Latent Variable Modeling and Statistical Learning

Yunxiao Chen

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Latent Variable Modeling and Statistical Learning

Yunxiao Chen

Latent variable models play an important role in psychological and educational measurement, which attempt to uncover the underlying structure of responses to test items. This thesis focuses on the development of statistical learning methods based on latent variable models, with applications to psychological and educational assessments. In that connection, the following problems are considered.

The first problem arises from a key assumption in latent variable modeling, namely the local independence assumption, which states that given an individual's latent variable (vector), his/her responses to items are independent. This assumption is likely violated in practice, as many other factors, such as the item wording and question order, may exert additional influence on the item responses. Any exploratory analysis that relies on this assumption may result in choosing too many nuisance latent factors that can neither be stably estimated nor reasonably interpreted. To address this issue, a family of models is proposed that relax the local independence assumption by combining the latent factor modeling and graphical modeling. Under this framework, the latent variables capture the across-the-board dependence among the item responses, while a second graphical structure characterizes the local dependence. In addition, the number of latent factors and the sparse graphical structure are both unknown and learned from data, based on a statistically solid and computationally efficient method.

The second problem is to learn the relationship between items and latent variables, a structure that is central to multidimensional measurement. In psychological and educational assessments, this relationship is typically specified by experts when items are written and

is incorporated into the model without further verification after data collection. Such a non-empirical approach may lead to model misspecification and substantial lack of model fit, resulting in erroneous interpretation of assessment results. Motivated by this, I consider to learn the item - latent variable relationship based on data. It is formulated as a latent variable selection problem, for which theoretical analysis and a computationally efficient algorithm are provided.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank my Ph.D. advisors, Professors Zhiliang Ying and Jingchen Liu, for their encouragement, support, and guildance during my graduate study. This work would never have been done without their help. Their guildance is on both academic and personal levels, from which I will benefit for my whole life.

I would like to thank my committee members, Professors Young-Sun Lee, Shaw-Hwa Lo, and Doctor Matthias von Davier, for their advice and helpful insights. I am also thankful to Professor Michael Sobel, Dood Kalicharan and other faculty and staff in the Department of Statistics at Columbia University for their support and assistance over the past years. For their support and friendship, I would like to thank my friends and fellow students in the department, including Haolei Weng, Xiaoou Li, Lu Meng, and many others.

Finally, my warmest thanks go to my family. I am grateful to my patients and grandparents, for their constant care and support. I especially thank my beloved wife, Ting, for her continued and unfailing love, support, and understanding. And thanks my lovely daughter, Lucia, for always brightening my day. They are the most important people in my world and I dedicate this thesis to them.

To my family

Chapter 1

Introduction

Psychological assessment refers to a way of testing people about their behavior, personality, and capabilities to draw conclusions using combinations of techniques (Groth-Marnat, 2009). Its goal is to develop good understanding of the individuals. Conventionally, psychological assessments include psychiatric evaluation based on a standard questionnaire and the assessment of examinees' strengths and weaknesses on different skills based on their responses to test items. These assessments, properly administrated, help to prepare customized treatments to individuals with specific mental disorders and to impact the learning process of students by providing feedback. From a general sense, psychological assessment is everywhere in our life. For example, the preference of online consumers can be measured by their shopping history, which could help the E-store make individualized recommendations to boost profits. For another example, the political attitudes of senators may be measured based on their voting behavior, which can be used to make prediction. Therefore, although only applications in psychological and educational measurement are discussed in this thesis, the proposed models and methods can be potentially applied in many other fields.

Model-based psychological assessment refers to making psychological assessment based on a statistical model. The key is to properly model the relationship between individuals' underlying characteristics and their responses to items and thus make valid inference of the individuals. The advantages of model-based assessment is of two-fold. First, instead of

scoring the examinees/patients based on the subjective judgement, a measurement model provides a data-driven scoring rule that is more objective. Second, under the statistical hypothesis testing framework, a psychological theory becomes testable if it can be formulated as a measurement model. Different statistical models have been proposed for psychological assessment. Classical test theory (Gulliksen, 1950; Spearman, 1907, 1913) has been the most popular measurement model for most of the 20th century. It has defined the standard for test development, beginning with the initial explosion of testing in the 1930s. Instead of modeling item-level responses, the classical test theory models an individual's observed total score, by assuming it to be the true score plus some measurement error.

Item response theory (IRT; Rasch, 1960; Lord and Novick, 1968) is another important family of measurement models. It is generally regarded as being superior to classical test theory (Embretson and Reise, 2000) and has become the preferred method for developing scales in the United States, especially when high-stake decisions are demanded. For example, the Armed Services Vocational Aptitude Battery, the National Assessment of Education Progress (NAEP), the Scholastic Aptitude Test (SAT), and the Graduate Record Examination (GRE) apply IRT models to score the individuals. Unlike the classical test theory in which the items are interchangeable, the item response theory considers the items to be heterogenous. For example, items may differ in terms of their difficulty levels and therefore provide different measurement information. The latent trait of an individual is scored based on both their responses to test items and the administrated items. The early IRT applications involved primarily unidimensional IRT models that measure only a undimensional latent trait that may represent for example the math ability or the degree of depression of an individual. In subsequent developments, several multidimensional IRT models (e.g. McKinley and Reckase, 1982; Bock et al., 1988; Revuelta, 2014) have been proposed for modeling the response process driven by multiple latent traits.

Diagnostic classification models (DCMs) form another family of measurement models that have recently gained prominence in educational assessment, psychiatric evaluation, and many other disciplines (Rupp and Templin, 2008b; Rupp et al., 2010). In particular, the general diagnostic models, proposed in von Davier and Yamamoto (2004) and von Davier (2008), provide a general modeling framework that includes many popular diag-

nostic classification models and item response theory models as special cases (von Davier, 2008, 2014a,b). Instead of an estimate of the general ability in unidimensional IRT models and estimates of several subscores in multidimensional IRT models, a diagnostic test provides each individual with a profile detailing the concepts or skills (often called attributes) that he/she has mastered. In educational assessment, such clear-cut feedback could have a significant impact on learning process by providing students and teachers detailed information on students' strengths and weaknesses. Not only in educational assessment, diagnostic classification models are generally suitable whenever statistically-driven classifications of individuals according to multiple attributes are sought.

Using the IRT models and DCMs as the basic vehicles, this thesis attempts to solve two problems. The first is a modeling framework for exploratory analysis without the local independence assumption. The local independence assumption, which will be mathematically defined in Chapter 2, is made in most of the latent variable models. It says that given an individual's latent variable (vector), all the responses are independent. Under this assumption, response data from the administration of real tests are likely to require a large number of latent variables to accurately represent the variation in the data. Most of these latent variables may represent minor factors that are not the target of measurement, such as the item wording and question order (Knowles and Condon, 2000; Schwarz, 1999). On the other hand, for the purpose of interpretation, models with a small number of latent variables are desired, hoping to capture the major factors measured by the test. Unfortunately, these models may be lack of fit (Yen, 1984; Chen and Thissen, 1997; Sireci et al., 1991; Tuerlinckx and De Boeck, 2001; Braeken et al., 2007). This dilemma can be fixed by relaxing the local independence assumption. Chapter 3 discusses a modeling framework that allows for local dependence, by combining the latent factor modeling and graphical modeling. Under this framework, the latent variables capture the across-the-board dependence among the item responses, while a second graphical structure characterizes the local dependence. In addition, the number of latent factors and the sparse graphical structure are both unknown and learned from data, based on a statistically solid and computationally efficient method.

The second one is a data-driven method to learn the relationship between items and latent variables, a structure that is central to multidimensional measurement. In psycho-

logical and educational assessments, this relationship is typically specified by experts when items are written and is incorporated into the model without further verification after data collection. Such a non-empirical approach may lead to model misspecification and substantial lack of model fit, resulting in erroneous interpretation of assessment results. Motivated by this, I consider to learn the item - latent variable relationship based on data. It is formulated as a latent variable selection problem, for which theoretical analysis and a computationally efficient algorithm are provided. This problem will be discussed in Chapter 4. Part of this work can be found in Chen et al. (2015).

The rest of the thesis will be organized as follows. In Chapter 2, a unified framework for the measurement models is introduced, which includes the IRT models and DCMs as special cases. Under this framework, we motivate the works in Chapter 3 and 4, by discussing some common problems in latent variable measurement models. Then several popular IRT models and DCMs are reviewed, some of which will be used as the building blocks of the subsequent analysis. In Chapter 3, the latent and undirected graphical measurement models are proposed to relax the local independence assumption in most latent variable models. In addition, the theoretical and computational analysis are provided for the statistical inference based on the proposed model. In Chapter 4, we consider to learn the item - latent variable relationship based on data. It is formulated as a latent variable selection problem, for which theoretical analysis and a computationally efficient algorithm are provided.

Chapter 2

Introduction to Latent Variable Models for Measurement

2.1 A Latent Variable Model Framework for Measurement

2.1.1 A Latent Variable Model Framework for Measurement

Consider that there are N individuals, each of whom responds to J items. An individual's responses to items are denoted as $\mathbf{Y} = (Y_1, ..., Y_J)$, where Y_j are categorical response $Y_j \in \{0, ..., c_j\}$ for $c_j + 1$ categories. An item j is called a binary item when $c_j = 1$. In addition, let $\mathbf{y} = (y_1, ..., y_J)$ be an observation of $\mathbf{Y}$. All these measurement models assume that an individual's responses to items are driven by his/her latent (unobserved) characteristics $\boldsymbol{\theta} \in \mathbb{R}^K$ (boldfaced $\boldsymbol{\theta}$ will be used when $K > 1$ and unboldfaced θ will be used when $K = 1$). Different assumptions are made on $\boldsymbol{\theta} = (\theta_1, ..., \theta_K) \in \mathbb{R}^K$ in different measurement models and we distinguish them as follows.

Unidimensional IRT:	$\theta \in \mathbb{R}^1 (K = 1)$ and θ is a continuous random variable.
Multidimensional IRT:	$\boldsymbol{\theta} \in \mathbb{R}^K (K > 1)$ and $\boldsymbol{\theta}$ is a continuous random vector.
DCM:	$\boldsymbol{\theta} \in \mathbb{R}^K (K > 1)$ and $\boldsymbol{\theta}$ is a discrete random vector. For most of the DCMs, $\theta_k \in \{0, 1\}$, representing nonmastery ($\theta_k = 0$) and mastery ($\theta_k = 1$) of a skill.

$\mathbf{Y}$, $\mathbf{y}$ and $\boldsymbol{\theta}$ are individual-specific and we will later use subscript to indicate different individuals. That is $\mathbf{Y}_i$ and $\boldsymbol{\theta}_i$ are the responses and the latent characteristics of individual i and $\mathbf{y}_i$ is an observation of $\mathbf{Y}_i$. $\boldsymbol{\theta}_i$s are assumed to be independent and identically distributed random variables (vectors), with density

$$\boldsymbol{\theta}_i \sim f(\boldsymbol{\theta}),$$

where f has a dominating measure μ and f could be either known or unknown. μ is typically a Lebesgue measure for IRT models and a counting measure for DCMs, because $\boldsymbol{\theta}$ is continuous for IRT models and discrete for DCMs. Given $\boldsymbol{\theta}$, the remainder of model specification becomes a regression problem

$$\mathbf{Y}|\boldsymbol{\theta} \sim \varphi(\mathbf{y}|\boldsymbol{\theta}),$$

where $\varphi(\mathbf{y}|\boldsymbol{\theta})$ is typically parametric and varies for different models. To simplify the specification of $\varphi(\mathbf{y}|\boldsymbol{\theta})$, most of the measurement models make the *local independence assumption*. That is, Y_1, Y_2, ..., Y_J are independent given $\boldsymbol{\theta}$ and

$$Y_j|\boldsymbol{\theta} \sim \varphi_j(y_j|\boldsymbol{\theta}).$$

As a consequence, $\varphi(\mathbf{y}|\boldsymbol{\theta})$ takes a product form:

$$\varphi(\mathbf{y}|\boldsymbol{\theta}) = \prod_{j=1}^{J} \varphi_j(y_j|\boldsymbol{\theta}). \tag{2.1}$$

This is a strong assumption, which is not satisfied in many situations. For example, in educational testing, sets of items (testlets) often come from a single common stimulus (e.g. a reading comprehension passage). In this setting, all items given to an examinee are unlikely to be conditionally independent (given examinee proficiency). Besides, other factors such as the item wording and question order may cause local dependence as well (Knowles and Condon, 2000; Schwarz, 1999). When the local independence assumption fails, the regression function $\varphi(\mathbf{y}|\boldsymbol{\theta})$ is not well approximated by the product form in (2.1), resulting in substantial lack of fit. The discussion in Chapter 2 will be around relaxing this assumption by combining latent variable and graphical modeling.

The goal of the measurement model is to make inference on the individuals' $\boldsymbol{\theta}_i$s. Under this framework, if both the marginal distribution $f(\boldsymbol{\theta})$ and the conditional distribution $\varphi(\mathbf{y}|\boldsymbol{\theta})$ are known, $\boldsymbol{\theta}_i$ is best characterized by the posterior distribution

$$p(\boldsymbol{\theta}_i|\mathbf{y}_i) \propto f(\boldsymbol{\theta}_i)\varphi(\mathbf{y}_i|\boldsymbol{\theta}_i).$$

Unfortunately, φ is typically unknown and f is sometimes unknown as well. As a consequence, a two-stage procedure is often used (Bock and Aitkin, 1981; Embretson and Reise, 2000):

Step 1: Estimate f and φ based on data and obtain $\hat{f}$ and $\hat{\varphi}$.

Step 2: Make inference on $\boldsymbol{\theta}_i$ based on

$$\hat{p}(\boldsymbol{\theta}_i|\mathbf{y}_i) \propto \hat{f}(\boldsymbol{\theta}_i)\hat{\varphi}(\mathbf{y}_i|\boldsymbol{\theta}_i).$$

In step 1, f and φ are usually estimated by the marginal maximal likelihood estimator (Bock and Aitkin, 1981; Johnson et al., 2007) and this two-stage procedure is related to the empirical Bayes method in statistics (Robbins, 1956; Laird and Ware, 1982; Morris, 1983). Under the latent variable model framework, φ and f are typically estimated based on the marginal likelihood

$$L(\varphi, f) = \prod_{i=1}^{N} \int \varphi(\mathbf{y}_i|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \mu(d\boldsymbol{\theta}),$$

where $\boldsymbol{\theta}_i$'s are marginalized out from the joint distribution as they are not observable.

2.1.2 Log-Linear Models

Following the description above, when in particular $\varphi_j(y_j|\boldsymbol{\theta})$s are following a log-linear model, the corresponding latent variable model is called a log-linear model. Log-linear models are commonly discussed in the context of categorical data analysis (Agresti, 1996). Many of the popular item response theory and diagnostic classification models can be viewed as special cases of the log-linear measurement model (see e.g. von Davier, 2008; Henson et al., 2009; de la Torre, 2011). In addition, although the normal ogive IRT models are not within the log-linear model framework, they can be well approximated by log-linear models, due to

the relationship between the normal and logistic distribution functions (Haley, 1952). Introducing this log-linear model framework helps us better understand the specific measurement models and will facilitate our subsequent discussion.

Binary variables. When $Y_j \in \{0, 1\}$ is a binary variable, the log-linear model can be viewed as a logistic regression model:

$$\text{logit}(P(Y_j = 1|\boldsymbol{\theta})) = \beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k}\theta_k + \sum_{1 \le k_1 < k_2 \le K} \beta_{j,k_1k_2}\theta_{k_1}\theta_{k_2} + \cdots + \beta_{j,12\cdots K} \prod_{k=1}^{K} \theta_k, \quad (2.2)$$

where

$$\text{logit}(P(Y_j = 1|\boldsymbol{\theta})) = \log\left(\frac{P(Y_j = 1|\boldsymbol{\theta})}{P(Y_j = 0|\boldsymbol{\theta})}\right).$$

This is a log-linear model with all the interactions of $\theta_1, ..., \theta_K$. For many IRT models, the interaction coefficients are constrained to be 0 and for some DCMs, the interaction terms are needed. Specifically, the function

$$\phi(\boldsymbol{\theta}) = P(Y_j = 1|\boldsymbol{\theta})$$

is usually called the item response function, or item characteristic curve (Tucker, 1946).

Categorical variables. When $c_j > 1$, the model naturally extends the binary case by

$$\log\left(\frac{P(Y_j = y|\boldsymbol{\theta})}{P(Y_j = 0|\boldsymbol{\theta})}\right) = \beta^y_{j,0} + \sum_{k=1}^{K} \beta^y_{j,k}\theta_k + \sum_{1 \le k_1 < k_2 \le K} \beta^y_{j,k_1k_2}\theta_{k_1}\theta_{k_2} + \cdots + \beta^y_{j,12\cdots K} \prod_{k=1}^{K} \theta_k$$

for $y \in 1, ..., c_j$. This is consistent with (2.2) when $c_j = 1$.

This log-linear model framework contains many well-known item response theory models, such as the unidimensional and multidimensional versions of the Rasch model (Rasch, 1960), two parameter logistic model (2PL; Birnbaum, 1968), partial credit model (PCM; Masters, 1982), generalized partial credit model (GPCM; Muraki, 1992), and the nominal response model (NRM; Bock, 1972). For all these item response theory models, interaction terms do not appear. Many popular diagnostic classification models can be formulated under this framework as well, such as the conjunctive DINA and NIDA models (Junker, 1999; Tatsuoka, 2002; de la Torre and Douglas, 2004), the compensatory DINO and NIDO models (Templin and Henson, 2006), the C-RUM model (Hartz, 2002), and the G-DINA

model (de la Torre, 2011). For a specific model, there are model-specific constraints on the log-linear model parameters.

