $\bar{q}_{j_i,h} = 1$ only if $h \notin \mathcal{H}_k$. This implies that for all $\boldsymbol{\alpha} \in \mathcal{A}_3$, there must be $\boldsymbol{\alpha} \succeq \bar{\boldsymbol{q}}_{j_i}$ and $\bar{\theta}_{j_i,\boldsymbol{\alpha}} = \bar{\theta}_{j_i}^+$. So we have shown that for any $\boldsymbol{\alpha} \in \{0,1\}^K$, there must be $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 0$, and the claim that $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} = 0$ is proved. And we have $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = 0$.

Next, we consider $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}$. Due to the last two terms in (E.23), $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \in \mathcal{A}_4$ with $\mathcal{A}_4$ defined as

$$\mathcal{A}_4 = \{\boldsymbol{\alpha} :\, \boldsymbol{\alpha} \succeq \boldsymbol{q}_j \;\forall\, j > K \text{ s.t. } q_{j,k} = 0; \quad \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_j \;\forall\, j > K \text{ s.t. } q_{j,k} = 1\}.$$

We claim that for any $\boldsymbol{\alpha} \in \mathcal{A}_4$, there is $\boldsymbol{\alpha} \succeq \boldsymbol{q}_m$ for all $m \notin \mathcal{H}_k$. This claim is true because $\boldsymbol{\alpha} \in \mathcal{A}_4$ implies $\alpha_m = 1$ for all attribute m such that $\mathbf{Q}^\star_{\boldsymbol{\cdot},m} \nprec Q^\star_{\boldsymbol{\cdot},k}$. Recall our inductive hypothesis made in Step 3.1 that $\bar{\theta}_m^- = \theta_m^-$ for all attribute m that satisfies $Q^\star_{\boldsymbol{\cdot},m} \prec Q^\star_{\boldsymbol{\cdot},k}$, then we have $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha}$ further belongs to the following set $\mathcal{A}_5$,

$$\begin{aligned}
\mathcal{A}_5 = \{\boldsymbol{\alpha} :\, & \alpha_m = 1 \;\forall m \in [K] \text{ s.t. } Q^\star_{\boldsymbol{\cdot},m} \nprec Q^\star_{\boldsymbol{\cdot},k} \text{ (due to the last two terms in (E.23))}; \\
& \alpha_m = 1 \;\forall m \in [K] \text{ s.t. } Q^\star_{\boldsymbol{\cdot},m} \prec Q^\star_{\boldsymbol{\cdot},k} \text{ and } m \notin \mathcal{H}_k \\
& \qquad\qquad \text{(due to the 2nd term in (E.23))}\} \\
= \{\boldsymbol{\alpha} :\, & \alpha_m = 1 \;\forall m \in [K] \text{ s.t. } m \notin \mathcal{H}_k\},
\end{aligned}$$

where the last equality uses Lemma E.1 that $Q^\star_{\boldsymbol{\cdot},m} \nprec Q^\star_{\boldsymbol{\cdot},k}$ implies $k \not\to m$. From $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \in \mathcal{A}_5$, we have that for all $\boldsymbol{\alpha} \in \mathcal{A}_5$, there is $\boldsymbol{\alpha} \succeq \boldsymbol{q}_m$ for any attribute $m \notin \mathcal{H}_k$, and hence $\theta_{m,\boldsymbol{\alpha}} = \theta_m^+$.

Furthermore, we claim that if $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ (which implies $\boldsymbol{\alpha} \in \mathcal{A}_5$), we have $\boldsymbol{\alpha} \nsucceq \boldsymbol{q}_k$ for the following reason. For $\boldsymbol{\alpha} \in \mathcal{A}_5$, there is $\alpha_m = 1$ for all $m \notin \mathcal{H}_k$. Consider the item j_1 with $q_{j_1,k} = 1$ and $q_{j_1,h} = 0$ for all $h \in \mathcal{H}_k$, and for this j_1, there is

$$\boldsymbol{\alpha} \succeq \boldsymbol{q}_{j_1} - (0, \ldots, 0, \underbrace{1}_{\text{column } k}, 0, \ldots, 0). \tag{E.24}$$

Then since $\theta^+_{j_1}\boldsymbol{e}_{j_1}$ is included in (E.23), in order to have $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ we must have $\boldsymbol{\alpha} \not\succeq \boldsymbol{q}_{j_1}$. Combined with the above (E.24), we obtain $\alpha_k = 0$ and $\theta_{k,\boldsymbol{\alpha}} = \theta^-_k$. Denote $\mathcal{A}_6 = \mathcal{A}_5 \cap \{\boldsymbol{\alpha} : \alpha_k = 0\}$, and we have $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \in \mathcal{A}_6$. Importantly, any $\boldsymbol{\alpha}$ in $\mathcal{A}_6$ does not violate the attribute hierarchy since $\alpha_k = 0$ does not contradict $\alpha_m = 1$ for $m \notin \mathcal{H}_k$ as specified in $\mathcal{A}_5$. Therefore $p_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha} \in \mathcal{A}_6$ under the attribute hierarchy.

Finally, with (E.23), we conclude that

$$T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = \begin{cases} (\theta^-_k - \bar\theta^-_k) \prod_{\substack{m \leq K:\, m \neq k \\ m \notin \mathcal{H}_k}} (\theta^+_m - \bar\theta^-_m) \prod_{j > K:\, q_{j,k}=1} (\theta^-_j - \theta^+_j) \prod_{j > K:\, q_{j,k}=0} (\theta^+_j - \theta^-_j), & \boldsymbol{\alpha} \in \mathcal{A}_6; \\ 0, & \text{otherwise.} \end{cases}$$

and further

$$\begin{aligned} 0 &= T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} \\ &= (\theta^-_k - \bar\theta^-_k) \prod_{\substack{m \leq K:\, m \neq k \\ m \notin \mathcal{H}_k}} (\theta^+_m - \bar\theta^-_m) \prod_{j > K:\, q_{j,k}=1} (\theta^-_j - \theta^+_j) \prod_{j > K:\, q_{j,k}=0} (\theta^+_j - \theta^-_j) \Big(\sum_{\boldsymbol{\alpha} \in \mathcal{A}_6} p_{\boldsymbol{\alpha}} \Big). \end{aligned}$$

Then since in the last paragraph we have shown $\sum_{\boldsymbol{\alpha} \in \mathcal{A}_6} p_{\boldsymbol{\alpha}} > 0$, the only potentially zero factor in the above display could only be $(\theta^-_k - \bar\theta^-_k)$. Now we have obtained $\bar\theta^-_k = \theta^-_k$, and the proof of Step 3.2.3 is complete.

Step 3.3. Now we complete the inductive argument in the current Step 3 and conclude $\bar\theta^-_k = \theta^-_k$ for all attribute $k \in [K]$. By completing the induction, we have obtained one more useful byproduct in the proof of Step 3, which is (E.16) that

$\bar{q}_{h,m} = 0$ for any attribute $m \nrightarrow h$. This exactly means under the true attribute hierarchy and the induced attribute pattern set $\mathcal{A}$, the first K items of $\bar{Q}$ is equivalent to the identity matrix I_K. Namely, we obtain

$$\bar{Q}_{1:K,\boldsymbol{\cdot}} \overset{\mathcal{E}}{\sim} I_K. \tag{E.25}$$

Step 4. In this step we prove $\bar{Q} \overset{\mathcal{E}}{\sim} Q$. Without loss of generality, we assume the columns of $\bar{Q}$ is arranged in the order $(b_1, b_2, \dots, b_K)$. Recall that $\mathcal{A} \subseteq \{0,1\}^K$ denotes the set of attribute patterns that respect the specified attribute hierarchy. For each $j \in \{K+1, \dots, J\}$, in the following two parts (i) and (ii), we first prove

$$\mathcal{A}_* := \{\boldsymbol{\alpha} \in \mathcal{A} : \boldsymbol{\alpha} \succeq \bar{\boldsymbol{q}}_j, \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_j\} = \varnothing$$

in (i); and then prove

$$\mathcal{A}_{**} := \{\boldsymbol{\alpha} \in \mathcal{A} : \boldsymbol{\alpha} \nsucceq \bar{\boldsymbol{q}}_j, \boldsymbol{\alpha} \succeq \boldsymbol{q}_j\} = \varnothing$$

in (ii). Together, these two conclusions would imply $\bar{\boldsymbol{q}}_j \overset{\mathcal{E}}{\sim} \boldsymbol{q}_j$.

(i) We use proof by contradiction and suppose $\mathcal{A}_* = \{\boldsymbol{\alpha} \in \mathcal{A} : \boldsymbol{\alpha} \succeq \bar{\boldsymbol{q}}_j, \boldsymbol{\alpha} \nsucceq \boldsymbol{q}_j\} \neq \varnothing$ for some $j \in \{K+1, \dots, J\}$. Then $\sum_{\boldsymbol{\alpha} \in \mathcal{A}_*} p_{\boldsymbol{\alpha}} > 0$. Define

$$\boldsymbol{\theta}^* = \sum_{k \le K:\, \bar{q}_{j,k}=1} \theta_k^- \boldsymbol{e}_k + \theta_j^+ \boldsymbol{e}_j, \tag{E.26}$$

then $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$ and hence $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}} \bar{\boldsymbol{p}} = 0$. Based on Step 2 and 3, we have $\bar{\theta}_j^+ = \theta_j^+$ and $\bar{\theta}_k^- = \theta_k^-$ for the j and any k with $\bar{q}_{j,k} = 1$ used in (E.26). Therefore, due to the first summation term in (E.26), $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha}$ satisfies $\alpha_k = 1$ for all k s.t. $\bar{q}_{j,k} = 1$ (i.e., $\boldsymbol{\alpha} \succeq \bar{\boldsymbol{q}}_j$); and due to the second term $\theta_j^+ \boldsymbol{e}_j$ in (E.26), $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\theta_{j,\boldsymbol{\alpha}} = \theta_j^-$ (i.e., $\boldsymbol{\alpha} \nsucceq \boldsymbol{q}_j$). In summary,

$T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \in \mathcal{A}_*$, so

$$T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = \prod_{k \leq K:\, q_{j,k}=1} (\theta_k^+ - \theta_k^-)(\theta_j^- - \theta_j^+)\Big(\sum_{\boldsymbol{\alpha} \in \mathcal{A}_*} p_{\boldsymbol{\alpha}}\Big) \neq 0,$$

which contradicts $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} = 0$. This contradiction means $\mathcal{A}_* = \varnothing$ must hold.

(ii) We also use proof by contradiction and suppose $\mathcal{A}_{**} = \{\boldsymbol{\alpha} \in \mathcal{A} : \boldsymbol{\alpha} \not\succeq \bar{\boldsymbol{q}}_j,\, \boldsymbol{\alpha} \succeq \boldsymbol{q}_j\} \neq \varnothing$ for some $j \in \{K+1, \ldots, J\}$. Then there exists $\boldsymbol{\alpha} \in \mathcal{A}$ with $\boldsymbol{\alpha} \succeq \boldsymbol{q}_j$ but $\boldsymbol{\alpha} \not\succeq \bar{\boldsymbol{q}}_j$, which implies there exists some attribute $k \in [K]$ s.t. $\bar{q}_{j,k} = 1$ and $q_{j,k} = 0$. Based on the above relation, we apply Lemma E.3 to obtain $\bar{\theta}_k^+ = \theta_k^+$. Define

$$\boldsymbol{\theta}^* = \bar{\theta}_j^- \boldsymbol{e}_j + \bar{\theta}_k^+ \boldsymbol{e}_k + \sum_{\substack{m \leq K: \\ k \not\to m}} \theta_m^- \boldsymbol{e}_m, \tag{E.27}$$

then based on the first two terms in (E.27), we have $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$. So $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} = 0$ and further $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = 0$. Now consider $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}}$, then $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha}$ belongs to the set $\mathcal{A}_7$ defined as

$$\mathcal{A}_7 = \{\boldsymbol{\alpha} \in \mathcal{A} : \alpha_k = 0;\ \alpha_m = 1\ \forall\ k \not\to m\}, \tag{E.28}$$

then this $\mathcal{A}_7 \neq \varnothing$ because the $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_K^*)$ defined as follows belongs to $\mathcal{A}_7$. The $\boldsymbol{\alpha}^*$ takes the form $\alpha_k^* = 0$, $\alpha_\ell^* = 0$ for all $k \to \ell$, and $\alpha_m^* = 1$ for all $k \not\to m$. The $\boldsymbol{\alpha}^*$ also satisfies $\boldsymbol{\alpha}^* \succeq \boldsymbol{q}_j$ for the following reason. Since $q_{j,k} = 0$, then under the attribute hierarchy this $\boldsymbol{q}_j$ is equivalent to a $\widehat{\boldsymbol{q}}_j$ with $\widehat{q}_{j,k} = 0$ and $\widehat{q}_{j,\ell} = 0$ for all ℓ s.t. $k \to \ell$. Therefore for the defined $\boldsymbol{\alpha}^* \in \mathcal{A}$ that respects the attribute hierarchy, there must be $\boldsymbol{\alpha}^* \succeq \widehat{\boldsymbol{q}}_j$ Since Lemma E.1 establishes that we can consider without loss of generality the case where each row vector of Q respects the attribute hierarchy, we have the conclusion that equivalently,

$\boldsymbol{\alpha}^* \succeq \boldsymbol{q}_j$. And further there is $\sum_{\boldsymbol{\alpha}\in\mathcal{A}_7} p_{\boldsymbol{\alpha}} \geq p_{\boldsymbol{\alpha}^*} > 0$. Now we have

$$0 = T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = (\theta_j^+ - \bar{\theta}_j^-)(\theta_k^- - \bar{\theta}_k^+) \prod_{\substack{m\leq K:\\ k\not\to m}} (\theta_m^+ - \theta_m^-)\Big(\sum_{\boldsymbol{\alpha}\in\mathcal{A}_7} p_{\boldsymbol{\alpha}}\Big),$$

which leads to a contradiction since each factor in the above term is nonzero. So we have proved the $\mathcal{A}_{**}$ defined earlier must also be an empty set.

As stated before, based on the (i) and (ii) shown above, we obtain $\bar{\boldsymbol{q}}_j \overset{\mathcal{E}}{\sim} \boldsymbol{q}_j$ for every item $j \in \{K+1, \ldots, J\}$.

In summary, by far we have obtained $\bar{\theta}_k^- = \theta_k^-$ for all $k \in [K]$, $\bar{\theta}_j^+ = \theta_j^+$ for all $j \in \{K+1, \ldots, J\}$, and $\bar{Q} \overset{\mathcal{E}}{\sim} Q$.

Step 5. We next show $\bar{\theta}_k^+ = \theta_k^+$ for all $k \in [K]$ and $\bar{\theta}_j^- = \theta_j^-$ for all $j \in \{K+1, \ldots, J\}$, and $\bar{\boldsymbol{p}} = \boldsymbol{p}$.

Step 5.1. In this step, we show $\bar{\theta}_k^+ = \theta_k^+$ for all $k \in [K]$. By Condition C, there exists some item $j > K$ s.t. $q_{j,k} = 1$, and we denote this item by j_k. Define

$$\boldsymbol{\theta}^* = \sum_{\substack{h\leq K:\\ h\neq k}} \bar{\theta}_h^- \boldsymbol{e}_h + \bar{\theta}_{j_k}^- \boldsymbol{e}_{j_k} + \sum_{\substack{j>K:\\ j\neq j_k}} \theta_j^- \boldsymbol{e}_j, \tag{E.29}$$

then $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ if and only if $\boldsymbol{\alpha} = \mathbf{1}_K$. This is because considering the the last term of summation in (E.29), we have $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \succeq \boldsymbol{q}_{\mathcal{J}}$ where $\mathcal{J} := \{K+1, \ldots, J\} \setminus \{j_k\}$; and by Condition C there is $\boldsymbol{q}_{\mathcal{J}} = \mathbf{1}_K$. Specifically,

$$T_{\boldsymbol{r}^*,\mathbf{1}_K} = \prod_{\substack{h\leq K:\\ h\neq k}} (\theta_h^+ - \bar{\theta}_h^-)(\theta_{j_k}^+ - \theta_{j_k}^-) \prod_{\substack{j>K:\\ j\neq j_k}} (\theta_j^+ - \theta_j^-),$$

and there is $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = T_{\boldsymbol{r}^*,\mathbf{1}_K} p_{\mathbf{1}_K} \neq 0$. So by (E.3) we have $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} \neq 0$. Further, the element $\bar{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}}$ could be potentially nonzero only if $\boldsymbol{\alpha} = \mathbf{1}_K$. This is because considering the first two terms ($\sum_{\substack{h\leq K:\\ h\neq k}} \bar{\theta}_h^- \boldsymbol{e}_h$ and $\bar{\theta}_{j_k}^- \boldsymbol{e}_{j_k}$) in $\boldsymbol{\theta}^*$ defined in (E.29), there is $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}} \neq 0$ only if $\boldsymbol{\alpha} \succeq \max(\max_{\substack{h\leq K:\\ h\neq k}} \bar{\boldsymbol{q}}_h, \bar{\boldsymbol{q}}_{j_k})$; and since $\bar{q}_{j_k,k} = 1$ there must be $\max(\max_{\substack{h\leq K:\\ h\neq k}} \bar{\boldsymbol{q}}_h, \bar{\boldsymbol{q}}_{j_k}) = \mathbf{1}_K$. Therefore we have

$$\bar{\theta}_k^+ = \frac{\overline{T}_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\cdot}}\bar{\boldsymbol{p}}}{\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}}} = \frac{T_{\boldsymbol{r}^*+\boldsymbol{e}_k,\boldsymbol{\cdot}}\boldsymbol{p}}{T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p}} = \theta_k^+.$$