2.2 Confirmatory and Exploratory Analyses and Q-matrix

A central quantity in the specification of a multidimensional measurement model (when $K > 1$) is the test design matrix, which describes the relationship between the items and the components of the latent characteristics $\boldsymbol{\theta}$. Following the literature of diagnostic classification models, we call this design matrix the Q-matrix (Tatsuoka, 1983). It is also known as the loading structure in the IRT literature. The design matrix $Q = (q_{jk})_{J\times K}$ is a J by K matrix with zero-one entries, each of which indicates whether an item is associated to a dimension of $\boldsymbol{\theta}$. More precisely, each row of Q represents an item and each column represents a dimension of $\boldsymbol{\theta}$ (e.g. a skill). For example, the Q-matrix of a math test that measures the subtraction and multiplication skills (represented by θ_1 and θ_2 respectively) is specified in Table 2.1. In this example, the first item "$7-2$" only measures the subtraction skill and therefore the first row of Q is $(1, 0)$. The third item "$(7-2)\times 2$" measures both skills and thus the corresponding row is $(1, 1)$.

Example 1. *An example of the Q-matrix of a math test that measures the subtraction and multiplication skills.*

$Q=$

	subtraction	multiplication
1. $7-2$	1	0
2. 5×2	0	1
3. $(7-2)\times 2$	1	1

.

Table 2.1: An example of the Q-matrix.

When the Q-matrix is known and incorporated into the measurement model, the model is called a *confirmatory model*. Under the log-linear model framework, incorporating the Q-matrix is equivalent to setting constraints to the model parameters. More precisely, in

the binary case,

$$\beta_{j,k_1\cdots k_T} = 0 \text{ if } \prod_{t=1}^{T} q_{j,k_t} = 0,$$

where $\{k_1, \cdots k_T\} \subset \{1, 2, ..., K\}$ is a subset of latent variables. It means that if θ_k is irrelevant to an item j ($q_{jk} = 0$), the coefficients of θ_k and interactions having θ_k are all 0. For example, for the item "$7 - 2$" in Table 2.1, $\beta_{1,2} = \beta_{1,12} = 0$ and for the item "5×2", $\beta_{2,1} = \beta_{2,12} = 0$. In the categorical response case,

$$\beta^{y}_{j,k_1\cdots k_T} = 0 \text{ if } \prod_{t=1}^{T} q_{j,k_t} = 0, \text{ for all } y \in \{1, 2, ..., c_j\}.$$

Under these constraints, it is guaranteed that two individuals have the same item response function if they have all required skills according to $\mathbf{q}_j$:

$$\varphi_j(y|\boldsymbol{\theta}) = \varphi_j(y|\boldsymbol{\theta}') \text{ for all } y, \text{ if } \boldsymbol{\theta} \geq \mathbf{q}_j \text{ and } \boldsymbol{\theta}' \geq \mathbf{q}_j,$$

where $\boldsymbol{\theta} \geq \mathbf{q}_j$ means that $\theta_k \geq q_{jk}$ for all k. In other words, having additional irrelevant skills does not change the item response distribution.

An exploratory analysis does not incorporate the Q-matrix into the model, which is considered when there is no clear hypothesis about the structure or when an unconstrained solution is preferred to verify the design. The Q-matrix is central to many multidimensional measurement models. In psychological and educational assessments, the Q-matrix is typically specified by experts when items are written and is incorporated into the model without further verification after data collection. Such a non-empirical approach may lead to model misspecification and substantial lack of model fit, resulting in erroneous interpretation of assessment results. A natural idea arises: can we recover the Q-matrix from data by starting from an exploratory model? This is a challenging problem due to the complicated latent structure and we will further discuss it in Chapter 4.

2.3 Item Response Theory Models

In what follows, I give a brief introduction to popular IRT models under the log-linear model framework. A comprehensive review of the unidimensional and multidimensional IRT models can be found in Embretson and Reise (2000) and Reckase (2009).

Rasch Model (Rasch, 1960). The Rasch model is a unidimensional IRT model ($K = 1$) for binary responses, where

$$\text{logit}(P(Y_j = 1|\theta)) = \theta - b_j,$$

or equivalently the item response function

$$\phi(\theta) = \frac{e^{\theta - b_j}}{1 + e^{\theta - b_j}}.$$

In other words, $\beta_{j,0} = -b_j$ and $\beta_{j,1} = 1$ under the log-linear model framework. For a given b_j, $\phi(\theta)$ is a monotone function of θ. For example, in educational testing, it means that an examinee with a higher ability has a higher probability to answer an item correctly. The parameter b_j is known as the difficulty parameter or the location parameter that shows where the item response function achieves its central value between its lower and upper asymptotes. A larger value of b_j indicates more difficult items, with smaller success probabilities given the same ability level θ, while a smaller value of b_j indicates the reverse. In particular, when $\theta = b_j$, the correct response probability takes value $1/2$. Figure 2.1 gives three item response functions for the Rasch model with $b_j = -1, 0, 1$.

Two Parameter Logistic Model (2PL; Birnbaum, 1968). The limitation of the Rasch model is that all the item characteristic curves have the same shape and therefore is not flexible enough. Birnbaum (1968) proposed a slightly more complex model

$$\text{logit}(P(Y_j = 1|\theta)) = a_j(\theta - b_j),$$

or equivalently the item response function

$$\phi(\theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}},$$

where a_j, known as the discrimination parameter, is a parameter related to the maximum slope of the item characteristic curve. More precisely,

$$\phi'(\theta) = a_j\phi(\theta)(1 - \phi(\theta)).$$

Since $\phi(\theta)(1 - \phi(\theta))$ is maximized when $\theta = b_j$ with maximum $\frac{1}{4}$,

$$a_j = 4 \max_{\theta} \phi'(\theta),$$

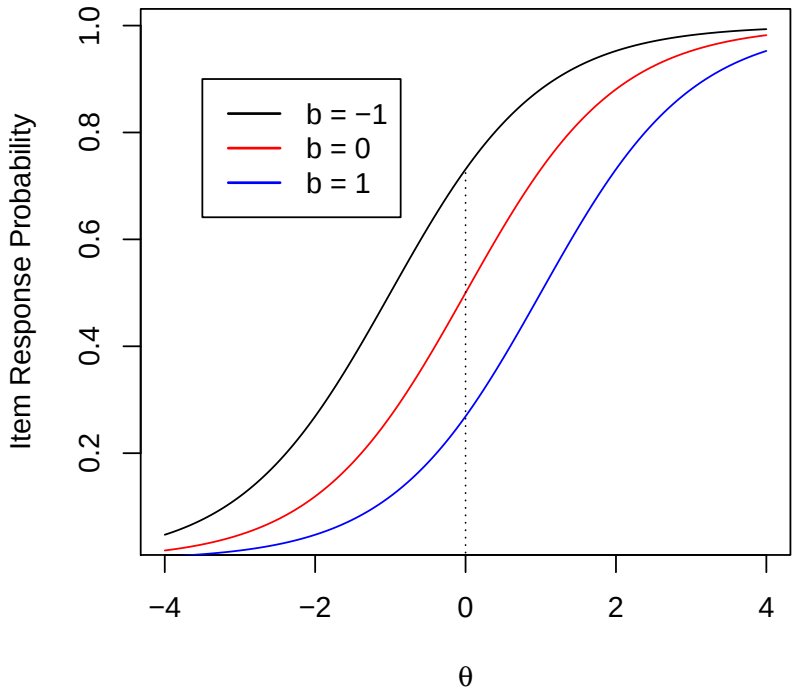

Figure 2.1: The item characteristic curve for the Rasch model.

In other words, a_j is 4 times of the maximal slope of the item characteristic curve. Figure 2.2 gives the item response curves for the 2PL model under different values of a_j and $b_j = 0$. As we can see, as the discrimination parameter a_j increases, the slope of the item characteristic curve at θ close to 0 becomes more steep. In the limiting case $a_j = +\infty$ (and when $b_j = 0$),

$$\phi(\theta) = \begin{cases} 0 & \theta < 0, \\ 0.5 & \theta = 0, \\ 1 & \theta > 0. \end{cases}$$

This limiting case is known as the Guttman scale (Guttman, 1944), which is related to the diagnostic classification models that will be discussed in a sequel. In addition, when $a_j = 1$ for all j, the 2PL model becomes the Rasch model. The discrimination parameter a_j has another interpretation based on the Fisher information theory, where the Fisher information

for a specific item j is defined as

$$I(\theta) = \frac{d^2 \log(\phi(\theta))}{d\theta^2},$$

which measures the efficiency of the item for estimating an examinee with true ability θ. For the 2PL model,

$$I(\theta) = a_j^2 \phi(\theta)(1 - \phi(\theta)).$$

Therefore, for a given item, the Fisher information is maximized at $\theta = b_j$ and the maximal information is $\frac{a_j^2}{4}$, which connects the discrimination parameter with the Fisher information. In figure 2.3, the Fisher information $I(\theta)$ is presented for $a_j = 1, 2, 5$ and $b_j = 0$. As we can see, the Fisher information takes a bell shape, with a peak at b_j. In addition, the height of the peak is decided by a_j. In particular, if $a_j = 0$, the Fisher information $I(\theta) = 0$ and Y_j becomes independent of θ. In other words, the item does not contain any information about θ in this case.

Partial Credit Model (PCM; Masters, 1982). The partial credit model is designed for polytomous response data and it can be viewed as an extension of the Rasch model. It was originally developed for analyzing test items that require multiple steps and for which it is important to assign partial credit for completing several steps in the solution process. Thus, the partial credit model lends itself naturally to describing item responses to achievement tests (e.g., math problems) where partially correct answers are possible. The PCM is also appropriate for analyzing attitude or personality scale responses where subjects rate their beliefs, or respond to statements on a multi-point scale (see Masters and Wright, 1997).

More precisely, for $y \geq 1$, the probability

$$P(Y_j = y|\theta, Y_j \in \{y - 1, y\})$$

is modeled by a Rasch model

$$\text{logit}(P(Y_j = y|\theta, Y_j \in \{y - 1, y\})) = \theta - b_j^y.$$

According to the model specification above, it is not difficult to show that

$$\log\left(\frac{P(Y_j = y|\theta)}{P(Y_j = 0|\theta)}\right) = y\theta - \sum_{l=1}^{y} b_j^l.$$

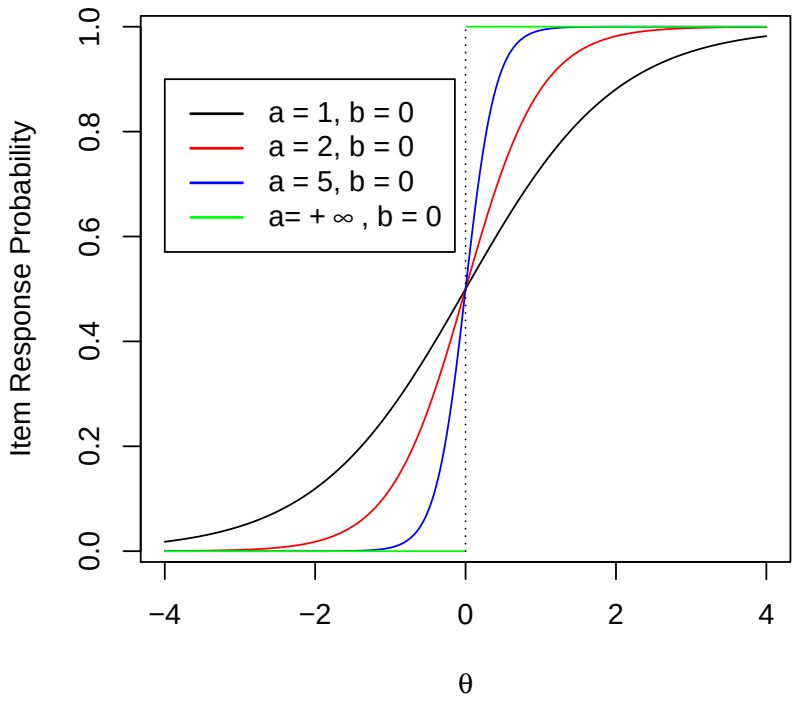

Figure 2.2: The item characteristic curve for the 2PL model.

It means that under the log-linear model framework,

$$\beta^y_{j,0} = -\sum_{k=1}^{y} b^k_j \quad \text{and} \quad \beta^y_{j,1} = y.$$

The score characteristic functions shown in Figure 2.4 are for an item with three score categories, 0, 1, and 2 and item parameters $b^1_j = 0$ and $b^2_j = 2$. The curves in the figure show the probability of each score category for an individual at different θ levels. For example, if θ is equal to -1, the probability of a score of 0 is 0.72, 0.27, and 0.01. A score of 0 is the most likely one at this θ-level, but the other scores are also possible.

Generalized Partial Credit Model (GPCM; Muraki, 1992). Similar to the extension from the Rasch model to 2PL model, the generalized partial credit model extends the

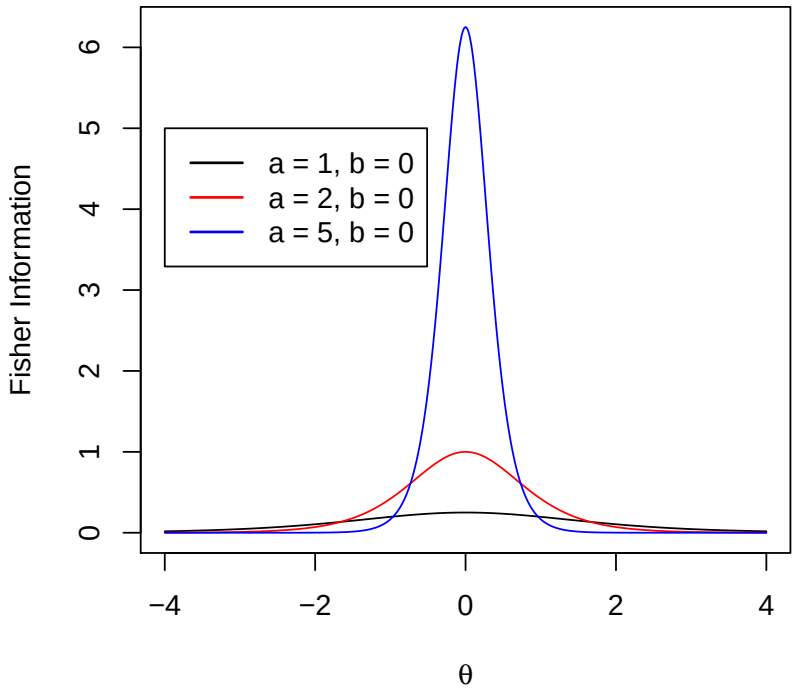

Figure 2.3: The item Fisher information for the 2PL model.

partial credit model by adding an item specific discrimination parameter a_j. That is

$$\text{logit}(P(Y_j = y|\theta, Y_j \in \{y-1, y\})) = a_j(\theta - b_j^y),$$

where a_j does not depend on the score category k. As a result,

$$\log\left(\frac{P(Y_j = y|\theta)}{P(Y_j = 0|\theta)}\right) = ya_j\theta - a_j\sum_{l=1}^{y} b_j^l.$$

Then we have under the log-linear model framework,

$$\beta_{j,0}^y = -a_j\sum_{k=1}^{y} b_j^k \quad \text{and} \quad \beta_{j,1}^y = ya_j.$$

The generalized partial credit model is more flexible than the partial credit model and it has become one of the most popular models for polytomous response data. For example, the National Assessment of Education Progress (NAEP), that is the largest nationally

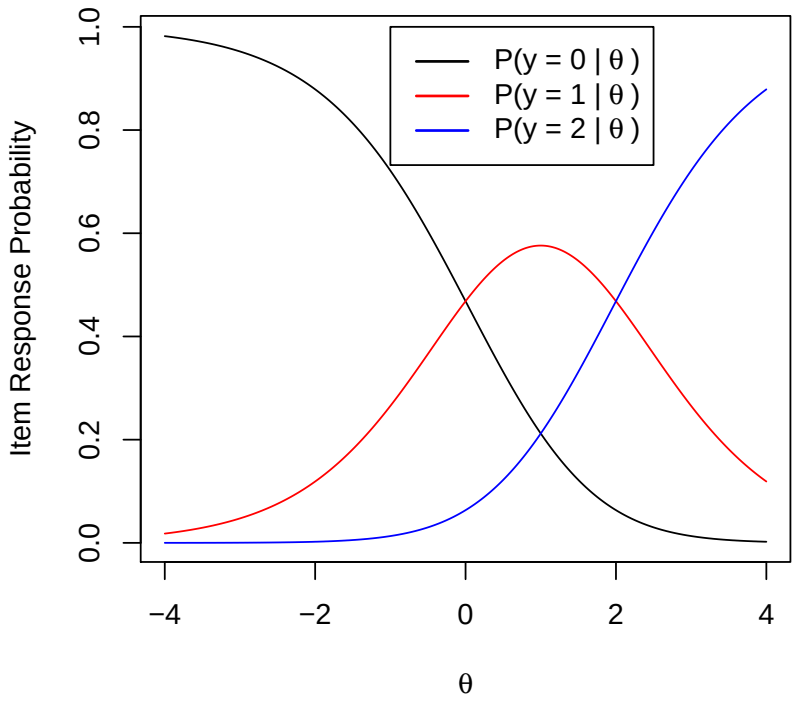

Figure 2.4: Score category probabilities for a three-category item under the PCM.

representative and continuing assessment of American students in various subject areas, uses the generalized partial credit model to analyze the constructed-response questions (Allen et al., 1999).

Nominal Response Model (NRM; Bock, 1972). Bock (1972) proposed a general IRT model that can be used to characterize item responses when responses are not necessarily ordered along the trait continuum (i.e. nominal responses). An example of a nominal response item is as follows.

Example 2. *Which political party would you identify yourself with?*
0. Democrat 1. Republican 2. Independent 3. Unaffiliated

The NRM model assumes that

$$\log\left(\frac{P(Y_j = y|\theta)}{P(Y_j = 0|\theta)}\right) = a_j^y(\theta - b_j^y).$$

for $y \in \{1, ..., c_j\}$. Under the log-linear model framework,

$$\beta_{j,0}^y = -a_j^y b_j^y \text{ and } \beta_{j,1}^y = a_j^y.$$

In fact, the PCM and GPCM models above are special cases of the nominal response model (Thissen and Steinberg, 1986) and the NRM model may be used for any items with multiple response options, which includes applications in personality and attitude assessment (Thissen, 1993).