Step 5.2. In this step we show $\bar{\theta}_j^- = \theta_j^-$ for all $j \in \{K+1, \ldots, J\}$. Consider an arbitrary $j > K$, then there exists an attribute k such that $q_{j,k} = 1$. Define $\boldsymbol{\theta}^* = \theta_k^+ \boldsymbol{e}_k$, and note that in Step 5.1 we obtained $\bar{\theta}_k^+ = \theta_k^+$ and in Step 3 we obtained $\bar{\theta}_k^- = \theta_k^-$. Then with this $\boldsymbol{\theta}^*$, there is

$$0 \neq (\theta_k^- - \theta_k^+)\Big(\sum_{\boldsymbol{\alpha}\in\mathcal{A}:\, \boldsymbol{\alpha}\nsucceq\boldsymbol{q}_k} p_{\boldsymbol{\alpha}}\Big) = T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = \overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}} = (\theta_k^- - \theta_k^+)\Big(\sum_{\boldsymbol{\alpha}\in\mathcal{A}:\, \boldsymbol{\alpha}\nsucceq\boldsymbol{q}_k} \bar{p}_{\boldsymbol{\alpha}}\Big),$$

and note that for any $\boldsymbol{\alpha} \nsucceq \boldsymbol{q}_k$, there must be $\boldsymbol{\alpha} \nsucceq \boldsymbol{q}_j$ since $q_{j,k} = 1$. Now consider the item j, we have

$$\bar{\theta}_j^- = \frac{\overline{T}_{\boldsymbol{r}^*+\boldsymbol{e}_j,\boldsymbol{\cdot}}\bar{\boldsymbol{p}}}{\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}}} = \frac{\overline{T}_{\boldsymbol{r}^*+\boldsymbol{e}_j,\boldsymbol{\cdot}}\boldsymbol{p}}{\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p}} = \theta_j^-.$$

Since j is arbitrary from $\{K+1, \ldots, J\}$, we have obtained $\bar{\theta}_j^- = \theta_j^-$ for all $j \in \{K+1, \ldots, J\}$.

Step 6. In this step we show that for $\Gamma(Q, \mathcal{A})$ and the alternative Γ-matrix $\overline{\Gamma}$ (also denoted by $\Gamma(\bar{Q}, \bar{\mathcal{A}})$ where $\bar{\mathcal{A}}$ is the set corresponding to those columns in $\overline{\Gamma}$ with nonzero proportion parameters in $\bar{\boldsymbol{p}}$), the column vectors in $\Gamma(\bar{Q}, \bar{\mathcal{A}})$ that correspond to $\bar{p}_{\boldsymbol{\alpha}} > 0$ are identical to $\Gamma(Q, \mathcal{A})$; furthermore, $\bar{p}_{\pi(\boldsymbol{\alpha})} = p_{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha} \in \mathcal{A}$, where $\pi : \mathcal{A} \to \bar{\mathcal{A}}$ is a one-to-one map. For an arbitrary $\boldsymbol{\alpha} \in \mathcal{A}$, define

$$\boldsymbol{\theta}^* = \sum_{\substack{k\leq K:\\ \boldsymbol{\alpha}\succeq\boldsymbol{q}_k}} \theta_k^- \boldsymbol{e}_k + \sum_{\substack{m\leq K:\\ \boldsymbol{\alpha}\nsucceq\boldsymbol{q}_m}} \theta_m^+ \boldsymbol{e}_m. \tag{E.30}$$

Then for any $\boldsymbol{\alpha}^* \in \mathcal{A}$, the $T_{\boldsymbol{r}^*,\boldsymbol{\alpha}^*} \neq 0$ (equivalently, $\overline{T}_{\boldsymbol{r}^*,\boldsymbol{\alpha}^*} \neq 0$) if and only if $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}$, because $Q^0 = Q_{1:K,\boldsymbol{\cdot}} \overset{\mathcal{E}}{\sim} I_K$. Then $T_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\boldsymbol{p} = \overline{T}_{\boldsymbol{r}^*,\boldsymbol{\cdot}}\bar{\boldsymbol{p}}$ gives

$$\prod_{\substack{k\leq K:\\ \alpha_k=1}}(\theta_k^+ - \theta_k^-)\prod_{\substack{m\leq K:\\ \alpha_m=0}}(\theta_m^- - \theta_m^+)p_{\boldsymbol{\alpha}} = \prod_{\substack{k\leq K:\\ \alpha_k=1}}(\theta_k^+ - \theta_k^-)\prod_{\substack{m\leq K:\\ \alpha_m=0}}(\theta_m^- - \theta_m^+)\bar{p}_{\pi(\boldsymbol{\alpha})},$$

and we obtain $\bar{p}_{\pi(\boldsymbol{\alpha})} = p_{\boldsymbol{\alpha}}$, Since $\sum_{\boldsymbol{\alpha}\in\{0,1\}^K}\bar{p}_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in\mathcal{A}}p_{\boldsymbol{\alpha}} = 1$, the equality $\bar{p}_{\boldsymbol{\alpha}} = p_{\boldsymbol{\alpha}}$ for any $\boldsymbol{\alpha}\in\mathcal{A}$ also implies $\bar{p}_{\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha}\in\{0,1\}^K\setminus\mathcal{A}$. So $\Gamma(\bar{Q},\bar{\mathcal{A}}) = \Gamma(Q,\mathcal{A})$ also holds. This completes the proof of Step 6.

Now we have shown $\Gamma(Q,\mathcal{A}) = \Gamma(\bar{Q},\bar{\mathcal{A}})$, $\bar{\boldsymbol{\theta}}^+ = \boldsymbol{\theta}^+$, $\bar{\boldsymbol{g}} = \boldsymbol{g}$, $\bar{\boldsymbol{p}} = \boldsymbol{p}$. This completes the proof of the sufficiency of Conditions A, B and C.

We next show that Condition A is necessary for identifying $(\Gamma(Q,\mathcal{A}), \boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$. We use proof by contradiction and assume that Condition A does not hold. Recall that the type of modification of Q described in Condition A is "Operation" A, which sets every $q_{j,k}$ to zero if $q_{j,h} = 1$ and $k \to h$. Denote the resulting matrix by Q^A. If Condition A fails to hold, then Q^A lacks an identity submatrix I_K. Without loss of generality, suppose Q^A does not contain any row vector in the form $\boldsymbol{e}_h$ for some $h\in[K]$. Combined with the definition of Operation A, this means for any $\boldsymbol{q}$-vector with $q_{j,h} = 1$, in the original Q there must be $q_{j,\ell} = 1$ for some $\ell \not\to h$. Then the following two attribute patterns in $\mathcal{A}$ will lead to the same column vectors in $\Gamma(Q,\mathcal{A})$: $\boldsymbol{\alpha}_1 := \mathbf{0}_K$ and $\boldsymbol{\alpha}_2 := (\alpha_{2,1},\ldots,\alpha_{2,K})$ where $\alpha_{2,h} = 1$, $\alpha_{2,k} = 1$ for all $k\to h$, and $\alpha_{2,\ell} = 0$ for all $\ell\not\to h$. The fact that $\Gamma_{:,\boldsymbol{\alpha}_1} = \Gamma_{:,\boldsymbol{\alpha}_2}$ directly results in that $p_{\boldsymbol{\alpha}_1}$ and $p_{\boldsymbol{\alpha}_2}$ can be at best identified up to their sum, even if all the item parameters $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$ are identified and known. This proves the necessity of Condition A.