Multidimensional 2PL Model (M2PL; McKinley and Reckase, 1982). In many applications, a set of test items may measure multiple skills (or mental disorders, etc.). The multidimensional two parameter logistic (M2PL) model is one of the most popular models for measuring multiple latent traits when responses are binary. The M2PL model is a log-linear model,

$$\text{logit}(P(Y_j = 1|\boldsymbol{\theta})) = d_j + \mathbf{a}_j^\top \boldsymbol{\theta}, \tag{2.3}$$

or equavalently, the item response function,

$$\phi(\boldsymbol{\theta}) = \frac{e^{d_j + \mathbf{a}_j^\top \boldsymbol{\theta}}}{1 + e^{d_j + \mathbf{a}_j^\top \boldsymbol{\theta}}}.$$

This model is labeled as a compensatory model, as the success probability only depends on the linear combination $d_j + \mathbf{a}_j^\top \boldsymbol{\theta}$ and thus a low ability on one dimension (e.g. small θ_1) can be compensated by a high ability on another dimension (e.g. large θ_2). Figure 2.5 represents the surface plot for item response probability under the M2PL model for an item with $\mathbf{a}_j = (0.5, 1)$ and $d_j = 0$. For example, under this setting, $\boldsymbol{\theta} = (-2, 1)$ and $\boldsymbol{\theta'} = (0, 0)$ both yield a success probability 0.5. In this example, the first dimension of $\boldsymbol{\theta}$ is compensated by its second dimension, so that it has the same item response probability as $\boldsymbol{\theta'}$.

The M2PL model can be used in either a confirmatory or an exploratory manner. For a confirmatory M2PL model, the design matrix Q is available and the parameters $\mathbf{a}_j$s are constrained as

$$a_{jk} = 0 \text{ if } q_{jk} = 0.$$

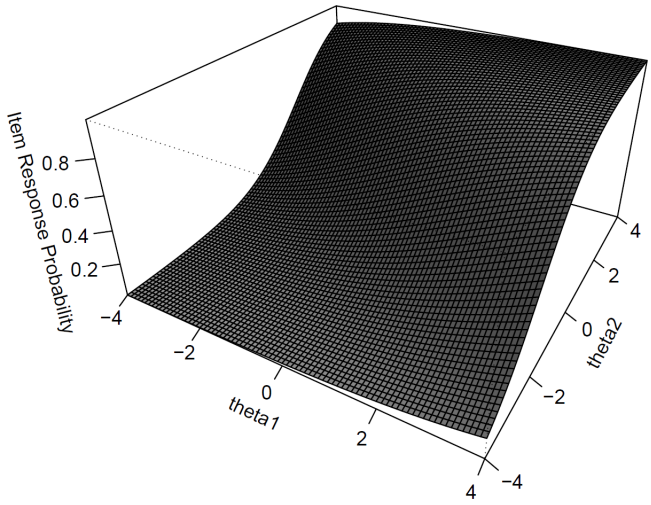

Figure 2.5: Surface plot for item response probability for an item under the M2PL model.

Multidimensional GPCM (MGPCM; Yao and Schwarz, 2006). The multidimensional generalized partial credit model (MGPCM) extends the GPCM to measure multiple latent traits:

$$\log\left(\frac{P(Y_j = y|\boldsymbol{\theta})}{P(Y_j = 0|\boldsymbol{\theta})}\right) = y\mathbf{a}_j^\top\boldsymbol{\theta} + d_j^y.$$

It is a log-linear model with

$$\beta_{j,0}^y = d_j^y \quad , \quad \beta_{j,k}^y = ya_{j,k},$$

and the parameters corresponding to the interaction terms are all 0. Similar to the M2PL model, a confirmatory testing design can be incorporated into the MGPCM model by setting

$$a_{jk} = 0 \;\text{ if }\; q_{jk} = 0.$$

Multidimensional NRM (MNRM; Revuelta, 2014). The multidimensional nominal response model extends the NRM to the multidimensional case:

$$\log\left(\frac{P(Y_j = y|\boldsymbol{\theta})}{P(Y_j = 0|\boldsymbol{\theta})}\right) = (\mathbf{a}_j^y)^\top\boldsymbol{\theta} + d_j^y, \tag{2.4}$$

where $\mathbf{a}_j^y = (a_{j1}^y, ..., a_{jK}^y)^\top$ is a K-dimensional vector containing the slope parameters of item j and the response category y. It is a log-linear model with

$$\beta_{j,0}^y = d_j^y \quad , \quad \beta_{j,k}^y = a_{j,k}^y,$$

and the parameters corresponding to the interaction terms are all 0. A confirmatory testing design can be incorporated into the MNRM model by setting

$$\|\mathbf{a}_{jk}\| = 0 \text{ if } q_{jk} = 0,$$

where $\mathbf{a}_{jk} = (a_{jk}^1, ...a_{jk}^{c_j})^\top$ is the vector of slope parameters corresponding to jth item and kth dimension of the latent vector and $\|\cdot\|$ is the Euclidean norm.

2.4 Diagnostic Classification Models

We review several diagnostic classification models that assume $\theta_k \in \{0, 1\}$, representing nonmastery ($\theta_k = 0$) and mastery ($\theta_k = 1$) of a skill. These models are confirmatory in nature and therefore their specification typically involves the Q-matrix. All these models can be parameterized under the log-linear model framework. A comprehensive review of diagnostic classification models can be found in Rupp et al. (2010).

DINA Model (Junker, 1999). The DINA model assumes a conjunctive relationship among attributes. It is necessary to possess all the attributes indicated by the Q-matrix to be capable of providing a positive response. In addition, having additional unnecessary attributes does not compensate for the lack of necessary attributes. This assumption is sometimes reasonable in practice; for example, it is believed that both subtraction and multiplication skills are necessary for answering item 3 in Table 2.1 and an examinee who lacks one of them is not able to solve the item. The DINA model is popular in the educational testing applications and is often employed for modeling exam problem solving processes. For each item j and attribute vector $\boldsymbol{\theta}$, we define the ideal response

$$\xi_{DINA}^j(\boldsymbol{\theta}, Q) = \prod_{k=1}^{K} (\theta_k)^{q_{jk}} = 1_{(\theta_k \geq q_{jk} \text{ for all } k)} \tag{2.5}$$

that is, whether $\boldsymbol{\theta}$ has all the attributes required by item j. For each item, there are two additional parameters s_j and g_j that are known as the slipping and guessing parameters, where s_j quantifies the probability of making a mistake when the subject is able to solve the item and g_j quantifies the probability of correctly answering the item by guessing. More

precisely, the response probability $P(Y_j = 1|\boldsymbol{\theta})$ takes the form

$$P(Y_j = 1|\boldsymbol{\theta}) = (1 - s_j)^{\xi^j_{DINA}(\boldsymbol{\theta},Q)} g_j^{1-\xi^j_{DINA}(\boldsymbol{\theta},Q)}. \tag{2.6}$$

If $\xi^j_{DINA}(\boldsymbol{\theta}, Q) = 1$ (the subject is capable of solving a problem), then the positive response probability is $1 - s_j$; otherwise, the probability is g_j. In addition, the marginal distribution $f(\boldsymbol{\theta})$ is assumed to be unknown and follows a categorical distribution. That is,

$$f(\boldsymbol{\theta}) = p_{\boldsymbol{\theta}}, \quad \boldsymbol{\theta} \in \{0, 1\}^K$$

representing the proportion of attribute profile $\boldsymbol{\theta}$ in the population. The vector $\mathbf{p} = (p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \{0, 1\}^K)$ is estimated based on the data.

The DINA model can be reparameterized sparsely under the log-linear model framework. Let q_{jk_1}, ..., q_{jk_T} be all the nonzero entries of $\mathbf{q}_j = (q_{j1}, ..., q_{jK})$. Then,

$$\beta_{j,0} = \text{logit}(g_j)$$

and

$$\beta_{j,k_1 \cdots k_T} = \text{logit}(1 - s_j) - \text{logit}(g_j)$$

are the only nonzero coefficients for the log-linear model.

DINO Model (Templin and Henson, 2006). The DINO model can be viewed as the compensatory counterpart of the DINA model. It assumes that it is enough to have any of the required attributes to correctly respond to the item. It is often employed in the application of psychiatric assessment, for which the positive response to a diagnostic question (item) could be due to the presence of one disorder (attributes) among several. The ideal response of the DINO model assumes that

$$\xi^j_{DINO}(\boldsymbol{\theta}, Q) = 1 - \prod_{k=1}^{K} (1 - \theta_k)^{q_{jk}} = 1_{(\theta_k \geq q_{jk} \text{ for at least one } k)}. \tag{2.7}$$

Similar to the DINA model, the positive response probability is

$$P(Y_j = 1|\boldsymbol{\theta}) = (1 - s_j)^{\xi^j_{DINO}(\boldsymbol{\theta},Q)} g_j^{1-\xi^j_{DINO}(\boldsymbol{\theta},Q)}.$$

The analysis with respect to the DINO model can be translated to the DINA model, by making use of their duality described as follows.

Proposition 1 (Duality between DINA and DINO Model)**.** *Consider a response vector* $\mathbf{Y} = (Y_1, ..., Y_J)$ *following a DINA model with latent attribute* $\boldsymbol{\theta}$ *and* $\mathbf{Y'} = (Y'_1, ..., Y'_J)$ *following the DINO model with latent attribute* $\boldsymbol{\theta'}$*. Their slipping and guessing parameters are denoted by* s_j*,* g_j*,* s'_j*, and* g'_j*, respectively. If* $1 - s_j = g'_j$*,* $g_j = 1 - s'_j$*, and* $\theta_k = 1 - \theta'_k$*, for* $j = 1, ..., J$ *and* $k = 1, ..., K$*, then* $\mathbf{Y}$ *and* $\mathbf{Y'}$ *are identically distributed.*

This proposition is straightforward to verify through the specifications of the two models. Thus, we omit the detailed proof.

C-RUM Model (Hartz, 2002). The compensatory reparameterized unified model (C-RUM) has the same structure as in the M2PL model, except that the elements of $\boldsymbol{\theta}$ are binary. The model can be specified as a log-linear model with no interaction terms:

$$\text{logit}(P(Y_j = 1|\boldsymbol{\theta})) = \beta_{j,0} + \sum_{k=1}^{K} \beta_{j,k}\theta_k,$$

where $\beta_{j,k} = 0$ when $q_{jk} = 0$.

G-DINA Model (de la Torre, 2011). The G-DINA model under the logistic link function is taking form of the saturated log-linear model:

$$\text{logit}(P(Y_j = 1|\boldsymbol{\theta})) = \beta_{j,0} + \sum_{k=1}^{K} \beta_{k,j}\theta_k + \sum_{1 \le k_1 < k_2 \le K} \beta_{j,k_1k_2}\theta_{k_1}\theta_{k_2} + \cdots + \beta_{j,12\cdots K} \prod_{k=1}^{K} \theta_k,$$

where

$$\beta_{j,k_1\cdots k_T} = 0 \text{ when } \prod_{t=1}^{T} q_{j,k_t} = 0$$

for any subset $\{k_1, \cdots, k_T\} \subseteq \{1, ..., K\}$.

Chapter 3

Latent and Undirected Graphical Measurement Models and Exploratory Analysis

3.1 Introduction

Nowadays, computer-based psychological and educational tests are prevalent, in which a large number of responses with complex dependence structure is often observed. As a consequence, a low-dimensional latent vector is often insufficient to capture all the variations of the responses. Contextual factors, such as the item wording and question order, may exert additional influence on the item response (Knowles and Condon, 2000; Schwarz, 1999; Yen, 1993). Furthermore, it is conceivable that problem solving and task accomplishing are often very complicated cognitive processes and are hardly completely governed by only a few latent factors, especially in the presence of high-dimensional responses. Keeping K small will violate the local independence assumption, which will result in series model lack of fit (Yen, 1984; Chen and Thissen, 1997; Sireci et al., 1991; Tuerlinckx and De Boeck, 2001; Braeken et al., 2007). On the other hand, including many latent factors typically results in a model that is difficult to interpret or to obtain stable estimations.

From the technical aspect, a low-dimensional factor model is not rich enough to capture

all the dependence of the responses. A number of approaches have been developed to deal with local dependence, for example, using shared random effects (e.g. Gibbons and Hedeker, 1992; Wilson and Adams, 1995; Bradlow et al., 1999; Wang and Wilson, 2005; Jeon et al., 2013), fixed interaction parameters (e.g. Hoskens and De Boeck, 1997; Haberman, 2007), and copula models (e.g. Braeken et al., 2007; Braeken, 2011). However, all these methods essentially assume the local dependence pattern to be known and are therefore not suitable for exploring the local dependence structure.

In this chapter, we extend the multidimensional item response theory model, in particular, the M2PL model for binary data and the MNRM for categorical data. The key feature of the proposed model is to handle high-dimensional responses while maintaining a low-dimensional latent structure. We achieve this by including an additional graphical component to the model to capture the remaining dependence that is not explained by the low-dimensional latent vector. Different from existing methods, our procedure is purely exploratory, in the sense that both the number of latent factors and the graphical structure of local dependence will be selected based on the data. In the proposed model, the dependence among the responses comes from two sources. The responses from a subject share the same latent vector. In addition, the responses are also correlated through a graphical structure. From the inference viewpoint, it is important to separate these two sources of dependence. To do so, we make inference based on the following belief. A low-dimensional latent vector model is largely correct and majority of the dependence among the responses is induced by the common latent vector. There is a small remainder due to the graphical structure. Technical statements of this assumption will be described in Section 3.4.

This modeling framework is related to the problem of decomposing the sum of a sparse matrix and a low-rank matrix, first studied in Candès et al. (2011) and in Chandrasekaran et al. (2011) independently. In their analysis, a convex program is proposed for this matrix decomposition problem and sufficient conditions are given under which the low rank matrix L and sparse matrix S can be exactly recovered when observing $M = L + S$ via the convex program. Their results have many applications, such as face recognition (Candès et al., 2011) and video surveillance (Candès et al., 2011; Bouwmans and Zahzah, 2014). Then Zhou et al. (2010) extends the analysis in Candès et al. (2011) by considering that

$M = L + S + Z$, where Z is a noise term. Chandrasekaran et al. (2012) considers the statistical inference of a multivariate Gaussian model whose precision matrix admits the form of a low-rank matrix plus a sparse matrix, which has been applied to analyze neural circuits (Yatsenko et al., 2015).

The rest of this chapter is organized as follows. In Section 3.2, undirected graphical models for binary and categorical data are introduced. In Section 3.3, the latent and undirected graphical measurement models are proposed. The exploratory analysis based on the proposed models and its statistical properties are presented in Section 3.4 and the computation is discussed in Section 3.5. Suggestions are provided for the selection of tuning parameters in Section 3.6 and finally simulation and real data analysis are presented.

3.2 Undirected Graphical Models for Binary/Categorical Responses

A building block of the proposed models is the undirected graphical models, also called the Markov random field (Lauritzen, 1996; Pearl, 1988). We begin by describing some useful notations about undirected graphs. A graph $G = (V, E)$ is formed by a collection of vertices $V = \{1, 2, ..., J\}$, and a collection of edges $E \subset V \times V$. Each edge consists of a pair of vertices $i, j \in E$ and a graph is undirected in sense that $(i, j) \in E$ if and only if $(j, i) \in E$. Two vertices i and j are connected in graph G if there exist vertices $v_1, v_2, ..., v_T \in V$, such that $(i, v_1), (v_1, v_2), ..., (v_{T-1}, v_T), (v_T, j) \in E$; in other words, if there exists a path $(i, v_1, v_2, ..., v_T, j)$ connecting vertices i and j. In addition, we say the path passes vertex set $C \subset V$ if there exists $v_t \in \{v_1, v_2, ..., v_T\}$, such that $v_t \in C$. Furthermore, for mutually disjoint subsets $A, B, C \subset V$, we say A and B are separated by C if any path from a vertex in A to a vertex in B passes C.

We then discuss undirected graphical models. In order to define an undirected graphical model, we associate with each vertex $j \in V$ a random variable Y_j. In addition, let $\mathbf{Y}_A = \{Y_j : j \in A\}$ be the random vector associated with a vertex set $A \subset V$. We write $\mathbf{Y}_A \perp \mathbf{Y}_B | \mathbf{Y}_C$, if random vectors $\mathbf{Y}_A$ and $\mathbf{Y}_B$ are independent given $\mathbf{Y}_C$, for $A, B, C \subset V$. An undirected graphical model associates the graph separation in G with the conditional

independence relationships among $(Y_j : j \in V)$. More precisely, an undirected graphical model consists of a collection of probability distributions of $(Y_j : j \in V)$ that satisfy:

$$\mathbf{Y}_A \perp \mathbf{Y}_B | \mathbf{Y}_C,$$

if A and B are separated by C, for any mutually disjoint subsets $A, B, C \subset V$. In particular, the pairwise Markov property is satisfied:

$$Y_i \perp Y_j | \mathbf{Y}_{V/\{i,j\}}$$

for $(i, j) \notin E$. An example of an undirected graph is showed in Figure 3.1. Corresponding to this graph, we have several conditional independence relationships, such as

$$Y_1 \perp (Y_4, Y_5) \mid (Y_2, Y_3) \quad \text{and} \quad Y_5 \perp (Y_1, Y_2, Y_3) \mid Y_4.$$

The undirected graphical model provides a flexible framework for modeling the dependence structure of the random variables and the Markov properties bring good interpretation to the dependence structure. In particular, the latent variable models under the local independence assumption discussed in Section 2.1.1 can be viewed as an undirected graphical model for $(\mathbf{Y}, \boldsymbol{\theta})$ with a graphical structure in Figure 3.2, where there is no edge between any pair of Y_i and Y_j.