As for the last claim in part (i) that Conditions A, B and C are necessary and sufficient for identifiability of $(Q,\boldsymbol{p},\boldsymbol{\theta}^+,\boldsymbol{\theta}^-)$ where there is no hierarchy (i.e., $p_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha}\in\{0,1\}^K$), it directly follows from the result in Theorem 1 in Gu and Xu (2020b).

Proof of part (ii). *We first show that if Q contains a submatrix I_K other than satisfying A, B and C, then $(\mathcal{A}, \boldsymbol{\theta}^+, \boldsymbol{\theta}^-, \boldsymbol{p})$ are jointly identifiable.* Based on the conclusion of part (i), it suffices to show that if Q contains an I_K, then $\mathcal{A}$ are identifiable from $\Gamma(Q, \mathcal{A})$. That is, we will show that if $\Gamma(Q, \mathcal{A}) = \Gamma(\bar{Q}, \bar{\mathcal{A}})$ with both Q and $\bar{Q}$ containing a submatrix I_K, then $\mathcal{A} = \bar{\mathcal{A}}$. Note that when Q contains an I_K, the $J \times 2^K$ matrix $\Gamma(Q, \{0,1\}^K)$ has 2^K distinct column vectors Gu and Xu (2020a). Without loss of generality, suppose the first K rows of Q and $\bar{Q}$ are both I_K. Then $\Gamma_{1:K,:}(Q, \mathcal{A}) = \Gamma_{1:K,:}(\bar{Q}, \bar{\mathcal{A}})$ exactly implies $\mathcal{A} = \bar{\mathcal{A}}$, due to this distinctiveness of the 2^K ideal response vectors of the 2^K latent patterns under an identity matrix.

We next show that in order to identify an arbitrary $\mathcal{A}$, it is necessary for Q to contain an I_K. Suppose Q does not contain an I_K, then based on the concept of $\boldsymbol{p}$-partial identifiability in Gu and Xu (2020a), certain patterns would become equivalent in that they lead to the same column vectors in $\Gamma(Q, \{0,1\}^K)$, hence there must exist some $\mathcal{A}$ that is not identifiable. □

E.2 Computational details for Chapter VI

E.2.1 Details of Extended Bayesian Information Criterion (EBIC)

Consider the objective function (6.6). For a $\lambda \in (-\infty, 0)$, denote the estimated set of patterns by $\mathcal{A}_\lambda = \{\boldsymbol{\alpha} \in \mathcal{A}_{\text{candi}} : \widehat{p}_{\boldsymbol{\alpha}} > \rho_N,\ (\widehat{\boldsymbol{\Theta}}, \widehat{\boldsymbol{p}}) = \arg\max_{\boldsymbol{\Theta},\boldsymbol{p}} \ell_\lambda^{\text{2nd}}(\boldsymbol{\Theta}, \boldsymbol{p})\}$. Here $\rho_N > 0$ is the threshold for selecting latent patterns, and we take a sample size dependent $\rho_N = 1/(2N)$ in all the experiments. The EBIC proposed in Chen and Chen (2008) has the form

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

where a larger EBIC parameter $\gamma \in [0,1]$ would encourage a more parsimonious model (i.e., fewer selected latent patterns). We take $\gamma = 1$ for the greatest amount of parsimony. Then $\mathcal{A}_{\text{final}}$ is taken to be the particular $\mathcal{A}_\lambda$ that achieves the smallest EBIC value.

E.2.2 Structure learning with missing data and binary matrix completion

When there exists missing data in a HLAM, what is observed is not a complete $N \times J$ binary matrix $\boldsymbol{R}$; instead, it is $\boldsymbol{R}$ with missing entries. Denote by $\Omega \subseteq [N]\times[J]$ the set of indices of the observed entries. Then the original objective functions presented in (6.5) and (6.6) should be replaced by the following two objective functions $\ell^{\text{1st},\Omega}$ and $\ell_\lambda^{\text{2nd},\Omega}$, respectively:

$$
\begin{aligned}
\ell^{\text{1st},\Omega}(Q,\, \mathbf{A},\, \boldsymbol{\Theta} \mid \boldsymbol{R}) = &\sum_{(i,j)\in\Omega} \Big[r_{i,j} \Big(\prod_k a_{i,k}^{q_{j,k}} \log \theta_j^+ + (1 - \prod_k a_{i,k}^{q_{j,k}}) \log \theta_j^- \Big) \\
&+ (1 - r_{i,j}) \Big(\prod_k a_{i,k}^{q_{j,k}} \log(1-\theta_j^+) + (1 - \prod_k a_{i,k}^{q_{j,k}}) \log(1-\theta_j^-) \Big) \Big];\\
\ell_\lambda^{\text{2nd},\Omega}(\boldsymbol{p},\, \boldsymbol{\Theta} \mid \boldsymbol{R},\, \widehat{Q}) = &\sum_{i=1}^N \log \Big\{ \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{candi}}} p_{\boldsymbol{\alpha}} \prod_{j:\,(i,j)\in\Omega} \theta_{j,\boldsymbol{\alpha}}^{r_{i,j}} (1-\theta_{j,\boldsymbol{\alpha}})^{1-r_{i,j}} \Big\} + \\
&\lambda \sum_{\boldsymbol{\alpha}\in\mathcal{A}_{\text{candi}}} \log_\rho(p_{\boldsymbol{\alpha}}).
\end{aligned} \tag{E.31}
$$

With missing values in $\boldsymbol{R}$, the ADG-EM Algorithm 4 and the PEM algorithm in Gu and Xu (2019a) should be replaced by the following Algorithm 5 and Algorithm 6.

The algorithms for structure learning with missing data can be also used for binary matrix completion. Note that even if not all entries of $\boldsymbol{R}$ are observed, the entire $N\times K$ matrix $\mathbf{A}$ and the entire $J \times K$ matrix Q can be estimated from Algorithms 5-6, as long as there are at least *some* observed entries in each row and each column of $\boldsymbol{R}$. Then naturally, based on the estimated $\mathbf{A}$ and Q, a complete $N \times J$ matrix $\widehat{\boldsymbol{R}} = (\widehat{r}_{i,j})$

Algorithm 5: ADG-EM with missing data: Q estimation and dimension reduction

Data: Responses $\boldsymbol{R}$ with the set of indices of observed entries $\Omega \subseteq [N] \times [J]$.
Initialize attribute patterns $(a_{i,k})_{N\times K} \in \{0,1\}^{N\times K}$; and structural matrix $(q_{j,k})_{J\times K} \in \{0,1\}^{J\times K}$.
Initialize parameters $\boldsymbol{\theta}^+$ and $\boldsymbol{\theta}^-$. Set $t=1$, $\mathbf{A}^{\text{ave}} = \mathbf{0}$.
while *not converged* **do**
  for $(i,j) \in \Omega$ **do** $\psi_{i,j} \leftarrow r_{i,j} \log[\theta_j^+/\theta_j^-] + (1-r_{i,j}) \log[(1-\theta_j^+)/(1-\theta_j^-)]$;
  $\mathbf{A}^{\text{s}} \leftarrow \mathbf{0}$, $Q^{\text{s}} \leftarrow \mathbf{0}$.
  for $r \in [2M]$ **do**
    for $(i,k) \in [N] \times [K]$ **do**
      Draw $a_{i,k} \sim \text{Bernoulli}\Big(\sigma\Big(\sum_{j:\,(i,j)\in\Omega} q_{j,k} \prod_{m\neq k} a_{i,m}^{q_{j,m}} \psi_{i,j}\Big)\Big)$;
    if $r > M$ **then** $\mathbf{A}^{\text{s}} \leftarrow \mathbf{A}^{\text{s}} + \mathbf{A}$;
  $\mathbf{A}^{\text{ave}} \leftarrow \frac{1}{t}\mathbf{A}^{\text{s}}/M + \Big(1-\frac{1}{t}\Big)\mathbf{A}^{\text{ave}}$; $t \leftarrow t+1$.
  for $r \in [2M]$ **do**
    for $(j,k) \in [J] \times [K]$ **do**
      Draw $q_{j,k} \sim \text{Bernoulli}\Big(\sigma\Big(-\sum_{i:\,(i,j)\in\Omega}(1-a_{i,k}) \prod_{m\neq k} a_{i,m}^{q_{j,m}} \psi_{i,j}\Big)\Big)$;
    if $r > M$ **then** $Q^{\text{s}} \leftarrow Q^{\text{s}} + Q$;
  $Q = I(Q^{\text{s}}/M > \frac{1}{2})$ element-wisely; $\mathbf{I}^{\text{ave}} = \Big(\prod_k \{a_{i,k}^{\text{ave}}\}^{q_{j,k}}\Big)_{N\times J}$;
  for $j \in [J]$ **do**

$$\theta_j^+ \leftarrow \frac{\sum_{i:\,(i,j)\in\Omega} r_{i,j} I_{i,j}^{\text{ave}}}{\sum_{i:\,(i,j)\in\Omega} I_{i,j}^{\text{ave}}}, \quad \theta_j^- \leftarrow \frac{\sum_{i:\,(i,j)\in\Omega} r_{i,j}(1-I_{i,j}^{\text{ave}})}{\sum_{i:\,(i,j)\in\Omega}(1-I_{i,j}^{\text{ave}})};$$