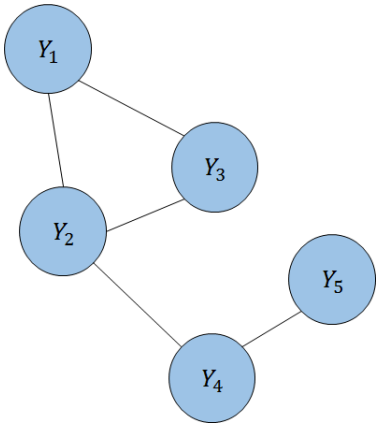

Figure 3.1: An example of an undirected graph.

In what follows, two prototypical undirected graphical models are introduced, including the Ising model (Ising, 1925) for binary data and its extension to categorical data (Wainwright and Jordan, 2008).

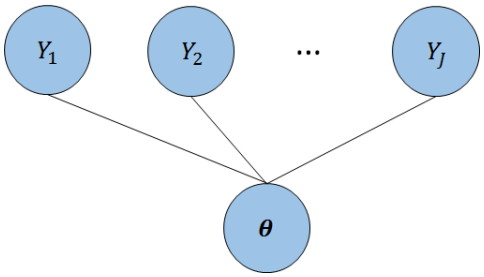

Figure 3.2: The undirected graphical structure of latent variable models under the local independence assumption.

Ising Model (Ising, 1925). We associate a random variable $Y_j \in \{0, 1\}$ to each vertex $j \in V$. The graphical structure encodes the conditional dependence structure among $Y_1, ..., Y_J$. The Ising model parameterizes an undirected graph via the exponential family, admitting the following probability mass function

$$P(\mathbf{Y} = \mathbf{y}) = \frac{1}{z(S)} \exp\left\{\frac{1}{2}\mathbf{y}^\top S\mathbf{y}\right\}, \tag{3.1}$$

where $S = (s_{ij})$ is a J by J symmetric matrix, i.e., $s_{ij} = s_{ji}$, and $z(S)$ is the normalizing constant

$$z(S) = \sum_{\mathbf{y}\in\{0,1\}^J} \exp\left\{\frac{1}{2}\mathbf{y}^\top S\mathbf{y}\right\}. \tag{3.2}$$

The matrix S maps to a graphical structure. There is an edge between vertices i and j, $(i, j) \in E$, if and only if $s_{ij} = s_{ji} \neq 0$. Based on the probability mass function (3.1), it is easy to check that Y_i and Y_j are conditionally independent given all other Y_l, $l \neq i$ or j, if $s_{ij} = 0$.

Categorical variables (Wainwright and Jordan, 2008). In order to generalize the Ising model to categorical response data, we define the indicator vectors

$$\mathbf{1}_{y_j} = (1_{\{y_j=1\}}, ..., 1_{\{y_j=c_j\}})^\top, \text{ and } \mathbf{1}_{\mathbf{y}} = (\mathbf{1}_{y_1}^\top, ..., \mathbf{1}_{y_J}^\top)^\top.$$

Then the generalized Ising (GIsing) model is defined as

$$P(\mathbf{Y} = \mathbf{y}) = \frac{1}{z(S)} \exp\left\{\frac{1}{2}\mathbf{1}_{\mathbf{y}}^\top S\mathbf{1}_{\mathbf{y}}\right\}, \tag{3.3}$$

where $S = (S_{ij})$ is a symmetric block matrix with blocks S_{ij}s, satisfying when $i \neq j$

$$S_{ij} = \begin{pmatrix} s_{ij}^{11} & s_{ij}^{12} & \cdots & s_{ij}^{1c_j} \\ s_{ij}^{21} & s_{ij}^{22} & \cdots & s_{ij}^{2c_j} \\ \vdots & \vdots & & \vdots \\ s_{ij}^{c_i 1} & s_{ij}^{c_i 2} & \cdots & s_{ij}^{c_i c_j} \end{pmatrix}$$

and when $i = j$

$$S_{jj} = \begin{pmatrix} s_{jj}^{11} & 0 & \cdots & 0 \\ 0 & s_{jj}^{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{jj}^{c_j c_j} \end{pmatrix}.$$

S is still symmetric or equivalently

$$S_{ij} = S_{ji}^{\top}$$

Similar to the binary case, the matrix S corresponds to the graphical structure. In particular, there is an edge between vertices i and j, $(i, j) \in E$, if and only if $\|S_{ij}\|_F \neq 0$. Here $\|S_{ij}\|_F = \sqrt{\sum_{l=1}^{c_i} \sum_{m=1}^{c_j} (s_{ij}^{lm})^2}$ is the matrix Frobenius norm. In other words, for the GIsing model, there is no edge between i and j, if and only if S_{ij} is a zero matrix. When $c_j = 1$ for all j, this model is identical to the Ising model.

3.3 Modeling Local Dependence

We consider to model the local dependence structure. The key idea is to combine the latent variable models with undirected graphical models described above. Specifically, we use the M2PL and MNRM models, two flexible and popular multidimensional IRT models, as the latent variable model components for binary and categorical response data respectively. It is possible to replace these two models by others. For example, one may combine the MGPCM with the GIsing model for polytomous response data.

M2PL-Ising Model. When the responses are binary, we consider to combine the M2PL model with the Ising model. To do so, we present the M2PL model in (2.3) as

$$P(Y_j = y_j|\boldsymbol{\theta}) = \frac{e^{(d_j+\mathbf{a}_j^\top\boldsymbol{\theta})y_j}}{1+e^{d_j+\mathbf{a}_j^\top\boldsymbol{\theta}}} \propto e^{(d_j+\mathbf{a}_j^\top\boldsymbol{\theta})y_j}.$$

Under the local independence assumption, the joint conditional distribution is

$$P(\mathbf{Y}=\mathbf{y}|\boldsymbol{\theta}) \propto \exp\Big\{\sum_{j=1}^{J}(d_j+\mathbf{a}_j^\top\boldsymbol{\theta})y_j\Big\} = \exp\Big\{\boldsymbol{\theta}^\top A^\top\mathbf{y}+\mathbf{d}^\top\mathbf{y}\Big\},$$

where $A=(\mathbf{a}_1,...,\mathbf{a}_J)=(a_{jk})_{J\times K}$ and $\mathbf{d}=(d_1,...,d_J)^\top$. In the above representation, the probability mass function of the Ising model in (3.1) can be similarly written as

$$P(\mathbf{Y}=\mathbf{y}) \propto \exp\Big\{\frac{1}{2}\mathbf{y}^\top S\mathbf{y}\Big\}.$$

To model local dependence, the key idea is to model the conditional distribution of $\mathbf{Y}$ given $\boldsymbol{\theta}$ as an Ising model. Specifically, we combine these two models and write

$$P(\mathbf{Y}=\mathbf{y}|\boldsymbol{\theta}) \propto \exp\Big\{\boldsymbol{\theta}^\top A^\top\mathbf{y}+\frac{1}{2}\mathbf{y}^\top S\mathbf{y}\Big\}. \tag{3.4}$$

We remove the term $\mathbf{d}^\top\mathbf{y}$ because it is absorbed into the squared terms of $\frac{1}{2}\mathbf{y}^\top S\mathbf{y}$. Notice that $y_j \in \{0,1\}$ and thus $y_j = y_j^2$. The squared terms in the (3.4) becomes linear $\sum_{j=1}^{J} s_{jj}y_j^2 = \sum_{j=1}^{J} s_{jj}y_j$. We further impose a prior distribution $f(\cdot)$ on $\boldsymbol{\theta}$ such that the joint distribution of $(\mathbf{Y},\boldsymbol{\theta})$ given the parameters (A,S) is

$$p(\mathbf{y},\boldsymbol{\theta}|A,S) = P(\mathbf{Y}=\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \propto \exp\Big\{-\frac{1}{2}\|\boldsymbol{\theta}\|^2+\boldsymbol{\theta}^\top A^\top\mathbf{y}+\frac{1}{2}\mathbf{y}^\top S\mathbf{y}\Big\}, \tag{3.5}$$

where $\|\cdot\|$ is the vector Euclidian norm. We name this model as the *M2PL-Ising model*.

Remark 1. *Throughout this chapter, we frequently use the notation "$\propto$" to define probability density or mass functions. It means that left-hand side and the right-hand side are different by a factor that depends only on the parameters (behind " $|$") and is free of the value of the random variable. The constant can be obtained by summing or integrating out the random variable. Such a constant sometimes could be difficult to evaluate, which will be discussed in the sequel.*

Define the normalizing constant

$$z(A,S) = \sum_{\mathbf{y}\in\{0,1\}^J} \int_{\mathbb{R}^K} \exp\Big\{ -\frac{1}{2}\|\boldsymbol{\theta}\|^2 + \boldsymbol{\theta}^\top A^\top \mathbf{y} + \frac{1}{2}\mathbf{y}^\top S\mathbf{y}\Big\} d\boldsymbol{\theta}.$$

The joint distribution of a single observation $(\mathbf{y}, \boldsymbol{\theta})$ is

$$p(\mathbf{y}, \boldsymbol{\theta}|A,S) = \frac{1}{z(A,S)} \exp\Big\{ -\frac{1}{2}\|\boldsymbol{\theta}\|^2 + \boldsymbol{\theta}^\top A^\top \mathbf{y} + \frac{1}{2}\mathbf{y}^\top S\mathbf{y}\Big\}. \tag{3.6}$$

The normalizing constant $z(A,S)$ is not easy to compute and thus the evaluation of the above likelihood is not straightforward. We will address this issue momentarily.

Both the M2PL model and the Ising model are special cases of (3.4). By setting $a_{jk} = 0$, (3.4) recovers the Ising model with parameter matrix S; by setting $s_{ij} = 0$ for $i \neq j$, (3.4) is equivalent to the M2PL model. Furthermore, conditional on $\boldsymbol{\theta}$, $\mathbf{Y}$ follows an Ising model. In particular,

$$P(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta}) \propto \exp\Big\{\frac{1}{2}\mathbf{y}^\top S(\boldsymbol{\theta})\mathbf{y}\Big\}$$

where $s_{ij}(\boldsymbol{\theta}) = s_{ij}$ for $i \neq j$ and $s_{jj}(\boldsymbol{\theta}) = s_{jj} + 2\mathbf{a}_j^\top \boldsymbol{\theta}$. $S(\boldsymbol{\theta})$ maps to the same graph as S. Thus, the graphical structure S captures the remaining dependence that is not explained by the latent vector. The graphical structure of $(\mathbf{Y}, \boldsymbol{\theta})$ is illustrated in Figure 3.3, where the blue edges correspond to the local dependence structure described by S and the black edges correspond to the rows of the loading matrix A.

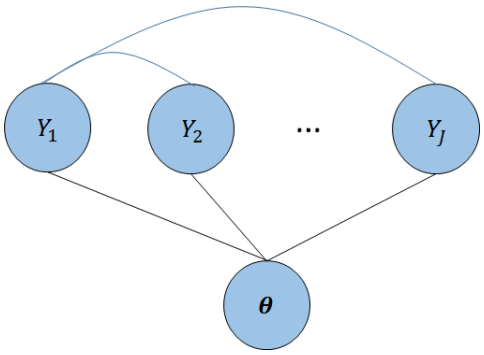

Figure 3.3: The undirected graphical structure of the proposed latent variable and undirected graphical model.

Lastly, we consider the marginal joint distribution of $\mathbf{Y}$ with the latent vector $\boldsymbol{\theta}$ inte-

grated out. More precisely,

$$p(\mathbf{y}|A,S) = P(\mathbf{Y}=\mathbf{y}) = \int_{\mathbb{R}^K} p(\mathbf{y},\boldsymbol{\theta}|A,S)d\boldsymbol{\theta} = \frac{(2\pi)^{K/2}}{z(A,S)}\exp\Big\{\frac{1}{2}\mathbf{y}^\top(AA^\top+S)\mathbf{y}\Big\}. \quad (3.7)$$

The latent vector $\boldsymbol{\theta}$ is not directly observed. Our subsequent analysis is mostly based on the above marginal likelihood. Notice that the loading matrix A enters the likelihood function $p(\mathbf{y}|A,S)$ in the form of $AA^\top$. Therefore, A is not identifiable by itself. We reparameterize and define $L = AA^\top$. We slightly abuse the notation and write

$$p(\mathbf{y}|L,S) = \frac{1}{z(L,S)}\exp\Big\{\frac{1}{2}\mathbf{y}^\top(L+S)\mathbf{y}\Big\}.$$

This is mostly because the latent vector $\boldsymbol{\theta}$ is not directly observed and its loading matrix A can only be identified up to a non-degenerate transformation. One may also notice that there also exists an identifiability issue between L and S. These two matrices enters the marginal likelihood function in the form of $L+S$. In particular, the matrix L characterizes the dependence among $\mathbf{Y}$ that is due to the latent structure and S characterizes that of the graphical structure. In the analysis, assumptions will be imposed on the parameter space so that L and S are separable from each other based on the data and the prior knowledge that L is low rank and S is sparse imposed by the L_1 and nuclear norm regularization.

MNRM-GIsing Model. The same idea can be applied to model categorical response data. Let

$$A_j = \begin{pmatrix} a_{j1}^1 & a_{j2}^1 & \cdots & a_{jK}^1 \\ a_{j1}^2 & a_{j2}^2 & \cdots & a_{jK}^2 \\ \vdots & \vdots & \cdots & \vdots \\ a_{j1}^{c_j} & a_{j2}^{c_j} & \cdots & a_{jK}^{c_j} \end{pmatrix}_{c_j\times K} \quad \text{and} \quad A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_J \end{pmatrix}_{(\sum_{j=1}^J c_j)\times K}.$$

The MNRM model in (2.4) can be presented as

$$P(Y_j = y_j|\boldsymbol{\theta}) \propto \exp\Big\{\boldsymbol{\theta}^\top A_j^\top \mathbf{1}_{y_j} + \mathbf{d}_j^\top \mathbf{1}_{y_j}\Big\},$$

where $\mathbf{d}_j = (d_j^1, ..., d_j^{c_j})^\top$. Under the local independence assumption, the joint conditional distribution is

$$P(\mathbf{Y}=\mathbf{y}|\boldsymbol{\theta}) \propto \exp\Big\{\boldsymbol{\theta}^\top A^\top \mathbf{1}_{\mathbf{y}} + \mathbf{d}^\top \mathbf{1}_{\mathbf{y}}\Big\},$$

where $\mathbf{d} = (\mathbf{d}_1^\top, ..., \mathbf{d}_J^\top)^\top$. We combine the MNRM and GIsing models and write

$$P(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta}) \propto \exp\Big\{\boldsymbol{\theta}^\top A^\top \mathbf{1}_\mathbf{y} + \frac{1}{2}\mathbf{1}_\mathbf{y}^\top S \mathbf{1}_\mathbf{y}\Big\}, \tag{3.8}$$

where the S matrix is defined in (3.3). Similar to the treatment in the M2PL-Ising model, we remove the term $\mathbf{d}^\top \mathbf{1}_\mathbf{y}$ and impose a prior distribution $f(\cdot)$ on $\boldsymbol{\theta}$ such that the joint distribution of $(\mathbf{Y}, \boldsymbol{\theta})$ given the parameters (A, S) is

$$p(\mathbf{y}, \boldsymbol{\theta}|A, S) = P(\mathbf{Y} = \mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \propto \exp\Big\{ -\frac{1}{2}\|\boldsymbol{\theta}\|^2 + \boldsymbol{\theta}^\top A^\top \mathbf{1}_\mathbf{y} + \frac{1}{2}\mathbf{1}_\mathbf{y}^\top S \mathbf{1}_\mathbf{y}\Big\}.$$

We name this model the MNRM-GIsing model. In particular, (3.8) becomes the MNRM model by setting S_{ij} to be zero matrices for all $i \neq j$ and (3.8) becomes the GIsing model by setting A to be a zero matrix. Furthermore, conditional on $\boldsymbol{\theta}$, $\mathbf{Y}$ follows the GIsing model. Slightly different from the binary case, the graphical structure of local dependence in this model is no longer captured by single entries of S. Instead, it is characterized by the blocks of S:

$$(i, j) \in E \quad \text{if and only if} \quad \|S_{ij}\|_F = \|S_{ji}\|_F \neq 0.$$

Lastly, the marginal distribution of $\mathbf{Y}$ under the proposed model can be written as

$$p(\mathbf{y}|L, S) = P(\mathbf{Y} = \mathbf{y}) = \frac{1}{z(L, S)} \exp\Big\{\frac{1}{2}\mathbf{1}_\mathbf{y}^\top (L + S)\mathbf{1}_\mathbf{y}\Big\}, \tag{3.9}$$

where $L = AA^\top$. In this model, both L and S are $C \times C$ matrices ($C = \sum_{j=1}^J c_j$). When $c_j = 1$ for all j, this model becomes identical to the binary model above.

3.4 Estimation

In this section, we address issues concerning the estimation of the M2PL-Ising and MNRM-GIsing models described above, including evaluation of the likelihood function, dimension estimation/reduction of the latent vector, estimation of the graphical structure, the identifiability of parameters, and lastly oracle property of the proposed estimator. We will focus on the analysis of the M2PL-Ising model and the results can be extended to the MNRM-GIsing model.

3.4.1 Regularized Pseudo-likelihood Estimators

M2PL-Ising Model. To begin with, we assume that all of the parameters including the dimension of the latent vector $\boldsymbol{\theta}$ are unknown. The unknown parameters are L and S.

The first issue concerning the estimation is that the evaluation of the marginal likelihood function (3.7) involves the normalizing constant $z(A, S)$ whose computational complexity grows exponentially fast with the dimension J. For a reasonable dimension J, the computation of $z(A, S)$ is practically infeasible. We take a slightly different approach by considering the conditional likelihood of Y_j given $\mathbf{Y}_{-j}$. Let $L = (l_{ij})$ and $S = (s_{ij})$. Based on the marginal likelihood function (3.7), the conditional distribution of Y_j given $\mathbf{Y}_{-j}$ admits the form of a logistic regression model

$$P(Y_j = 1|\mathbf{Y}_{-j} = \mathbf{y}_{-j}, L, S) = \frac{\exp\{\frac{l_{jj}+s_{jj}}{2} + \sum_{i\neq j}(l_{ij} + s_{ij})y_i\}}{1 + \exp\{\frac{l_{jj}+s_{jj}}{2} + \sum_{i\neq j}(l_{ij} + s_{ij})y_i\}}. \tag{3.10}$$

This closed form conditional distribution is crucial for our inference. We let

$$\mathcal{L}_j(L, S; \mathbf{y}) \triangleq P(Y_j = y_j|\mathbf{Y}_{-j} = \mathbf{y}_{-j}, L, S) \tag{3.11}$$

be the conditional likelihood for Y_j given $\mathbf{Y}_{-j}$. Our estimation is based on a pseudo-likelihood function by multiplying all the conditional likelihood together. We use subscript i to denote independent observations. The pseudo-likelihood of N independent observations is

$$\mathcal{L}(L, S) = \prod_{i=1}^{N}\prod_{j=1}^{J}\mathcal{L}_j(L, S; \mathbf{y}_i), \tag{3.12}$$

where $\mathbf{y}_i$ is observation i.