$\widehat{\mathbf{A}} = I(\mathbf{A}^{\text{ave}} > \frac{1}{2})$ element-wisely.
Output: $\mathcal{A}_{\text{candi}}$, which includes the unique row vectors of $\widehat{\mathbf{A}}$, and binary structural matrix $\widehat{Q}$.
Then $(\widehat{Q}, \mathcal{A}_{\text{candi}})$ are fed to the PEM with missing data to maximize (6.6) and obtain $\mathcal{A}_{\text{final}}$.

with no missing entries can be reconstructed, by setting $\widehat{r}_{i,j}$ equal to

$$\widehat{r}_{i,j} := I\left(\widehat{\theta}_j^+ \cdot \Gamma_{\widehat{\boldsymbol{q}}_j,\widehat{\boldsymbol{a}}_i} + \widehat{\theta}_j^- \cdot (1-\Gamma_{\widehat{\boldsymbol{q}}_j,\widehat{\boldsymbol{a}}_i}) > \frac{1}{2}\right), \quad (i,j) \in [N]\times[J]. \tag{E.32}$$

which is the integer (0 or 1) nearest to the posterior mean of (i,j)th entry of $\boldsymbol{R}$.

Algorithm 6: PEM with missing data: Penalized EM for log-penalty with $\lambda \in (-\infty, 0)$

Data: Responses $\boldsymbol{R}$ with the set of indices of observed entries $\Omega \in [N] \times [J]$, and candidate attribute patterns $\mathcal{A}_{\text{candi}}$.

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

while *not converged* **do**

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

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

  for $l \in [|\mathcal{A}_{\text{candi}}|]$ **do** $\Delta_l = \max\{c,\ \lambda + \sum_{i=1}^N \varphi_{i,\boldsymbol{\alpha}_l}\}$; ($c > 0$ is pre-specified) ;

  $\boldsymbol{p} \leftarrow \boldsymbol{\Delta}/(\sum_l \Delta_l)$;

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

$$\theta_j^+ = \frac{\sum_{i:\,(i,j)\in\Omega}\sum_{\boldsymbol{\alpha}_l} R_{i,j}\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}_l}\varphi_{i,\boldsymbol{\alpha}_l}}{\sum_{i:\,(i,j)\in\Omega}\sum_{\boldsymbol{\alpha}_l}\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}_l}\varphi_{i,\boldsymbol{\alpha}_l}}, \quad \theta_j^- = \frac{\sum_{i:\,(i,j)\in\Omega}\sum_{\boldsymbol{\alpha}_l} R_{i,j}(1-\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}_l})\varphi_{i,\boldsymbol{\alpha}}}{\sum_{i:\,(i,j)\in\Omega}\sum_{\boldsymbol{\alpha}_l}(1-\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}_l})\varphi_{i,\boldsymbol{\alpha}_l}}.$$

Output: $\{\boldsymbol{\alpha}_l \in \mathcal{A}_{\text{candi}}:\ p_{\boldsymbol{\alpha}_l} > \rho\}$.

E.2.3 More details on the experiments in Section 5 of the main text

Denote the output of the ADG-EM Algorithm 4 by $(\widehat{Q}, \mathcal{A}_{\text{candi}})$. The definition of $\text{acc}[Q]^{\mathcal{A}}$ in Table 6.1 is

$$\text{acc}[Q]^{\mathcal{A}} = \frac{1}{J|\mathcal{A}|}\sum_{(j,\boldsymbol{\alpha})\in[J]\times\mathcal{A}} I(\Gamma_{\widehat{\boldsymbol{q}}_j,\boldsymbol{\alpha}} \neq \Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}}). \tag{E.33}$$

That is, since an $\mathcal{A}$ gives an equivalence class of Q-matrices, the accuracy of estimating Q should be evaluated by the accuracy of estimating the ideal response structure $\{\Gamma_{\boldsymbol{q}_j,\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathcal{A}\}$ under the true $\mathcal{A}$. More specifically, as long as the $J \times |\mathcal{A}|$ ideal response matrix is estimated accurately, the Q is already identified in the correct equivalence class.

We next present the statistical variations of the results in Table 6.1 of the main text. We choose to present the inter quartile range (i.e., difference between the 75th and the 25th percentiles) as a measure of variation as it is more robust to outliers. See Table E.1 for details.

Table E.1: Statistical variations of results presented in Table 6.1 in the main text. The "IQR" stands for the inter quartile range of the three accuracy measures: $\text{acc}[Q]^{\mathcal{A}}$, TPR, and $1-$FDR and that of the size of $\mathcal{A}_{\text{final}}$. All results are based on 200 runs. In each scenario, the IQR is presented below the original number of the accuracy or the size of $\mathcal{A}_{\text{final}}$ in parenthesis.

2^K	$\lvert\mathcal{A}_0\rvert$	noise	$(N,J)=(1200,120)$				$(N,J)=(1200,1200)$			
			$\text{acc}[Q]^{\mathcal{A}}$	TPR	1-FDR	$\lvert\mathcal{A}_{\text{final}}\rvert$	$\text{acc}[Q]^{\mathcal{A}}$	TPR	1-FDR	$\lvert\mathcal{A}_{\text{final}}\rvert$
2^8	10	20%	1.00	1.00	0.96	10	1.00	1.00	1.00	10
		IQR	(0.000)	(0.000)	(0.091)	(1)	(0.000)	(0.000)	(0.000)	(0)
		30%	1.00	1.00	0.96	10	1.00	1.00	0.68	15
		IQR	(0.000)	(0.000)	(0.091)	(1)	(0.000)	(0.000)	(0.089)	(2)
	15	20%	1.00	1.00	0.95	15	1.00	1.00	1.00	15
		IQR	(0.000)	(0.000)	(0.063)	(1)	(0.000)	(0.000)	(0.000)	(0)
		30%	1.00	0.99	0.94	16	1.00	1.00	0.80	19
		IQR	(0.003)	(0.000)	(0.118)	(2)	(0.001)	(0.000)	(0.132)	(3)
2^{15}	10	20%	0.98	0.91	0.90	10	0.99	0.99	0.97	10
		IQR	(0.002)	(0.000)	(0.000)	(0)	(0.000)	(0.000)	(0.000)	(0)
		30%	0.99	1.00	0.88	10	0.97	0.94	0.62	15
		IQR	(0.007)	(0.000)	(0.000)	(0)	(0.000)	(0.000)	(0.126)	(2)
	15	20%	0.99	0.96	0.95	15	1.00	1.00	0.99	15
		IQR	(0.001)	(0.000)	(0.000)	(0)	(0.001)	(0.000)	(0.000)	(0)
		30%	0.99	0.99	0.89	15	0.99	0.98	0.71	21
		IQR	(0.006)	(0.000)	(0.063)	(1)	(0.000)	(0.000)	(0.123)	(3.5)

As illustrated in Example VI.2, a set of allowed patterns $\mathcal{A} \subseteq \{0,1\}^K$ would give an equivalence class of Q-matrices, each of which would lead to identical $\Gamma(Q,\mathcal{A})$. More generally, the structural matrix Q and the attribute pattern matrix $\mathbf{A}$ are coupled together, such that fixing one of them would allow identifying the other up to an equivalence class. However, there is indeed a way to uniquely determine an $\mathcal{A}$ from $\Gamma(Q,\mathbf{A})$ if we impose one constraint on the structural matrix Q: to require Q to contain a submatrix I_K. This fact is shown in the part (ii) of Theorem VI.1. In particular, if for instance we constrain $Q_{1:K,} = I_K$, then $\mathcal{A}$ is uniquely determined from the $K \times |\mathcal{A}|$ ideal response matrix $\Gamma(Q_{1:K,},\mathcal{A})$. As discussed in the main text after Theorem VI.1, this phenomenon is analogous to the identification criteria for the factor loading matrix in factor analysis models, where the loading matrix is often required to include an identity submatrix or satisfy certain rank constraints (Anderson, 2009; Bai and Li, 2012). Therefore at least theoretically, in estimation, one needs to enforce the constraint that Q contains a submatrix I_K to uniquely determine $\mathbf{A}$ and the attribute hierarchy. In particular, in the simulation scenario $(N, J, K) = (1200, 120, 8)$ in Table 6.1, we enforce such a constraint after running the ADG-EM algorithm, and then use the $\mathcal{A}_{\text{candi}}$ and the constrained $\widehat{Q}$ as input to the second stage PEM algorithm. Interestingly, we observe that practically, this constraint is not needed when estimating more large-scale problems. For instance, for all the other simulation scenarios in Table 6.1 and Figure 6.3 with $(N, J) = (1200, 1200)$ or $K = 15$, we directly run PEM without such a constraint and the structure learning results presented there are indeed accurate.

For the experiment with the Austrian TIMSS 2011 real data that has $(N, J, K) = (1010, 47, 9)$, a tentative Q-matrix is provided in the `R` package `CDM`. This tentative Q has all the row vectors being unit vectors, i.e., for each $j \in \{1, \ldots, 47\}$, there is $\boldsymbol{q}_j = \boldsymbol{e}_k$ for some $k \in [K]$. To learn the attribute hierarchy from this dataset, we run Algorithm 5 and Algorithm 6 presented in Section B.2 of this supplementary material to handle the missing entries in $\boldsymbol{R}$. In particular, we use the tentative Q provided in the `R` package as the initial value for Algorithm 5 and enforce $\widehat{Q}$ to contain an identify matrix after running Algorithm 5 to obtain $\widehat{Q}$ and $\mathcal{A}_{\text{candi}}$. Finally, with the $\mathcal{A}_{\text{candi}}$ and the enforced $\widehat{Q}$ as input, we run Algorithm 6 and obtain the results presented in Figure 6.4 in the main text.

E.3 Large noisy binary matrix factorization/reconstruction

A nice byproduct of the proposed ADG-EM Algorithm 4 is a scalable algorithm for large-scale noisy binary matrix factorization/reconstruction and latent structure estimation. As discussed in the previous section, if there exists attribute hierarchy, the

Q-matrix can be identified up to the equivalence class determined by $\mathcal{A}$. The emphasis of the experiments in the main text is to estimate $\mathcal{A}$; while in the current experiments, we focus on the estimation of the structural matrix Q, and the reconstruction of the ideal response matrix based on the estimated Q.

We use Algorithm 4 to decompose the $N \times J$ large noisy binary matrix $\boldsymbol{R}$ generated under the AND-model. Specifically, $\boldsymbol{R} \approx \mathbf{A} \circ Q^\top$, where the "$\circ$" denotes the "AND" logical operation between each pair of $\boldsymbol{a}_i = (a_{i,1}, \ldots, a_{i,K})$ and $\boldsymbol{q}_j = (q_{j,1}, \ldots, q_{j,K})$, as introduced in (6.1); while the "$\approx$" allows for item level noises as quantified by $1 - \theta_j^+$ and θ_j^- in (6.2). In matrix form, we have

$$\begin{pmatrix} r_{1,1} & \cdots & \cdots & r_{1,J} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ r_{N,1} & \cdots & \cdots & r_{N,J} \end{pmatrix} \approx \begin{pmatrix} a_{1,1} & \vdots & a_{1,K} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ a_{N,1} & \vdots & a_{N,K} \end{pmatrix} \circ \begin{pmatrix} q_{1,1} & \cdots & \cdots & q_{J,1} \\ \cdots & \cdots & \cdots & \cdots \\ q_{1,K} & \cdots & \cdots & q_{J,K} \end{pmatrix}. \tag{E.34}$$

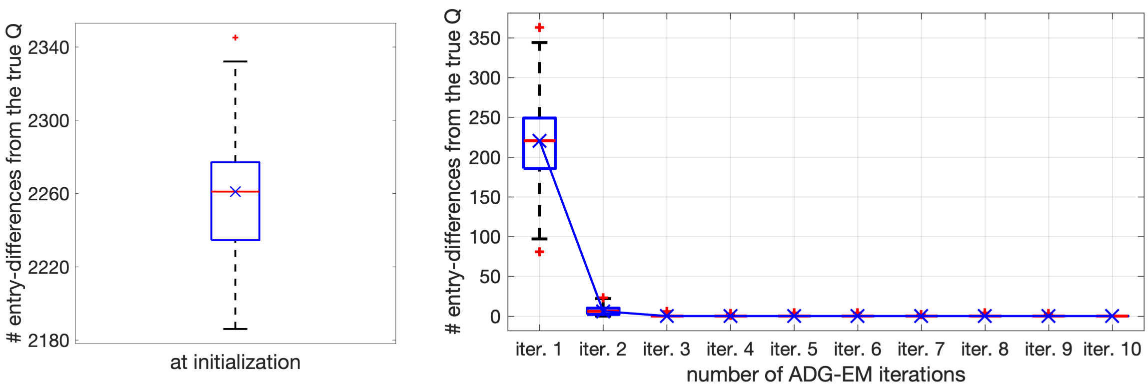
We perform simulations in the scenario $(N, J, K) = (1000, 1000, 7)$, where the true Q vertically stacks $J/(2K)$ copies of submatrix I_K and the remaining $J/2$ rows of Q vertically stacks an appropriate number of another $K \times K$ submatrix Q^2 in the following form,

$$Q^2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{7\times 7}. \tag{E.35}$$

The ground truth 1000×7 matrix Q is visualized in the bottom-right plot in Figure E.2, with color yellow representing value "1" and color blue representing value "0".

Then under this structural matrix Q, we generate response data $\boldsymbol{R}$ using noise parameters $1 - \theta_j^+ = \theta_j^- = 30\%$, and proportion parameters $p_{\boldsymbol{\alpha}} = 1/128$ for all $\boldsymbol{\alpha} \in \{0,1\}^7$. In the current scenario we would like to keep track of the estimation accuracy of Q itself, so we set $\mathcal{A} = \{0,1\}^K$ to be saturated such that $[Q]^{\mathcal{A}}$ contains only one element: Q itself. For each of 200 simulated datasets, we apply our ADG-EM Algorithm 4 alone to estimate Q and reconstruct the ideal case $\boldsymbol{R}$ using expression (E.32). The initializations $\{Q_{\text{ini}}\}$'s are obtained from randomly perturbing about one third entries in the true Q in each run. Instead of specifying a stopping criterion based on the convergence of the objective function, in the current experiment we just run exactly 10 stochastic EM iterations in Algorithm 4; we record the number of entry-differences between the estimated Q and the true Q_{true} along each EM iteration, and present the corresponding boxplot in Figure E.1(b). In addition, we record the number of entry-differences between Q_{true} and the initial value Q_{ini}, which is given as input to the algorithm, and present the boxplot based on 200 runs in Figure E.1(a).

The two boxplots in Figure E.1 show the superior convergence performance of the proposed ADG-EM algorithm. For each boxplot, the central mark denotes the median, and the bottom and top edges of the box are the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points that are not considered outliers, and the outliers are plotted with the '+'. Out of the $1000 \times 7 = 7000$ entries in the structural matrix, although the initialization of Q differs from the true one by more than 2000 entries on average, after just one stochastic EM iteration, the number of entry-differences between $Q_{\text{iter. 1}}$ and Q_{true} decreases to less than 300 entries in most cases. After just 3 stochastic EM iterations, for a vast majority of the 200 datasets, the Q_{true} is perfectly recovered and remains unchanged in further iterations of the algorithm. Indeed, after 10 iterations, for each of the 200 datasets, the Q_{true} is exactly recovered! Considering the relatively high noise rate 30% in $\boldsymbol{R}$ (i.e., 30% of the entries in the ideal response matrix are randomly flipped to form the observed $\boldsymbol{R}$) and the suboptimal initializations, such performance on latent structure estimation is impressive.