In the above pseudo-likelihood, L and S are the unknown parameters. Besides, the dimension of the latent vector and the conditional graphical structure implied by S are both unknown. We will estimate the set of edges E simultaneously along with S. As for the dimension of $\boldsymbol{\theta}$, to ensure identifiability, we assume that the loading matrix A is of full column rank; otherwise, we can always reduce the dimension K and make A full column rank. Thus, $L = AA^\top$ also has rank K. Notice that L is a positive semidefinite matrix, denoted by $L \succeq 0$. The rank of L is the same as the number of its non-zero eigenvalues. To estimate the conditional graph corresponding to S and the dimensional of the latent vector, we impose regularization on the off-diagonal entries of S and the eigenvalues of L.

As mentioned previously, the parameters L and S enter the likelihood function in the form of $L+S$. Thus, one cannot identify L from S based on the data only. We will impose further assumptions to ensure their identifiability based on the following rationale. We are in the situation that the M2PL model (with the local independence assumption) is largely correct. The latent vector accounts for most dependence/variation of the multivariate observation $\mathbf{Y}$. In the context of cognitive assessment, this is interpreted as that a person's responses to items are mostly driven by a few latent attributes. The remaining dependence is rather low. Thus, a crucial assumption in our estimation is that the graphical structure explains a small portion of the dependence in $\mathbf{Y}$. To quantify this assumption, we assume that the matrix S is sparse. In addition, the dimension of the latent vector stays low.

Based on the above discussion, we propose an estimator by optimizing a regularized pseudo-likelihood

$$(\hat{L}, \hat{S}) = \arg\min_{L,S} \Big\{ -\frac{1}{N} \log\{\mathcal{L}(L,S)\} + \gamma \sum_{i \neq j} |s_{ij}| + \delta \|L\|_* \Big\} \tag{3.13}$$

where $\mathcal{L}(L,S)$ is defined as in (3.12). The above optimization is subject to the constraint that L is positive semidefinite and S is symmetric. This optimization problem is convex and we will discuss the computation in Section 3.5. The tuning parameters γ and δ are positive and control the levels of regularization of the sparsity of S and the rank of L respectively. The estimators $\hat{L}$ and $\hat{S}$ depend on γ and δ. To simplify notation, we omit the indices γ and δ in the notation $\hat{L}$ and $\hat{S}$.

The first regularization term $\sum_{i \neq j} |s_{ij}|$ penalizes on the number of nonzero s_{ij}'s that is the same as the number of edges in the conditional Ising model. By increasing the regularization parameter γ, the number of nonzero off-diagonal elements decreases and thus the number of edges in the conditional graph also decreases. The L_1 penalty was originally proposed in Tibshirani (1996) and later in the context of graphical models (e.g. Meinshausen and Bühlmann, 2006; Friedman et al., 2008; Ravikumar et al., 2010). Notice that we do not penalize the diagonal terms in S because s_{jj} controls the marginal distribution of Y_j. As mentioned previously, the constant term d_j in the M2PL model is also absorbed in s_{jj}.

The second term $\|L\|_* \triangleq \text{Trace}(L)$ denotes the nuclear norm of a positive semidefinite matrix. Notice that $L = AA^\top$ is a positive semidefinite matrix and admits the following

eigendecomposition

$$L = T^\top \Lambda T$$

where T is an orthonormal matrix and $\Lambda = \text{diag}\{\lambda_1, ..., \lambda_J\}$ and $\lambda_j \geq 0$. The nuclear norm can be alternatively represented as

$$\|L\|_* = \sum_{j=1}^{J} |\lambda_j|.$$

Therefore, $\|L\|_*$ penalizes the number of nonzero eigenvalues of L, which is the same as the rank of L. This regularization is first proposed by Fazel et al. (2001) and its statistical properties have been studied in Bach (2008).

The regularized estimators $\hat{L}$ and $\hat{S}$ naturally yield estimators of the dimension of $\boldsymbol{\theta}$ and the conditional graph E. In particular, an estimator of the dimension of $\boldsymbol{\theta}$ is

$$\hat{K} = \text{rank}(\hat{L}) \tag{3.14}$$

and an estimator of the conditional graph is

$$\hat{E} = \{(i, j) : \hat{s}_{ij} \neq 0\}. \tag{3.15}$$

MNRM-GIsing Model. For the MNRM-GIsing model, we can similarly define the pseudo-likelihood $\mathcal{L}(L, S)$ as in (3.12), where both the L and S matrices are $C \times C$ ($C = \sum_{j=1}^{J} c_j$) and the conditional likelihood terms are obtained based on the joint marginal likelihood (3.9). Following the same rationale as above, we assume that L is a low rank matrix and S corresponds to a sparse graph. Specifically, S is block-wise sparse, meaning that there are a small number of (i, j) pairs that $\|S_{ij}\|_F \neq 0$. To recover the block-wise sparse structure, we consider to replace the L_1 (Lasso) regularization by the group Lasso regularization on S, because the L_1 regularization tends to make decisions based on the strength of individual entries, not a group of entries. In particular, the group Lasso regularization on S is

$$\sum_{i \neq j} \sqrt{c_i c_j} \|S_{ij}\|_F,$$

where $\|S_{ij}\|_F$ is the matrix Frobenius norm and the term $\sqrt{c_i c_j}$ terms account for the varying block sizes. The grouping effect comes from treating S_{ij} as a group in $\|S_{ij}\|_F$. Group Lasso

regularization was first proposed in Yuan and Lin (2006) and has been applied in learning graphical models (e.g. Ravikumar et al., 2010). In particular, our estimator will be based on the following regularized pseudo-likelihood,

$$(\hat{L}, \hat{S}) = \arg\min_{L,S} \Big\{ -\frac{1}{N} \log\{\mathcal{L}(L,S)\} + \gamma \sum_{i \neq j} \sqrt{c_i c_j} \|S_{ij}\|_F + \delta \|L\|_* \Big\}, \tag{3.16}$$

where the optimization is subject to the constraint that L is positive semidefinite and S is symmetric and γ and δ are two tuning parameters that control the low-rankness and sparsity. This is still a convex optimization problem.

The regularization term $\sum_{i \neq j} \|S_{ij}\|_F$ penalizes on the number of nonzero S_{ij} blocks that is the same as the number of edges in the conditional GIsing model. By increasing the regularization parameter γ, the number of nonzero off-diagonal blocks decreases, implying that the conditional graph becomes more sparse. The theoretical properties of this regularized estimator are omitted here, as we believe that the theory for the binary case can be extended to this model.

3.4.2 Properties of the Estimator

In this subsection, we presents the properties of the regularized estimator $(\hat{L}, \hat{S})$ defined in (3.13) as wells as the estimators in (3.14) and (3.15). Throughout the discussion, let L^* and S^* denote the true model parameters. For a differentiable manifold $\mathcal{M}$, we let $T_x\mathcal{M}$ denote its tangent space at $x \in \mathcal{M}$. We need the following concepts.

The pseudo-likelihood (and the likelihood) function depends on L and S through $L+S$. Define

$$h(L+S) = -\frac{1}{N} \log \mathcal{L}(L,S). \tag{3.17}$$

If we reparameterize $M = L + S$, its information associated with the pseudo-likelihood is given by

$$\mathcal{I}^* = \mathbb{E}\Big\{ \frac{\partial^2 h}{\partial^2 M}\Big|_{M=M^*} \Big\}$$

where $M^* = L^* + S^*$ is the true parameter matrix. Our first condition concerns this information matrix.

A1 The matrix $\mathcal{I}^*$ is positive definite restricted to the set

$$\mathcal{M} \triangleq \{M = L + S : L \succeq 0 \text{ and } S = S^\top\},$$

that is, for each vector $v \in \mathcal{M}$, $v^\top \mathcal{I}^* v \geq 0$ and the equality holds if and only if $v = 0$.

Let $\mathcal{S}^*$ be the set of symmetric matrix admitting the same sparsity as that of S^*, that is,

$$\mathcal{S}^* = \{S : S \text{ is a } J \times J \text{ symmetric matrix and } s_{ij} = 0 \text{ if } s^*_{ij} = 0 \text{ for all } i \neq j\}.$$

On considering that $\mathcal{S}^*$ is a submanifold on $\mathbb{R}^{J \times J}$, its tangent space at S^* is $\mathcal{S}^*$ itself, that is,

$$T_{S^*}\mathcal{S}^* = \mathcal{S}^*.$$

We refer to Lee (2009) for the definition of manifolds and their tangent spaces.

Define the set of matrices

$$\mathfrak{L} = \{L : L \text{ is positive semidefinite and } \text{rank}(L) \leq K\},$$

where $K = \text{rank}(L^*)$. The set $\mathfrak{L}$ is differentiable in a neighborhood of L^*. To define the tangent space of $\mathfrak{L}$ at L^*, we consider its eigendecomposition

$$L^* = U_1^* D_1^* U_1^{*\top}, \tag{3.18}$$

where U_1^* is a $J \times K$ matrix satisfying $U_1^{*\top} U_1^* = I_K$, I_K is the $K \times K$ identity matrix, and D_1^* is a $K \times K$ diagonal matrix consisting of the (positive) eigenvalues of L^*. Then, the tangent space of $\mathfrak{L}$ at L^* is

$$T_{L^*}\mathfrak{L} = \{U_1^* Y + Y^\top U_1^{*\top} : Y \text{ is a } K \times J \text{ matrix}\}.$$

Define a linear operator $\mathbf{F} : \mathcal{S}^* \times T_{L^*}\mathfrak{L} \to \mathcal{S}^* \times T_{L^*}\mathfrak{L}$,

$$\mathbf{F}(S, L) = (\mathbf{P}_{\mathcal{S}^*}\{\mathcal{I}^*(S + L)\}, \mathbf{P}_{T_{L^*}\mathfrak{L}}\{\mathcal{I}^*(S + L)\}).$$

In the above definition, $\mathcal{I}^*$ is a $J^2 \times J^2$ matrix. We slightly abuse the notation and let $\mathcal{I}^*(S + L)$ denote matrix-vector multiplication where S and L are vectorized and their elements are arranged in the same order as the order of the derivatives of $\mathcal{I}^*$. The map

$\mathbf{P}_{\mathcal{M}}(A)$ is the projection operator of matrix A on to the manifold $\mathcal{M}$ with respect to the inner product for matrices,

$$A \cdot B = \sum_{j=1}^{J} \sum_{l=1}^{J} A_{jl} B_{jl} = \text{Trace}(AB^{\top}).$$

That is, $\mathbf{P}_{\mathcal{M}}(A)$ is the matrix in $\mathcal{M}$ minimizing the distance to A induced by the matrix inner product "$\cdot$". We make the following assumptions on L^*, $\mathcal{S}^*$, and $\mathcal{L}^*$.

A2 The positive eigenvalues of L^* are distinct.

A3 The linear operator $\mathbf{F}$ is a bijective mapping. In other words,

$$\mathbf{F}^{-1} : \mathcal{S}^* \times T_{L^*}\mathfrak{L} \to \mathcal{S}^* \times T_{L^*}\mathfrak{L}$$

is a well defined linear mapping such that $\|\mathbf{F}^{-1}\| < \infty$. Here $\|\cdot\|$ is a norm for linear operators in $\mathbf{R}^{J\times J} \times \mathbf{R}^{J\times J} \to \mathbf{R}^{J\times J} \times \mathbf{R}^{J\times J}$. Notice that all norms for finite dimensional space are topologically equivalent.

A4 The intersection between $\mathcal{S}^*$ and $T_{L^*}\mathfrak{L}$ is trivial, that is, $\mathcal{S}^* \cap T_{L^*}\mathfrak{L} = \{\mathbf{0}_{J\times J}\}$, where $\mathbf{0}_{J\times J}$ is the $J \times J$ zero-matrix. This is the so-called transversality condition.

In addition, we consider a subspace of $T_{L^*}\mathfrak{L}$,

$$\mathcal{D}^* = \{U_1^* D_1' U_1^{*\top} : D_1' \text{ is a } K \times K \text{ diagonal matrix}\},$$

and the operator $\widetilde{\mathbf{F}} : \mathcal{S}^* \times \mathcal{D}^* \to \mathcal{S}^* \times \mathcal{D}^*$,

$$\widetilde{\mathbf{F}}(S', L') = (\mathbf{P}_{\mathcal{S}^*}\{\mathcal{I}^*(S' + L')\}, \mathbf{P}_{\mathcal{D}^*}\{\mathcal{I}^*(S' + L')\}), \tag{3.19}$$

for $S' \in \mathcal{S}^*$ and $L' \in \mathcal{D}^*$.

A5 The linear operator $\widetilde{\mathbf{F}}$ is a bijective mapping over $\mathcal{S}^* \times \mathcal{D}^*$.

Lastly, we present a condition that is similar to the irrepresentable condition for the linear model in (Zhao and Yu, 2006; Jia and Yu, 2010). We define a linear operator $\mathbf{F}^{\perp} : \mathcal{S}^* \times \mathcal{L}^* \to \mathcal{S}^{*\perp} \times \mathcal{L}^{*\perp}$,

$$\mathbf{F}^{\perp}(S', L') = (\mathbf{P}_{\mathcal{S}^{*\perp}}\{\mathcal{I}^*(S' + L')\}, \mathbf{P}_{(T_{L^*}\mathfrak{L})^{\perp}}\{\mathcal{I}^*(S' + L')\}).$$

For a linear subspace $\mathcal{M}$, $\mathcal{M}^{\perp}$ denotes its orthogonal complement in $\mathbb{R}^{J\times J}$. For a matrix $A = (a_{ij})$, we let the sign function apply to each of its element, that is

$$\text{sign}(A) = (\text{sign}(a_{ij})).$$

Furthermore, for each constant $\rho > 0$, define norm for a matrix couple (A, B) of appropriate dimensions such that

$$\|(A, B)\|_{\rho} = \max(\|A\|_{\infty}, \|B\|_2/\rho),$$

where $\|\cdot\|_{\infty}$ and $\|\cdot\|_2$ are the maximum and spectral norm respectively. The last conditions is stated as follows.

A6 There exists a positive constant ρ such that

$$\left\|\mathbf{F}^{\perp}\mathbf{F}^{-1}(\text{sign}(\mathbf{O}(S^*)), \rho U_1^* U_1^{*\top})\right\|_{\rho} < 1, \tag{3.20}$$

where $\mathbf{O}(S)$ is a $J \times J$ matrix such that it is identical to S except that its diagonal entries are all zero, that is, $\tilde{S} = (\tilde{s}_{ij}) = \mathbf{O}(S)$ where $\tilde{s}_{ij} = s_{ij}$ for $i \neq j$ and $\tilde{s}_{ii} = 0$

With these conditions, we are able to present the theoretical properties of our estimator, the proof of which is contained in Section 3.8.

Theorem 1. *Under the Assumptions A1-A6, if the tuning parameters γ_N and δ_N decrease at the rate $O(N^{-.5+\eta})$ for some sufficiently small positive constant η, and $\delta_N = \rho\gamma_N$ with ρ satisfying* (3.20)*, then the optimization problem* (3.13) *has a unique solution $(\hat{S}, \hat{L})$ that converges in probability to the true parameter (S^*, L^*). In addition, $(\hat{S}, \hat{L})$ recovers the sparse and low rank structure of (S^*, L^*) with a probability tending to* 1*, that is;*

$$\lim_{N\to\infty} P\Big\{\text{sign}(\mathbf{O}(\hat{S})) = \text{sign}(\mathbf{O}(S^*)), \text{rank}(\hat{L}) = \text{rank}(L^*)\Big\} = 1.$$

Remark 2. *We briefly discuss these assumptions. Assumption A4 is reflecting our belief that majority of the dependence among the responses is induced by the common latent vector and there is just a small remainder due to the graphical structure. Mathematically, if $L^* = A^*(A^*)^{\top}$ is a low rank matrix that is not sparse in a certain sense and S^* is a sparse matrix, their tangent spaces tend to be transversal. To see this, we introduce two quantities that characterize L^* and S^*. Let*

$$\xi = \max_j \|\mathbf{P}_{C(A^*)}e_j\| \quad \textit{and} \quad \mu = \max_j \|(s_{j1}^*, ..., s_{jJ}^*)^{\top}\|_0,$$

where $C(A^*)$ is the column space of L^*, e_j is the jth standard basis vector of $\mathbb{R}^J$, and $\|\cdot\|_0$ is the L_0 norm that counts the number of nonzero components in a vector. ξ is more likely to be small when K is small. In particular, the the minimal value ξ can take is $\sqrt{\frac{K}{J}}$. For a given K, ξ is small when the column space of A^* is not closely aligned with any of the coordinate axes, or equivalently there does not exist a linear combination of the latent factors that associates with only a small number of items. In other words, ξ is small when the responses are driven by a small number of latent factors and the effect of the factors spread out to the items. Furthermore, μ is roughly the maximal degree of the graph induced by S^*. A small μ implies that S^* is sparse, which happens when the latent vector captures majority of the dependence and leaving the conditional Ising model sparse. The smaller both ξ and μ are, the more transversal the tangent spaces of L^* and S^* are. In particular, the two spaces are transversal under the sufficient condition that $\xi\mu \leq \frac{1}{2}$ (Chandrasekaran et al., 2011). Besides A4, all other assumptions involve the Fisher information of the pseudo-likelihood. In particular, assumption A1 ensures the identifiability of $M = L + S$ based on the pseudo-likelihood and A6 is essentially a condition that is similar to the irrepresentable condition for the regularized estimator for the linear model (Zhao and Yu, 2006; Jia and Yu, 2010).