(a) # entry-differences between Q_{ini} and Q_{true}

(b) # entry-differences between Q_{iter} and Q_{true}

Figure E.1: Boxplots of entry-differences between estimated Q_{iter} (or Q_{ini}) and the true structural matrix Q_{true}, with size 1000×7. Results are based on 200 independent runs.

To obtain a better understanding of the performance of the ADG-EM algorithm, we next present two specific examples that visualize the intermediate results of the algorithm. Still in the setting described above, we simulate a 1000×1000 matrix $\boldsymbol{R}$ with noise rate $30\% = \theta_j^- = 1 - \theta_j^+$. In the first example, we use *randomly perturbed initialization for* $(Q, \mathbf{A})$, which is the same setting as the 200 runs behind Figure E.1. We present the results of Algorithm 4 together with its intermediate results along the first 4 iterations of the stochastic EM steps in Figure E.2. The 6 plots in the first row of Figure E.2 show the reconstruction of the data matrix $\boldsymbol{R}$, and the 6 plots in the second row of Figure E.2 show the estimation of the structural matrix Q. Specifically, after the t-th iteration, based on the $\widehat{Q}_{\text{iter. } t}$, the $\widehat{\boldsymbol{R}}_{\text{iter. } t}$ is reconstructed following Equation (E.32). The ground truth for $\boldsymbol{R}$ is just the $N \times J$ ideal response matrix in the noiseless case $\boldsymbol{R}_{\text{true}} = (r_{i,j}^{\text{true}})$, where $r_{i,j}^{\text{true}} = \Gamma_{\boldsymbol{q}_j, \boldsymbol{a}_i} = \prod_{k=1}^{K} a_{i,k}^{q_{j,k}}$. Along the first 3 stochastic EM iterations, the matrix Q change 2246, 275, 11 entries, respectively. Then from the 4th iteration to the 14th iteration when the stopping criterion is reached, we observe that all the entries of Q remain the same during the sampling in the E step. In the last several iterations the item parameters $(\boldsymbol{\theta}^+, \boldsymbol{\theta}^-)$ continued to change slightly and converge. Let $(r_{i,j}^{\text{observe}})$ and $(r_{i,j}^{\text{recons}})$ denote the observed noisy data matrix and the reconstructed data matrix in the end of the algorithm, respectively. Corresponding to the trial in Figure E.2, there is

$$\frac{1}{NJ}\sum_{(i,j)\in[N]\times[J]} I(r_{i,j}^{\text{true}} \neq r_{i,j}^{\text{observe}}) = 0.2995,$$
$$\frac{1}{NJ}\sum_{(i,j)\in[N]\times[J]} I(r_{i,j}^{\text{true}} \neq r_{i,j}^{\text{recons}}) = 5.21\times 10^{-5}.$$

In the above display, the 0.2995 reflects the noise rate in the observed data matrix corresponding to $1-\theta_j^+ = \theta_j^- = 0.3$ for each $j\in[J]$; and the 5.21×10^{-5} represents the error rate of reconstructing the $N\times J$ ideal response matrix, which is far smaller than the initial noise rate by several magnitudes. Indeed, there is no discernible difference between $\boldsymbol{R}_{\text{iter. 4}}$ and $\boldsymbol{R}_{\text{true}}$ based on the two rightmost plots in the first row of Figure E.2.

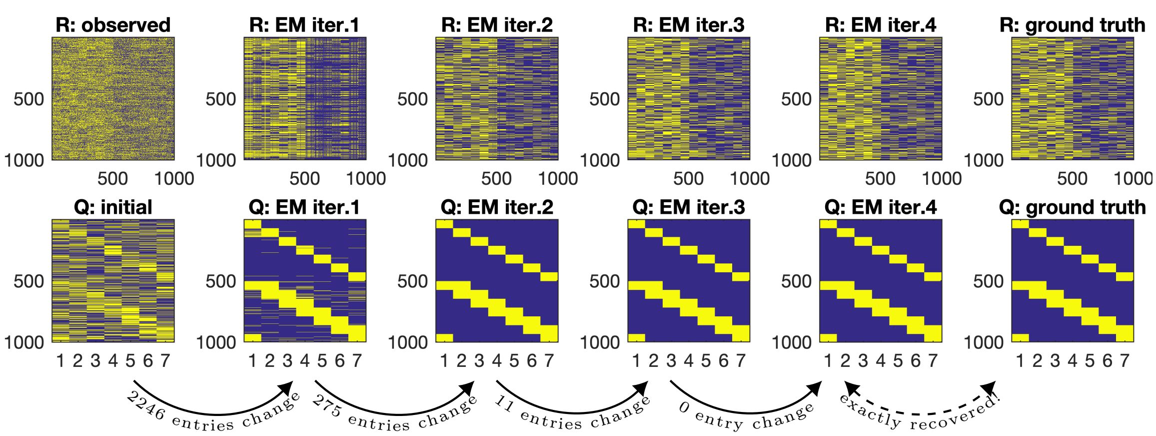


Figure E.2: Noisy binary matrix factorization and reconstruction with *randomly perturbed initialization.* Color yellow represents value "1" and color blue represents value "0". Only 3 stochastic EM iterations suffice for perfect estimation of the structural matrix Q.

In the second visualization example, we use *entirely random initialization* to obtain the $(Q_{\text{ini}}, \mathbf{A}_{\text{ini}})$ as input to Algorithm 4. The results of Algorithm 4 together

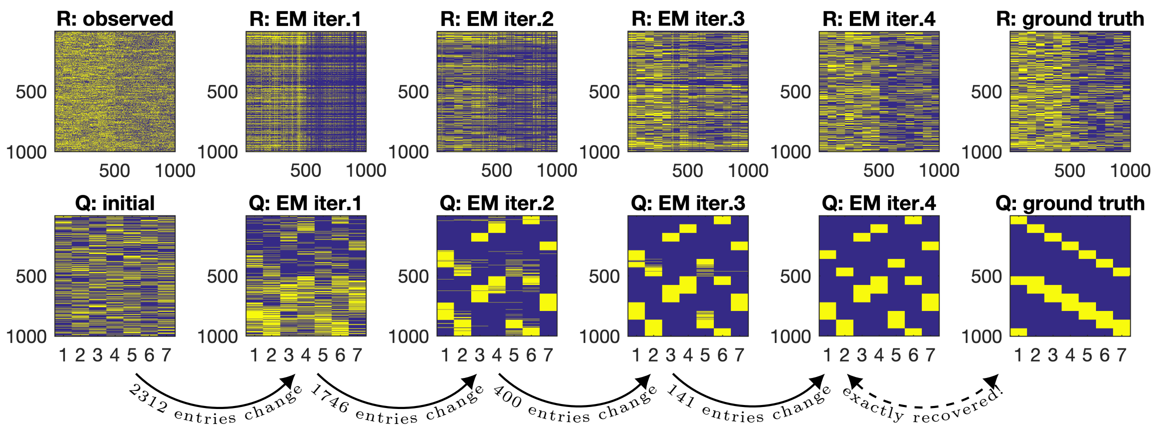


Figure E.3: Noisy binary matrix factorization and reconstruction with *entirely random initialization*. Color yellow represents value "1" and color blue represents value "0". Only 4 stochastic EM iterations of the proposed ADG-EM Algorithm 4 suffice for almost perfect decomposition and reconstruction. The stochastic Q after 4 iterations is identical to the true Q after column permutation.

with its intermediate results along the first 4 iterations of the stochastic EM steps are presented in Figure E.3. Along the first 4 stochastic EM iterations, the matrix Q changed 2312, 1746, 400, 141 entries, respectively. Then from the 5th iteration to the 18th iteration when the stopping criterion is reached, all the entries of Q remain the same during the sampling in the E step. With this entirely random initialization mechanism, we observe that the ADG-EM algorithm is not trapped in some sub-optimal local optimum; instead, the finally obtained $\widehat{Q}$ only differs from Q_{true} by a column permutation. This column permutation is the inevitable and trivial ambiguity associated with a latent attribute model with a structural matrix (Chen et al., 2015). The proposed ADG-EM algorithm also succeeds in this scenario. For Figure E.3, the reconstruction result for the data matrix $\boldsymbol{R}$ with noise rate 0.3 is

$$\frac{1}{NJ}\sum_{(i,j)\in[N]\times[J]} I(r_{i,j}^{\text{true}} \neq r_{i,j}^{\text{recons}}) = 7.20\times 10^{-5}.$$

One can also see from the above high reconstruction accuracy that estimating Q up to a column permutation does not compromise the reconstruction of $\boldsymbol{R}$ at all and that the reconstruction error is still at the magnitude of 10^{-5}.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

Anandkumar, A., Chaudhuri, K., Hsu, D., Kakade, S. M., Song, L., and Zhang, T. (2011). Spectral methods for learning multivariate latent tree structure. In *Advances in Neural Information Processing Systems*, pages 2025–2033.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832.