3.5 Computation

In this section, we discuss the computation of the regularized estimator in (3.13) and the algorithm can be slightly modified to solve (3.16). This optimization problem is nontrivial for two reasons. First, the coordinate-wise descent algorithms (Fu, 1998; Friedman et al., 2007), that are widely used in convex optimization problems with L_1 norm regularization, do not work well in this problem. The coordinate-wise decent algorithms optimize with respect to one parameter at a time while keeping the rest fixed to the current values. These algorithms are computationally fast when each iteration has a closed form update. However, when optimizing (3.13), the coordinate-wise decent update with respect to s_{ij} is not in a closed form (due to the logistic regression form), which greatly slows down the computation. Second, the optimization is constrained on the space where the matrix L is

positive semidefinite. As a consequence, it becomes a semidefinite programming problem, and a standard approach is the interior point methods (Boyd and Vandenberghe, 2004). However, the per-iteration computational cost and memory requirements of an interior point method are prohibitively high for this problem, especially when J is large.

Our method avoids these issues by taking advantage of the special structure of L_1 and nuclear norms by means of the alternating direction method of multiplier (ADMM; Boyd et al., 2011; Glowinski and Marroco, 1975; Gabay and Mercier, 1976). As a consequence, the proposed method is able to solve large problems efficiently. The key idea is to decompose the optimization of (3.13) into subproblems that can be solved efficiently.

Consider the optimization

$$\begin{aligned} &\min_x \{f(x) + g(z)\}, \\ &s.t. \;\; x = z, \end{aligned} \tag{3.21}$$

where f, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are closed and proper convex functions. Both f and g are possibly nondifferentiable. The alternating direction method of multiplier is an iterative algorithm that finds the minimizer of the objective function in (3.21). We define the proximal operator $\mathbf{Prox}_{\lambda,f}$: $\mathbb{R}^n \to \mathbb{R}^n$ as

$$\mathbf{Prox}_{\lambda,f}(v) = \arg\min_x f(x) + \frac{1}{2\lambda}\|x - v\|^2,$$

where $\|\cdot\|$ is the usual Euclidean norm on $\mathbb{R}^n$ and λ is a scale parameter that is a fixed positive constant in the algorithm.

The algorithm starts with some initial values $x^0, z^0, u^0 \in \mathbb{R}^n$. At the $(m+1)$th iteration, (x^m, z^m, u^m) is updated according to the following steps until convergence

Step 1: $x^{m+1} := \mathbf{prox}_{\lambda,f}(z^m - u^m)$;

Step 2: $z^{m+1} := \mathbf{prox}_{\lambda,g}(x^{m+1} + u^m)$;

Step 3: $u^{m+1} := u^m + x^{m+1} - z^{m+1}$.

The convergence properties of the ADMM are summarized in the following theorem from Boyd et al. (2011). Let p^* be the optimal value of the objective function in (3.21).

Theorem 2 (Boyd et al., 2011). *Assume the functions f, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are closed and proper convex functions. In addition, assume the Lagrangian*

$$L(x, z, y) = f(x) + g(z) + y^\top (x - z)$$

has a saddle point; that is there exist (x^*, z^*, y^*), not necessarily unique, for which

$$L(x^*, z^*, y) \leq L(x^*, z^*, y^*) \leq L(x, z, y^*).$$

Then the ADMM has the following convergence properties.

1. *Residual convergence. $x^m - z^m \to 0$ as $m \to \infty$, i.e., the iterates approach feasibility.*
2. *Objective convergence. $f(x^m) + g(z^m) \to p^*$ as $m \to \infty$, i.e., the objective function of the iterates approaches the optimal value.*

The assumption on the Lagrangian $L(x, z, y)$ is mild (see Chapter 5, Boyd and Vandenberghe, 2004). In particular, if (x^*, z^*) and y^* are primal and dual optimal points for the problem (3.21) in which strong duality holds, (x^*, z^*, y^*) forms a saddle-point for the Lagrangian.

We now adapt this algorithm to the optimization of the regularized pseudo-likelihood. In particular, we reparameterize $M = L + S$ and let $x = (M, L, S)$. Define

$$f(x) = \begin{cases} h(M) + \gamma \sum_{i \neq j} |s_{ij}| + \delta \|L\|_* & \text{if } L \succeq 0 \text{ and } S = S^\top; \\ +\infty & \text{otherwise} \end{cases}$$

and

$$g(x) = \begin{cases} 0 & \text{if } M = M^\top \text{ and } M = L + S; \\ +\infty & \text{otherwise.} \end{cases}$$

It is easy to verify that g is a closed and proper convex function. Furthermore, $h(M)$ is convex, since the pseudo-likelihood function is the sum of several log-likelihood functions of the logistic models that are all concave (See Chapter 7, Boyd and Vandenberghe, 2004). Because the L_1 and nuclear norms and the symmetric and positive semidefinite constraints are all convex, $f(x)$ is a convex function. Thanks to continuity, f is proper and closed. The optimization (3.13) can be written as

$$\min_x \ \{f(x) + g(x)\}.$$

We now present each of the three steps of the ADMM algorithm. Let

$$x^m = (M^m, L^m, S^m), \quad z^m = (\tilde{M}^m, \tilde{L}^m, \tilde{S}^m), \quad u^m = (U_M^m, U_L^m, U_S^m).$$

Step 1. We solve $x^{m+1} = \mathbf{prox}_{\lambda,f}(z^m - u^m)$. Due to the additive form of $f(\cdot)$, M^{m+1}, L^{m+1}, and S^{m+1} can be updated separately by proximal operators. More precisely,

$$\begin{aligned} M^{m+1} &= \mathbf{prox}_{\lambda,h}(\tilde{M}^m - U_M^m), \\ L^{m+1} &= \mathbf{prox}_{\lambda,f_1}(\tilde{L}^m - U_L^m), \\ S^{m+1} &= \mathbf{prox}_{\lambda,f_2}(\tilde{S}^m - U_S^m), \end{aligned} \tag{3.22}$$

where the function $h(\cdot)$ is defined in (3.17),

$$f_1(L) = \begin{cases} \delta\|L\|_* & \text{if } L \succeq 0, \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$f_2(S) = \begin{cases} \gamma \sum_{i \neq j} |s_{ij}| & \text{if } S = S^\top, \\ +\infty & \text{otherwise.} \end{cases}$$

We now discuss the evaluation of the three proximal operators. In particular, $\mathbf{prox}_{\lambda,f_1}(\tilde{L}^m - U_L^m)$ and $\mathbf{prox}_{\lambda,f_2}(\tilde{S}^m - U_S^m)$ can be solved in closed forms. More precisely, when $\tilde{L}^m - U_L^m$ and $\tilde{S}^m - U_S^m$ are both symmetric matrices (which is guaranteed when M^0, L^0, S^0, U_M^0, U_L^0, and U_S^0 are chosen to be symmetric),

$$L^{m+1} = T\text{diag}(\Lambda - \lambda\delta)_+ T^\top \tag{3.23}$$

where $\tilde{L}^m - U_L^m = T\Lambda T^\top$ is the eigenvalue decomposition and $\text{diag}(\Lambda - \lambda\delta)_+$ is a diagonal matrix with its jth diagonal element $(\Lambda_{jj} - \lambda\delta)_+$. The operation $(\Lambda_{jj} - \lambda\delta)_+$ is called eigenvalue thresholding. L^{m+1} is guaranteed to be positive-semidefinite according to (3.23).

In addition, S^{m+1} is updated as

$$s_{jj}^{m+1} = (\tilde{S}^m - U_S^m)_{jj}$$

and its off-diagonal entries are obtained by a soft-thresholding operator:

$$s_{ij}^{m+1} = \begin{cases} (\tilde{S}^m - U_S^m)_{ij} - \gamma\lambda & \text{if } (\tilde{S}^m - U_S^m)_{ij} > \gamma\lambda; \\ (\tilde{S}^m - U_S^m)_{ij} + \gamma\lambda & \text{if } (\tilde{S}^m - U_S^m)_{ij} < -\gamma\lambda; \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, solving $\mathbf{prox}_{\lambda,h}(\tilde{M}^m - U_M^m)$ is equivalent to solve J J-dimensional unconstrained convex optimization problems. To see it, we denote

$$M_j = (m_{1j}, ..., m_{Jj})^\top$$

as the jth column of a $J \times J$ matrix M. According to equation (3.10), the conditional likelihood $\mathcal{L}_j(L, S; \mathbf{y}_i)$ in (3.11) is a function of M_j only and we denote it as $\mathcal{L}_j(M_j; \mathbf{y}_i)$. As a result, $\mathbf{prox}_{\lambda,h}(\tilde{M}^m - U_M^m)$ can be decomposed into solving

$$\min_{M_j} -\frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_j(M_j; \mathbf{y}_i) + \frac{1}{2\lambda}\|M_j - (\tilde{M}^m - U_M^m)_j\|^2,$$

for $j = 1, 2, ..., J$, where $(\tilde{M}^m - U_M^m)_j$ represents the jth column of $\tilde{M}^m - U_M^m$. Each problem is a J-dimensional unconstrained convex optimization problem, that can be solved efficiently using a standard solver such as the Broyden-Fletcher-Goldfarb-Shanno method (see e.g. Gentle, 2009).

Step 2. We solve $z^{m+1} = \mathbf{prox}_{\lambda,g}(x^{m+1} + u^m)$. Denote $\bar{M} = M^{m+1} + U_M^m$, $\bar{L} = L^{m+1} + U_L^m$, and $\bar{S} = S^{m+1} + U_S^m$. Then evaluating $\mathbf{prox}_{\lambda,g}(x^{m+1} + u^m)$ becomes:

$$\begin{aligned} \min_{M,L,S} \quad & \frac{1}{2}\|M - \bar{M}\|_F^2 + \frac{1}{2}\|L - \bar{L}\|_F^2 + \frac{1}{2}\|S - \bar{S}\|_F^2 \\ \text{s.t.} \quad & M = M^\top, \\ & M = L + S. \end{aligned}$$

It is a quadratic programming problem with only equality constraints and thus can be solved in a closed form:

$$\begin{aligned} \tilde{M}^{m+1} &= \frac{1}{3}\bar{M} + \frac{1}{3}\bar{M}^\top + \frac{1}{3}\bar{L} + \frac{1}{3}\bar{S}, \\ \tilde{L}^{m+1} &= \frac{2}{3}\bar{L} + \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^\top - \frac{1}{3}\bar{S}, \\ \tilde{S}^{m+1} &= \frac{2}{3}\bar{S} + \frac{1}{6}\bar{M} + \frac{1}{6}\bar{M}^\top - \frac{1}{3}\bar{L}. \end{aligned}$$

Step 3 is a simple arithmetic. The advantage of the proposed algorithm is the low cost of computation and memory in each iteration. In particular, the nonsmooth L_1 and nuclear norm terms and the positive semidefinite constraint that cause trouble to a generic solver are efficiently handled by closed-form updates. In addition, the J^2 dimensional function $h(M)$ is decomposed to be the sum of J J-dimensional functions that can be optimized in parallel. As a consequence, this algorithm is efficient enough to solve large problems.

3.6 On the Choice of Tuning Parameters

Theorem 1 provides a theoretical guideline for choosing the regularization parameters γ and δ. Nonetheless, it leaves quite some room of freedom. In what follows, we provide more specific choices of γ and δ for the simulation study and real data analysis.

We consider to choose γ and δ to optimize model selection (in a certain sense). Information criteria, including the Akaike information criterion (AIC; Akaike, 1973) and Bayes information criterion (BIC; Schwarz, 1978), are widely used for model selection. The BIC is well-known to yield variable selection consistency in the asymptotic scenario in which the sample size N grows large while the number of parameters in the saturated model remains constant, while the AIC may fail because of overfitting (Shao, 1997). BIC is defined as

$$\mathrm{BIC}(\mathcal{M}) = -2 \log L_N(\hat{\beta}(\mathcal{M})) + |\mathcal{M}| \log N,$$

where $\mathcal{M}$ is the current model, $L_N(\hat{\beta}(\mathcal{M}))$ is the maximal likelihood given the model $\mathcal{M}$ and $|\mathcal{M}|$ is the number of free parameters in $\mathcal{M}$. As the evaluation of likelihood function is computationally infeasible for the proposed model, we replace the likelihood function in the BIC by the pseudo-likelihood function. To avoid ambiguity, we change the notation and use $\hat{L}^{\gamma,\delta}$ and $\hat{S}^{\gamma,\delta}$ to denote the estimator in (3.13) corresponding to regularization parameters γ and δ. Let

$$\begin{aligned}\mathcal{M}^{\gamma,\delta} = \{\, & L \text{ is positive semidefinite and } S \text{ is symmetric,} \\ & \mathrm{rank}(L) \le \mathrm{rank}(\hat{L}^{\gamma,\delta}) \text{ and } s_{ij} = 0 \text{ if } \hat{s}_{ij}^{\gamma,\delta} = 0 \text{ for all } i \ne j \,\}\end{aligned}$$

be the submodel selected by tuning parameters (γ, δ). It contained all models whose positive semidefinite matrix L has rank no larger than that of $\hat{L}^{\gamma,\delta}$ and whose symmetric matrix S

has the same support as $\hat{S}^{\gamma,\delta}$. We select the model that optimizes the Bayesian information criterion based on the pseudo-likelihood

$$\text{BIC}(\mathcal{M}^{\gamma,\delta}) = -2 \max_{(L,S)\in M^{\gamma,\delta}} \{\log \mathcal{L}(L,S)\} + |\mathcal{M}^{\gamma,\delta}| \log N. \tag{3.24}$$

When $\text{rank}(\hat{L}^{\gamma,\delta}) = K$, the number of parameters in $\mathcal{M}^{\gamma,\delta}$ is

$$|\mathcal{M}^{\gamma,\delta}| = \left(JK - \frac{(K-1)K}{2}\right) + \sum_{i\le j} 1_{\{\hat{s}_{ij}^{\gamma,\delta}\neq 0\}},$$

where the two terms are the numbers of free parameters in L and S respectively. Specifically, the number of free parameters in L is counted as follows. Let $L = U_1 D_1 U_1^\top$ be the eigendecomposition of L, where D_1 is a $K \times K$ diagonal matrix and columns of U_1 are unit-length eigenvectors of L. D_1 has K parameters and U_1 has $JK - \frac{K(K+1)}{2}$ parameters due to constraint $U_1^\top U_1 = I_K$. Combing them together, L has $JK - \frac{(K-1)K}{2}$ parameters.

Maximizing the pseudo-likelihood in (3.24) is no longer a convex optimization problem. However, it turns out that this nonconvex optimization can be solved stably using a numerical solver, with $(\hat{L}^{\gamma,\delta}, \hat{S}^{\gamma,\delta})$ as the starting point. The tuning parameters are finally selected by

$$(\hat{\gamma}, \hat{\delta}) = \arg\min_{\gamma,\delta} \text{BIC}(\mathcal{M}^{\gamma,\delta}).$$

In addition, the corresponding maximal pseudo-likelihood estimates of L and S are used as the final estimate of L and S:

$$(\hat{L}, \hat{S}) = \arg \max_{(L,S)\in M^{\hat{\delta},\hat{\gamma}}} \{\mathcal{L}(L,S)\}. \tag{3.25}$$

3.7 Simulation and Real Data Analysis

3.7.1 Simulation Study

We consider $J = 30$ items and sample size $N \in \{250, 500, 1000, 2000, 4000\}$ under the following three settings.

1. $K = 1$ latent variable. For the S-matrix, all off-diagonal elements are zero except for $s_{j,j+1}$ for $j = 1, 3, ..., 29$. There are in total 15 edges in the graph. This graph is equivalent to grouping the variables in pairs, {1,2}, {3,4}, ..., and {29, 30}. There is an edge between each pair.

2. $K = 1$ latent variable. For $j = 1, 4, ..., 28$, $s_{j,j+1}$, $s_{j,j+2}$, and $s_{j+1,j+2}$ are nonzero. There are 30 edges in the conditional graph. This is equivalent to grouping the variables in triples, {1,2,3}, {4,5,6}, ..., {28,29,30}. There edges among each triple.

3. $K = 2$, and the conditional graph is the same as situation 1.

The graphical representations corresponding the choices of S are visualized in Figure 3.4, where the upper and lower panels represent the settings of S in settings 1 and 2, respectively. For each model setting and each sample size, the simulation results are based on 50

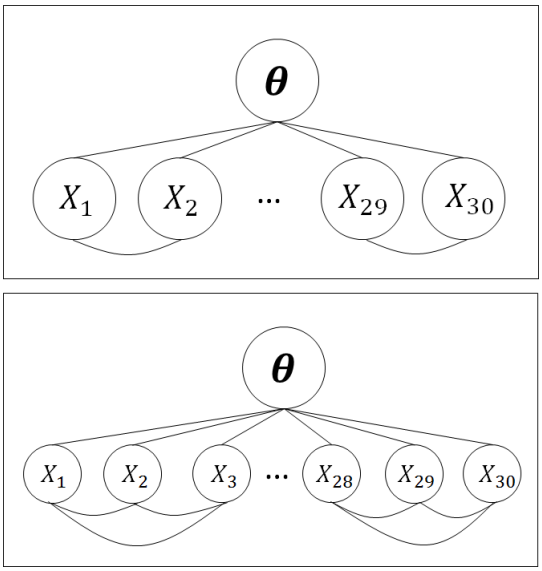

Figure 3.4: The graphical representation of the simulation settings.

independent data sets.

Data generation. To generate a sample from the latent graphical model, we first generate $\boldsymbol{\theta}$ from its marginal distribution

$$f(\boldsymbol{\theta}) \propto \sum_{\mathbf{y}\in\{0,1\}^J} \exp\left\{-\frac{1}{2}\|\boldsymbol{\theta}\|^2 + \mathbf{y}^\top A\boldsymbol{\theta} + \frac{1}{2}\mathbf{y}^\top S\mathbf{y}\right\}.$$

The above summation is computationally feasible because of the sparse graphical structure as in Figure 3.4 and the joint distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$ fall into pairs or triples. Then, $\boldsymbol{\theta}$ is sampled from the above marginal distribution by the Accept/Reject algorithm (see e.g. Chapter 5, Casella and Berger, 2002). The second step is to sample $\mathbf{Y}$ given $\boldsymbol{\theta}$ that is

a sparse Ising model. Once again, due to the simple structure of the graph, we generate conditional samples of $\mathbf{Y}$ in pairs or triples.

Evaluation criteria. To assess the performance of the regularized estimator on dimension reduction and the recovery of the graphical structure, we consider the criterion C_1. For a particular data set, $C_1 = 1$ if and only if there exists a pair of (γ, δ), such that $\text{rank}(\hat{L}^{\gamma,\delta}) = \text{rank}(L^*)$ and graph induced by $\hat{S}^{\gamma,\delta}$ is the same as that by S^*, where L^* and S^* are the true parameters. Furthermore, we evaluate the BIC-based tuning parameter selection method, based on criteria C_2, C_3, and C_4. Let $(\hat{L}, \hat{S})$ be the final estimates of the selected model, defined in (3.25). C_2 evaluates the estimation of the rank of L^*, defined as

$$C_2 = 1_{\{\text{rank}(\hat{L})=\text{rank}(L^*)\}}.$$

In addition, C_3 evaluates the positive selection rate of the graphical structure of S^*, defined as

$$C_3 = \frac{|\{(i,j) : i < j, \hat{s}_{ij} \neq 0, \text{ and } s^*_{ij} \neq 0\}|}{|\{(i,j) : i < j, s^*_{ij} \neq 0\}|}.$$

Furthermore, C_4 evaluates the false discovery rate, defined as

$$C_4 = \frac{|\{(i,j) : i < j, \hat{s}_{ij} \neq 0, \text{ and } s^*_{ij} = 0\}|}{|\{(ij) : i < j, s^*_{ij} = 0\}|}.$$

If the tuning parameter is reasonably selected, we expect that $C_2 = 1$, C_3 is close to 1, and C_4 is close to 0.

In Figure 3.5, the averages of C_1 over 50 replications versus the sample sizes are presented under all settings. Based on Figure 3.5, we observe that as the sample size becomes larger, the probability that the path of regularized estimator captures the true model increases and is close to 1 when the sample size is over 1000. The true model is likely to be missed by the solution path for small sample sizes. This is because the graphical structure is difficult to be captured completely when the sample size is small. On the other hand, the latent factor structure can be stably captured under all these sample sizes.

The results of model selection based on BIC are presented in Table 3.1, where the mean of C_2 and the means and standard errors of C_3, and C_4 over 50 replications are presented. According to these results, the BIC tends to choose a model that is close to the true one. In

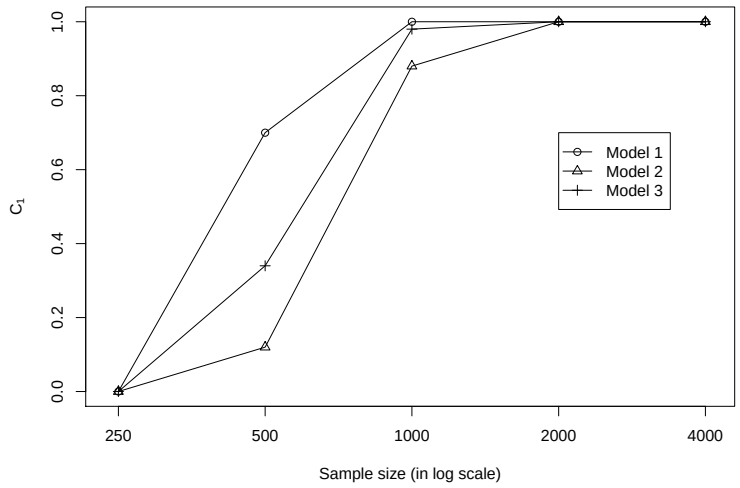

Figure 3.5: The mean of C_1 over 50 replications for all models and all sample sizes.

particular, according to C_2, the number of latent factors (i.e. the rank of L^*) can be stably recovered given a reasonable sample size. Specifically, the numbers of factors are recovered without error for all situations except when $N = 250$ for Model 3. For that case, the BIC may select a one-factor model, which is mainly caused by the small sample size. In addition, the edges in the conditional graph are recovered well according to C_3. Furthermore, based on C_4, a small number of false discoveries are made, meaning there exist edges mistakenly placed when they do not exist. Our result is reasonably well for such a challenging task (selection from 435 possible edges). According to Table 3.1, this issue gradually vanishes when the sample size becomes large.

3.7.2 Real Data Analysis: EPQ-R Data

We analyze the Psychoticism (P), Extraversion (E), and Neuroticism (N) scales of the Eysenck's Personality Questionnaire-Revised (EPQ-R: Eysenck et al., 1985; Eysenck and Barrett, 2013). The dataset contains the responses to 79 items from 824 female respondents from United Kingdom. This is initially a confirmatory analysis with three factors: Psychoticism (P), Extraversion (E), and Neuroticism (N). Among these 79 items, 32, 23, and 24 items are designed to measure the P, E, and N factors, respectively and these three

C_2	$N = 250$	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
Model1	100.0	100.0	100.0	100.0	100.0
Model2	100.0	100.0	100.0	100.0	100.0
Model3	78.0	100.0	100.0	100.0	100.0
C_3	$N = 250$	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
Model1	98.3(3.8)	100.0(0.0)	100.0(0.0)	100.0(0.0)	100.0(0.0)
Model2	92.7(5.1)	98.9(2.2)	100.0(0.0)	100.0(0.0)	100.0(0.0)
Model3	94.3(6.9)	99.6(1.5)	100(0.0)	100.0(0.0)	100.0(0.0)
C_4	$N = 250$	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
Model1	8.4(2.5)	6.7(1.6)	5.0(1.5)	4.3(1.5)	2.8(1.1)
Model2	8.6(2.7)	6.2(2.4)	0.1(0.4)	0.0(0.1)	0.0(0.1)
Model3	6.9(2.4)	5.5(1.4)	2.1(0.7)	0.3(0.3)	0.0(0.0)

Table 3.1: The mean and standard error in percentage (%) of C_2, C_3, and C_4.

sets of items are known as the P, E, and N scales. The specific questions can be found in the appendix of Eysenck et al. (1985). A typical item is "Are you rather lively?" The responses are "yes" or "no", coded as 1 or 0, respectively. Furthermore, the data have been preprocessed so that the negatively worded items are reversely scored (see Table 4 of Eysenck et al., 1985 for the scoring key).

We consider $\gamma \in (0, 0.02]$ and $\rho = \frac{\delta}{\gamma} \in (10, 20]$ and choose 20 grid values for both γ and ρ, so that there are 400 models on the solution path. A summary of the solution path is as follows. Among all 400 models, about 11% of the models have four or more factors, 57% of them are three-factor models, and the rest 32% of them have two or less factors. In addition, we define the graph sparsity level (GSP) of an estimated model as the estimated number of edges normalized by the total number of possible edges. A histogram of the GSP for all models on the path is presented in Figure 3.6. Furthermore, Figure 3.7 presents a box plot showing the number of factors - GSP relationship for these models, where the y

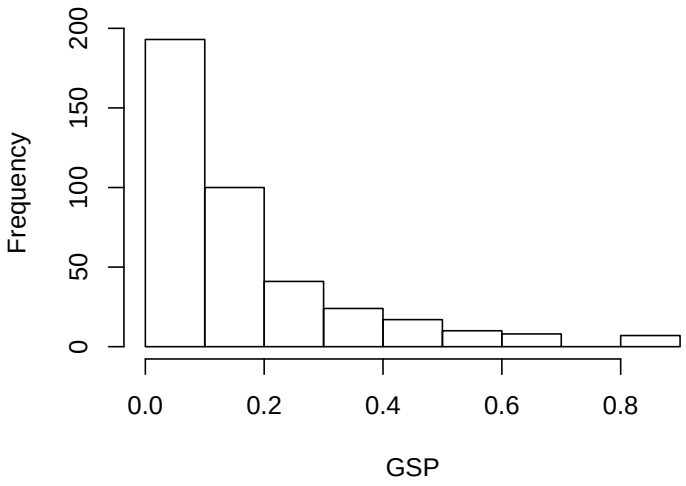

Figure 3.6: The histogram of graph sparsity levels for all 400 models.

axis represents the graph sparsity level.

In Table 3.2, we list the model fitting information of the ten models that have smallest BIC values. As we can see, all these models have three factors and have graph sparsity level at about 10%. In what follows, we investigate the model with the smallest BIC value. Two questions will be explored. First, what is the interpretation of the recovered three factors? Are they the Psychoticism, Extraversion, and Neuroticism factors? Second, what is the interpretation of the recovered graphical structure? In other words, what relationships between items is the graphical structure capturing?

To answer the first question, we need to pin down $\hat{A}$, which can only be identified up to a non-degenerate rotation. Rotational indeterminacy also appears in exploratory factor analysis, for which many rotational methods have been proposed to find a rotation. All these methods are trying to find a rotation such that the resulting loading matrix has as many entries close to zero as possible. This principle, first proposed in Thurstone (1947) and Cattell (1978), is based on the belief that each latent factor is only associated with a few number of items. We refer to Browne (2001) for a review of popular rotational methods. This rotational indeterminacy issue is also related to our discussion in Chapter 4. In this

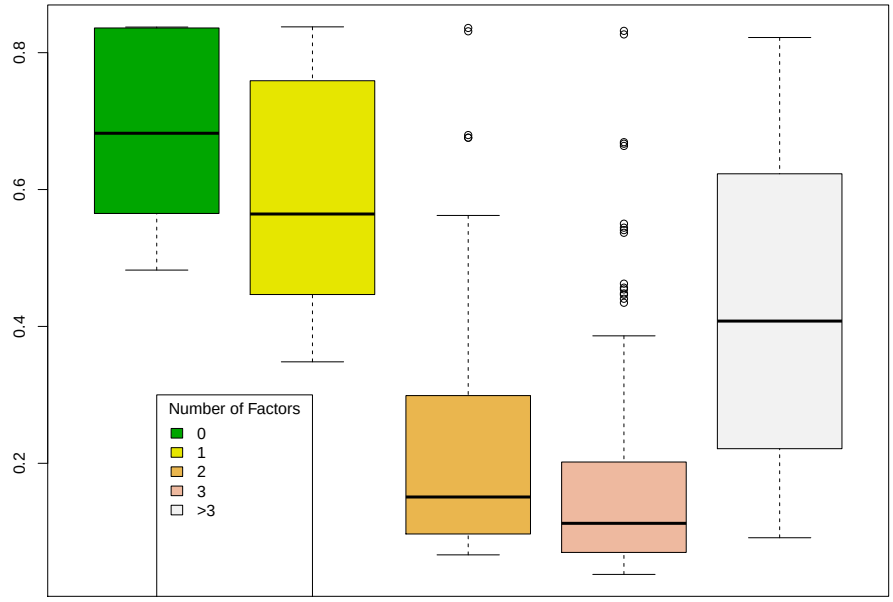

Figure 3.7: The number of factors VS the graph sparsity level for all models.

analysis, we obtain the estimated factor loading matrix $\hat{A}$, by using the varimax rotation (Kaiser, 1958), which is one of the most popular rotational methods in the literature of exploratory factor analysis. We then check the relationship between the latent factors identified by $\hat{A}$ and the three scales of EPQ-R. Based on model (3.6), the posterior mean of $\boldsymbol{\theta}_i$ is $E(\boldsymbol{\theta}_i|\mathbf{y}_i) = \mathbf{y}_i^\top A$. We replace A by its estimate $\hat{A}$, and use $\hat{\boldsymbol{\theta}}_i = \hat{A}^\top \mathbf{y}_i$ as an estimate of $\boldsymbol{\theta}_i$. In addition, we let T_i^P, T_i^E, and T_i^N be respondent i's total scores on the P, E, and N scales respectively. In Table 3.3, the sample correlation between $(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \hat{\theta}_{i3})$s and (T_i^P, T_i^E, T_i^N)s are calculated, where the diagonal entries being close to 1 implies that the three latent factors identified by the varimax rotation may be interpreted as the the Psychoticism, Extraversion, and Neuroticism factors respectively.

For the second question, the selected model has 346 edges (GSP = 11%). This graph measures the association among the items that is not attributable to the latent factors. Among the 346 edges, 91 are negative edges and 255 are positive. We first interpret the

	Pseudo-lik	BIC	K	Num-edge	GSP
1	-26177.6	55262.4	3	346	11%
2	-26207.9	55282.8	3	340	11%
3	-26338.4	55288.6	3	302	10%
4	-26157.1	55295.2	3	357	12%
5	-26133.0	55314.2	3	367	12%
6	-26319.5	55317.9	3	312	10%
7	-26017.4	55338.2	3	405	13%
8	-26297.9	55341.9	3	322	10%
9	-26499.1	55354.8	3	264	9%
10	-26264.1	55355.0	3	334	11%

Table 3.2: The logarithm of pseudo-likelihood, BIC, number of latent variables, number of edges, and graph sparsity levels for the top ten models.

positive ones. In Table 3.4, we present the ten item pairs that have the most positive edges. As we can see, items within a pair may share a common stimulus that is not completely attributable to the P, E, and N factors, resulting in additional dependence. For example, the first three pairs are about "party", "good manners", and "being lively", respectively. For some item pairs, the two items may express the same meaning under different wording, such as items "Do you stop to think things over before doing anything?" and "Do you generally 'look before you leap'?" in the fourth pair. In addition, an item itself may be the stimulus to the other. For example, for item pair 8, it is probably that a woman would like other people to be afraid of her, because her mother is (was) not a good woman.

We then turn to the negative item pairs. In Table 3.5, the ten item pairs with the most negative edges are presented. For these item pairs, it might be the case that the symptom in one item may trigger the one in the other in the opposite direction. For example, for the third pair, having a bad mother may cause a woman to lose interest in mixing with people.

	P	E	N
$\hat{\theta}_1$	**0.92**	0.29	-0.02
$\hat{\theta}_2$	0.06	**0.87**	-0.24
$\hat{\theta}_3$	0.11	-0.23	**0.85**

Table 3.3: The sample correlation between $(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \hat{\theta}_{i3})$s and (T_i^P, T_i^E, T_i^N)s.

In addition, items within some pairs have almost opposite meaning literally, such as item pair 10, "(R)Do good manners and cleanliness matter much to you?" and "Do you worry a lot about your looks?". It may also be the case that there is a common stimulus, that is not completely attributable to the latent factors, has opposite effects on the two symptoms. For example, for item pair 6, having social phobia may make a person not enjoy co-operating with others and not like mixing with others as well.

Finally, the interpretation of this graphical structure is not limited to the pairwise dependence. In particular, we check the cliques in the estimated graph. A clique is a subset of vertices of an undirected graph such that every two distinct vertices in the clique are connected. A maximal clique is a clique that cannot be extended by including one more adjacent vertex. For graphical models, random variables within a clique are usually considered to be highly dependent on each other. The estimated graph has 161 maximal cliques that have no less than three vertexes, including one 5-vertex clique, 32 4-vertex cliques, and 128 3-vertex cliques. In Table 3.6, we present the 5-vertex clique, two 4-vertex cliques, and two 3-vertex cliques. These 4-vertex and 3-vertex cliques are the ones having highest within-clique sum of $\hat{s}_{ij}$. It is interesting to observe that the maximal cliques of a graph, which is a geometric concept, identify meaningful item clusters. For example, the five cliques in Table 3.6 are about "communication with others", "think before action", "being nervous", "good manners", and "meeting people", respectively.

In summary, the propose method tends to recover sensible latent factor and local dependence structures of the EPQ-R. In particular, the model selected based on the regularized estimator and BIC receives good interpretion. In particular, the models yielding small BIC values are all three-factor models with similar graphical structures. Although we only report the model with the smallest BIC value, other models in Table 3.2 have very similar

results. Our results have two implications on the EPQ-R measurement. First, a three factor model that assumes local independence may not fit the EPQ-R data well and consequently a factor model based scoring rule (estimation of $\boldsymbol{\theta}_i$s) may be suboptimal for the EPQ-R measurement. Second, the estimated graph provides us diagnosis of the test items and let us think about revising the items. For example, if two items are highly locally dependent, we may consider to revise the wording of the items.

	$\hat{s}_{ij}$	Item	Scale	Item content
1	3.31	51	E	Can you easily get some life into a rather dull party?
		78	E	Can you get a patty going?
2	2.43	21	P	(R)Are good manners very important?
		41	P	(R)Do good manners and cleanliness matter much to you?
3	2.32	11	E	Are you rather lively?
		94	E	Do other people think of you as being very lively?
4	2.19	2	P	(R)Do you stop to think things over before doing anything?
		81	P	(R)Do you generally 'look before you leap'?
5	2.73	22	N	Are your feelings easily hurt?
		87	N	Are you easily hurt when people find fault with you ot the work you do?
6	1.97	35	N	Would you call yourself a nervous person?
		83	N	Do you suffer from 'nerves'?
7	1.83	6	E	Are you a talkative person?
		47	E	(R)Are you mostly quiet when you are with other people?
8	1.81	91	P	Would you like other people to be afraid of you?
		68	P	(R)Is (or was) your mother a good woman?
9	1.70	34	P	Do you have enemies who want to harm you?
		73	P	Are there several people who keep trying to avoid you?
10	1.69	24	E	(R)Do you tend to keep in the background on social occasions?
		47	E	(R)Are you mostly quiet when you are with other people?

Table 3.4: The top ten item pairs corresponding to the most positive edges. The item ID is consistent with Eysenck et al. (1985) and the reversely scored items are marked by "(R)".