Anandkumar, A., Hsu, D., Janzamin, M., and Kakade, S. (2015). When are overcomplete topic models identifiable? Uniqueness of tensor tucker decompositions with structured sparsity. *Journal of Machine Learning Research*, 16:2643–2694.

Anderson, T. (2009). *An introduction to multivariate statistical analysis, 3rd Ed.* Wiley India Pvt. Limited.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436–465.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2007). Bayesian nonparametric latent feature models. *Bayesian Statistics*, 8:1–25.

Bhaskara, A., Charikar, M., and Vijayaraghavan, A. (2014). Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Conference on Learning Theory*, pages 742–778.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

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

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

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

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

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

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

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

Chen, J. and de la Torre, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Frontiers in psychology*, 9:1474.

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

Chen, Y., Culpepper, S. A., Wang, S., and Douglas, J. (2018b). A hidden markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement*, 42(1):5–23.

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

Chérief-Abdellatif, B.-E. and Alquier, P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *arXiv preprint arXiv:1805.05054*.

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

Choi, K. M., Lee, Y.-S., and Park, Y. S. (2015). What cdm can tell about what students have learned: an analysis of timss eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education*, 11(6):1563–1577.

Choi, M. J., Tan, V. Y., Anandkumar, A., and Willsky, A. S. (2011). Learning latent tree graphical models. *Journal of Machine Learning Research*, 12(May):1771–1812.

Culpepper, S. A. (2015). Bayesian estimation of the dina model with gibbs sampling. *Journal of Educational and Behavioral Statistics*, 40(5):454–476.

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

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

de la Torre, J. and Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81:253–273.

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

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

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

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

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a bayesian extension of the DINA model. *Applied Psychological Measurement*, 36(6):447–468.

Deloria Knoll, M., Fu, W., Shi, Q., Prosperi, C., Wu, Z., Hammitt, L. L., Feikin, D. R., Baggett, H. C., Howie, S. R., Scott, J. A. G., et al. (2017). Bayesian estimation of pneumonia etiology: epidemiologic considerations and applications to the pneumonia etiology research for child health study. *Clinical infectious diseases*, 64(suppl_3):S213–S227.

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

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

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

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

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3):495–515.

Embretson, S. E. and Reise, S. P. (2013). *Item response theory*. Psychology Press.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, 1(2):502–537.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

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

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

Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, pages 475–482.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Gu, Y. and Xu, G. (2019a). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, 20(115):1–58.

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

Gu, Y. and Xu, G. (2020a). Partial identifiability of restricted latent class models. *The Annals of Statistics*. To appear.

Gu, Y. and Xu, G. (2020b). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica*. To appear.

Gyllenberg, M., Koski, T., Reilink, E., and Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31:542–548.

Hall, M. (1967). *Combinatorial theory*, volume 71. John Wiley & Sons.

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

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

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

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

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

Hsu, D., Kakade, S. M., and Liang, P. S. (2012). Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems*, pages 1511–1519.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research, and Evaluation*, 15(1):3.

Jaeger, J., Tatsuoka, C., Berns, S. M., and Varadi, F. (2006). Distinguishing neurocognitive functions in schizophrenia using partially ordered classification models. *Schizophrenia Bulletin*, 32(4):679–691.

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

Kisa, D., Van den Broeck, G., Choi, A., and Darwiche, A. (2014). Probabilistic sentential decision diagrams. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Kruskal, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138.

Larochelle, H. and Bengio, Y. (2008). Classification using discriminative restricted Boltzmann machines. In *Proceedings of the 25th International Conference on Machine Learning*, pages 536–543. ACM.

Larochelle, H., Bengio, Y., and Turian, J. (2010). Tractable multivariate binary density estimation and the restricted Boltzmann forest. *Neural computation*, 22(9):2285–2307.

Lazarsfeld, P. F. (1959). Latent structure analysis. *Psychology: A study of a science*, 3:476–543.

Lee, Y.-S., Park, Y. S., and Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in massachusetts, minnesota, and the us national sample using the timss 2007. *International Journal of Testing*, 11(2):144–177.

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (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., Xu, G., and Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, 36:548–564.

Liu, J., Xu, G., and Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*, 19(5A):1790–1817.

Ma, W. and de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69(3):253–275.

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

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

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, 21(4):331–347.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Miller, K., Jordan, M. I., and Griffiths, T. L. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems*, pages 1276–1284.

Mourad, R., Sinoquet, C., Zhang, N. L., Liu, T., and Leray, P. (2013). A survey on latent tree models and applications. *Journal of Artificial Intelligence Research*, 47:157–203.

O'Brien, K. L., Baggett, H. C., Brooks, W. A., Feikin, D. R., Hammitt, L. L., Higdon, M. M., Howie, S. R., Knoll, M. D., Kotloff, K. L., Levine, O. S., et al. (2019). Causes of severe pneumonia requiring hospital admission in children without hiv infection from africa and asia: the perch multi-country case-control study. *The Lancet*.

Poon, H. and Domingos, P. (2011). Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690. IEEE.

Rahman, T., Kothalkar, P., and Gogate, V. (2014). Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of chow-liu trees. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 630–645. Springer.

Ravanbakhsh, S., Póczos, B., and Greiner, R. (2016). Boolean matrix factorization and noisy completion via message passing. In *Proceedings of the 33rd International Conference on Machine Learning-Volume 48*, pages 945–954.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41.

Rhodes, J. A. (2010). A concise proof of Kruskal's theorem on tensor decomposition. *Linear Algebra and its Applications*, 432(7):1818–1824.

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

Rukat, T., Holmes, C. C., Titsias, M. K., and Yau, C. (2017). Bayesian Boolean matrix factorisation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2969–2978. JMLR. org.

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine learning*, pages 791–798. ACM.

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

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

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

Su, Y.-L., Choi, K., Lee, W., Choi, T., and McAninch, M. (2013). Hierarchical cognitive diagnostic analysis for timss 2003 mathematics. *Centre for Advanced Studies in Measurement and Assessment*, 35:1–71.

Tatsuoka, C. (2009). Diagnostic models as partially ordered sets. *Measurement*, 7:49–53.

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

Teicher, H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38:1300–1302.

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

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

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

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).

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

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

von Davier, M. and Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional diagnosticclassification modelsa commentary. *Psychometrika*, 79(2):340–346.

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

Wang, S., Yang, Y., Culpepper, S. A., and Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43(1):57–87.

Weiss, D. J. and Yoes, M. E. (1991). Item response theory. In *Advances in educational and psychological testing: Theory and applications*, pages 69–95. Springer.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, pages 339–362.

Wu, Z., Casciola-Rosen, L., Rosen, A., and Zeger, S. L. (2018). A Bayesian approach to restricted latent class models for scientifically-structured clustering of multivariate binary outcomes. *arXiv preprint arXiv:1808.08326*.

Wu, Z., Deloria-Knoll, M., Hammitt, L. L., Zeger, S. L., and for Child Health Core Team, P. E. R. (2016). Partially latent class models for case–control studies of childhood pneumonia aetiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 65(1):97–114.

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

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

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

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

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

Yang, Y., Pati, D., and Bhattacharya, A. (2018). α-variational inference with statistical guarantees. *The Annals of Statistics, to appear*.

Yen, I. E., Lee, W.-C., Chang, S.-E., Suggala, A. S., Lin, S.-D., and Ravikumar, P. (2017). Latent feature lasso. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3949–3957. JMLR. org.

Zhang, S. S., DeCarlo, L. T., and Ying, Z. (2013). Non-identifiability, equivalence classes, and attribute-specific classification in Q-matrix based cognitive diagnosis models. *arXiv preprint arXiv:1303.0426*.