	$\hat{s}_{ij}$	Item	Scale	Item content
1	-1.46	56	P	Do most things taste the same to you?
		58	E	Do you like mixing with people?
2	-1.16	68	P	(R)Is (or was) your mother a good woman?
		33	E	(R)Do you prefer reading to meeting people?
3	-1.07	68	P	(R)Is (or was) your mother a good woman?
		58	E	Do you like mixing with people?
4	-1.07	68	P	(R)Is (or was) your mother a good woman?
		13	N	Do you often worry about things you should not have done or said?
5	-0.99	79	P	(R)Do you try not to be rude to people?
		8	N	Do you ever feel just miserable for no reason?
6	-0.95	54	P	(R)Do you enjoy co-operating with others?
		58	E	Do you like mixing with people?
7	-0.93	52	N	Do you worry about your health?
		76	N	Have you ever wished that you were dead?
8	-0.88	52	N	Do you worry about your health?
		92	N	Are you sometimes bubbling over with energy and sometimes very sluggish?
9	-0.86	47	E	(R)Are you mostly quiet when you are with other people?
		97	N	Are you touchy about some things?
10	-0.81	41	P	(R)Do good manners and cleanliness matter much to you?
		74	N	Do you worry a lot about your looks?

Table 3.5: The top ten item pairs corresponding to the most negative edges. The item ID is consistent with Eysenck et al. (1985) and the reversely scored items are marked by "(R)".

	Item	Scale	Item content
1	6	E	Are you a talkative person?
	94	E	Do other people think of you as being very lively?
	47	E	(R)Are you mostly quiet when you are with other people?
	24	E	(R)Do you tend to keep in the background on social occasions?
	63	E	Do you nearly always have a 'ready answer' when people talk to you?
2	81	P	(R)Do you generally 'look before you leap'?
	2	P	(R)Do you stop to think things over before doing anything?
	69	E	Do you often make decisions on the spur of the moment?
	61	E	Have people said that you sometimes act too rashly?
3	35	N	Would you call yourself a nervous person?
	38	N	Are you a worrier?
	46	N	Would you call yourself tense or highly-strung?
	83	N	Do you suffer from nerves?
4	21	P	(R)Are good manners very important?
	14	P	(R)Do you dislike people who dont know how to behave themselves?
	41	P	(R)Do good manners and cleanliness matter much to you?
5	20	E	Do you enjoy meeting new people?
	33	E	(R)Do you prefer reading to meeting people?
	58	E	Do you like mixing with people?

Table 3.6: Examples of maximal cliques of the estimated graph. The item ID is consistent with Eysenck et al. (1985) and the reversely scored items are marked by "(R)".

3.8 Appendix of Chapter 3

3.8.1 Appendix A: Proof of Theorem 1

Throughout the proof, we will use κ as generic notation for large and not-so-important constants whose value may vary from place to place. Similarly, we use ε as generic notation for small positive constants. Furthermore, for two sequences of random variables a_N and b_N, we write $a_N = o_P(b_N)$ if $b_N/a_N \to 0$ in probability and $a_N = O_P(b_N)$ if a_N/b_N is tight. We also use the notations "$=_P$", "$<_P$", "$\leq_P$", "$>_P$" and "$\geq_P$" to indicate the equality and inequalities hold with a probability converging to one as N goes to infinity.

Proof Strategy. We first provide a sketch of the proof for the theorem. We introduce several notations and definitions. Let the eigendecomposition of L^* be $L^* = U^* D^* U^{*\top}$, such that U^* is a $J \times J$ orthogonal matrix and D^* is a $J \times J$ diagonal matrix whose first K diagonal elements are strictly positive. We write $U^* = [U_1^*, U_2^*]$ where U_1^* is the first K columns of U^*. Let D_1^* be the $K \times K$ diagonal matrix containing the nonzero diagonal elements of D^*. Define the localization set

$$\begin{aligned}\mathcal{M}_1 = \{(S, L): \quad & S = S^* + \delta_S, \quad \|\delta_S\|_\infty \leq \gamma_N^{1-2\eta}, \\ & L = UDU^\top, \quad \|U - U^*\|_\infty \leq \gamma_N^{1-\eta},\ U \text{ is a } J \times J \text{ orthogonal matrix}, \\ & \|D - D^*\|_\infty \leq \gamma_N^{1-2\eta},\ \text{and } D \text{ is a } J \times J \text{ diagonal matrix}\},\end{aligned}$$

and a subset

$$\begin{aligned}\mathcal{M}_2 = \{(S, L) : & S = S^* + \delta_S, \|\delta_S\|_\infty \leq \gamma_N^{1-2\eta}, \delta_S \in \mathcal{S}^*, \\ & L = U_1 D_1 U_1^\top, \|U_1 - U_1^*\|_\infty \leq \gamma_N^{1-\eta}, U_1 \text{ is a } J \times K \text{ matrix satisfying} \\ & U_1^\top U_1 = I_K, \|D_1 - D_1^*\|_\infty \leq \gamma_N^{1-2\eta},\ \text{and } D_1 \text{ is a } K \times K \text{ diagonal matrix}\}.\end{aligned}$$

Here, η is a positive constant that is sufficiently small. For each pair of $(S, L) \in \mathcal{M}_1$, S is close to S^*. Moreover, the eigendecoposition of L and L^* that are close to each other. As the sample size N grows large, the set $\mathcal{M}_1$ will tend to $\{(S^*, L^*)\}$. The set $\mathcal{M}_2$ is a subset of $\mathcal{M}_1$. For each pair of $(S, L) \in \mathcal{M}_2$, S has the same sparsity pattern as S^*, and L has the same rank as L^* for sufficiently large N.

The proof consists of two steps.

1. We first prove that with a probability converging to 1 the optimization problem (3.13) restricted to the subset $\mathcal{M}_2$ has a unique solution, which does not lie on the manifold boundary of $\mathcal{M}_2$. This part of proof is presented in Section 3.8.1.1.

2. We then show the unique solution restricted to $\mathcal{M}_2$ is also a solution to (3.13) on $\mathcal{M}_1$. It is further shown that with probability converging to 1 this solution is the unique solution to (3.13) subject to $(S, L) \in \mathcal{M}_1$. This part of proof is presented in Section 3.8.1.2.

The previous two steps together imply that the convex optimization problem (3.13) with the constraint $(S, L) \in \mathcal{M}_1$ has a unique solution that belongs to $\mathcal{M}_2$. Furthermore, this solution is an interior point of $\mathcal{M}_1$. Thanks to the convexity of the objective function, $(\hat{S}, \hat{L})$ is also the unique solution to the optimization problem (3.13). We conclude the proof by noticing that all $(S, L) \in \mathcal{M}_2$ converge to the true parameter (S^*, L^*) as $N \to \infty$ with the same sparsity and low rank structure.

3.8.1.1 Proof step 1

Denote by $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ a solution to the optimization problem

$$\begin{aligned} &\min_{(L,S)\in\mathcal{M}_2} \left\{ h_N(S+L) + \gamma_N \|\mathbf{O}(S)\|_1 + \delta_N \|L\|_* \right\} \\ &\text{subject to } L \text{ is positive semidefinite and } S \text{ is symmetric.} \end{aligned} \tag{3.26}$$

Recall here that the function h_N is defined in (3.17). We write the eigendecomposition $\hat{L}_{\mathcal{M}_2} = \hat{U}_{1,\mathcal{M}_2} \hat{D}_{1,\mathcal{M}_2} \hat{U}_{1,\mathcal{M}_2}^\top$, where $\hat{U}_{1,\mathcal{M}_2}^\top \hat{U}_{1,\mathcal{M}_2} = I_K$ and $\hat{D}_{1,\mathcal{M}_2}$ is a $K \times K$ diagonal matrix. To establish that $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ does not lie on the manifold boundary of $\mathcal{M}_2$, it is sufficient to show that

$$\|\hat{S}_{\mathcal{M}_2} - S^*\|_\infty <_P \gamma_N^{1-2\eta} \tag{3.27}$$

$$\|\hat{D}_{1,\mathcal{M}_2} - D_1^*\|_\infty <_P \gamma_N^{1-2\eta}. \tag{3.28}$$

$$\|\hat{U}_{1,\mathcal{M}_2} - U_1^*\|_\infty <_P \gamma_N^{1-\eta}. \tag{3.29}$$

To start with, we present a useful lemma.

Lemma 1. *Let*

$$\hat{\mathcal{D}} = \{\hat{U}_{1,\mathcal{M}_2} D_1' \hat{U}_{1,\mathcal{M}_2}^\top : D_1' \text{ is a } K \times K \text{ diagonal matrix}\}.$$

Consider the convex optimization problem

$$\min_{S \in \mathcal{S}^*, L \in \hat{\mathcal{D}}} \big\{ h_N(S+L) + \gamma_N \|\mathbf{O}(S)\|_1 + \delta_N \|L\|_* \big\}. \tag{3.30}$$

Then (3.30) *has a unique solution with probability converging to* 1*. Denote the solution by* $(\hat{S}_{\hat{\mathcal{D}}}, \hat{L}_{\hat{\mathcal{D}}})$ *with the singular decomposition* $\hat{L}_{\hat{\mathcal{D}}} = \hat{U}_{1,\mathcal{M}_2} \hat{D}_{1,\hat{\mathcal{D}}} \hat{U}_{1,\mathcal{M}_2}^\top$*. Then we have that there exists a constant* $\kappa > 0$ *such that* $\|\hat{S}_{\hat{\mathcal{D}}} - S^*\|_\infty \leq_P \kappa \gamma_N^{1-\eta}$ *and* $\|\hat{D}_{1,\hat{\mathcal{D}}} - D_1^*\|_\infty \leq_P \kappa \gamma_N^{1-\eta}$*.*

Because of the convexity of the objective function, a direct application of the above lemma is

$$(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2}) =_P (\hat{S}_{\hat{\mathcal{D}}}, \hat{L}_{\hat{\mathcal{D}}}) \text{ and } \hat{D}_{1,\mathcal{M}_2} =_P \hat{D}_{1,\hat{\mathcal{D}}}.$$

Thus, (3.27) and (3.28) are proved. We proceed to (3.29) by contradiction. If on the contrary $\|\hat{U}_{1,\mathcal{M}_2} - U_1^*\|_\infty = \gamma_N^{1-\eta}$, then later in the current section we will show that

$$h_N(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}}) + \gamma_N \|\mathbf{O}(\hat{S}_{\hat{\mathcal{D}}})\|_1 + \delta_N \|\hat{L}_{\hat{\mathcal{D}}}\|_* >_P h_N(S^* + L^*) + \gamma_N \|\mathbf{O}(S^*)\|_1 + \delta_N \|L^*\|_*. \tag{3.31}$$

We start with the Taylor expansion of $h_N(S+L)$ around S^* and L^*. Let $h(M) = \mathbb{E} h_N(M)$, then we have

$$h_N(M^* + \Delta) = h(M^*) + \frac{1}{2} v(\Delta)^\top \mathcal{I}^* v(\Delta) + R_N(\Delta), \tag{3.32}$$

where $M^* = S^* + L^*$, and the function $v : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}^{J^2} \times 1$ is a map that vectorizes matrices by rearranging their elements. Moreover, the term $R_N(\Delta)$ is the remainder term satisfying

$$R_N(\Delta) = R_N(\mathbf{0}_{J \times J}) + O_P(\|\Delta\|_\infty^3) + O_P(\frac{\|\Delta\|_\infty}{\sqrt{N}}) \text{ as } \Delta \to 0, N \to \infty. \tag{3.33}$$

Here, the term $R_N(\mathbf{0}_{J\times J})$ denotes $h_N(M^*) - h(M^*)$. The $O_P(\frac{\|\Delta\|_\infty}{\sqrt{N}})$ term corresponds to $\nabla h_N(M^*) \cdot \Delta$, and $O_P(\|\Delta\|_\infty^3)$ characterizes the remaining terms. Furthermore, the first and the second derivatives satisfy

$$\nabla R_N(\Delta) = O(\|\Delta\|_\infty^2) + O_P(\frac{1}{\sqrt{N}}) \text{ as } \Delta \to 0 \tag{3.34}$$

and

$$\nabla^2 R_N(\Delta) = O_P(\frac{1}{\sqrt{N}} + \|\Delta\|_\infty). \tag{3.35}$$

We plug $\Delta = \hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - S^* - L^*$ and $\Delta = \mathbf{0}_{J\times J}$ separately into (3.32) and take the difference between the resulting equations, then

$$\begin{aligned} &h_N(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}}) - h_N(S^* + L^*) \\ &= \frac{1}{2} v(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*))^\top \mathcal{I}^* v(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)) + R_N(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)) - R_N(\mathbf{0}_{J\times J}). \end{aligned} \tag{3.36}$$

We first establish a lower bound for $R_N(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)) - R_N(\mathbf{0}_{J\times J})$. According to Lemma 1, we have

$$\|\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)\|_\infty \leq_P \kappa \gamma_N^{1-\eta}, \tag{3.37}$$

with a possibly different κ. The above display and (3.33) yield

$$R_N(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)) - R_N(\mathbf{0}_{J\times J}) = O((\gamma_N^{1-\eta})^3) + O_P(\frac{\gamma_N^{1-\eta}}{\sqrt{N}}). \tag{3.38}$$

We proceed to a lower bound for the term $\frac{1}{2} v(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*))^\top \mathcal{I}^* v(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*))$ with the aid of the following two lemmas.

Lemma 2. *Under Assumption A4, there exists a positive constant ε such that*

$$\|S + L\|_\infty \geq \varepsilon \|L\|_\infty \text{ for all } (S, L) \in \mathcal{S}^* \times T_{L^*}\mathfrak{L}.$$

Lemma 3. *Let*

$$\Delta_L = U_1^* D_1^* (\hat{U}_{1,\mathcal{M}_2} - U_1^*)^\top + (\hat{U}_{1,\mathcal{M}_2} - U_1^*) D_1^* U_1^{*T} + U_1^* (\hat{D}_{1,\hat{\mathcal{D}}} - D_1^*) U_1^{*T}, \tag{3.39}$$

then we have

(i) $\Delta_L \in T_{L^*}\mathfrak{L}$.

(ii) There exists positive constant ε such that $\|\Delta_L\|_\infty >_P \varepsilon \gamma_N^{1-\eta}$, given $\|\hat{U}_{1,\mathcal{M}_2} - U_1^*\|_\infty = \gamma_N^{1-\eta}$.

(iii) $\|\hat{L}_{\hat{\mathcal{D}}} - L^* - \Delta_L\|_\infty \leq_P \kappa(\gamma_N^{1-\eta})^2$.

According to Lemma 2 and (i) and (iii) of Lemma 3 and noticing that $\hat{S}_{\hat{\mathcal{D}}} - S^* \in \mathcal{S}^*$, we have

$$\|\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)\|_\infty \geq_P \|\hat{S}_{\hat{\mathcal{D}}} - S^* + \Delta_L\|_\infty - \kappa(\gamma_N^{1-\eta})^2 \geq_P \varepsilon\|\Delta_L\|_\infty - \kappa(\gamma_N^{1-\eta})^2.$$

According to Lemma 3(ii), the above display further implies that

$$\|\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)\|_\infty >_P \varepsilon\gamma_N^{1-\eta}, \tag{3.40}$$

with a possibly different ε. According to assumption A1, $\mathcal{I}^*$ is positive definite. Therefore,

$$v(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*))^\top \mathcal{I}^* v(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)) > \varepsilon\|\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}} - (S^* + L^*)\| \geq_P \varepsilon^2(\gamma_N^{1-\eta})^2. \tag{3.41}$$

The second inequality of the above display is due to (3.40). (3.36), (3.38) and (3.41) give

$$h_N(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}}) - h_N(S^* + L^*) >_P \frac{\varepsilon^2}{2}(\gamma_N^{1-\eta})^2. \tag{3.42}$$

We proceed to the regularization terms in (3.31). For the L_1 penalty term, we have

$$\|\mathbf{O}(\hat{S}_{\hat{\mathcal{D}}})\|_1 - \|\mathbf{O}(S^*)\|_1 = \text{sign}(\mathbf{O}(S^*)) \cdot (\hat{S}_{\hat{\mathcal{D}}} - S^*) =_P O(\gamma_N^{1-\eta}). \tag{3.43}$$

The second equality in the above display is due to Lemma 1. For the nuclear norm term, we have

$$\|\hat{L}_{\hat{\mathcal{D}}}\|_* - \|L^*\|_* \geq -\|\hat{L}_{\hat{\mathcal{D}}} - L^*\|_* \geq -\kappa\gamma_N^{1-\eta}.$$

Again, the second inequality in the above display is because of Lemma 1. Notice that $\delta_N = \rho\gamma_N$. Equation (3.42), (3.43) and the above inequality imply

$$h_N(\hat{S}_{\hat{\mathcal{D}}} + \hat{L}_{\hat{\mathcal{D}}}) - h_N(S^* + L^*) + \gamma_N(\|\mathbf{O}(\hat{S}_{\hat{\mathcal{D}}})\|_1 - \|\mathbf{O}(S^*)\|_1) + \delta_N(\|\hat{L}_{\hat{\mathcal{D}}}\|_* - \|L^*\|_*)$$
$$>_P \varepsilon(\gamma_N^{1-\eta})^2 >_P 0,$$

with a possibly different ε. Notice that $(\hat{S}_{\hat{\mathcal{D}}}, \hat{L}_{\hat{\mathcal{D}}}) = (\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$, so we obtain (3.31) by rearranging terms in the above inequality, and this contradicts the definition of $\hat{S}_{\mathcal{M}_2}$ and $\hat{L}_{\mathcal{M}_2}$. This completes the proof for (3.29). The uniqueness of the solution is obtained according to the following lemma.

Lemma 4. *The solution to the optimization problem* (3.26) *is unique with a probability converging to* 1*. In addition,*

$$(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2}) = (S^*, L^*) + \mathbf{F}^{-1}(\gamma_N \text{sign}(\mathbf{O}(S^*)), \delta_N U_1^* U_1^{*T}) + o_P(\gamma_N^{1-\eta}), \tag{3.44}$$

as $N \to \infty$*.*

3.8.1.2 Proof step 2

In this section, we first show that $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ is a solution of the optimization problem

$$\begin{aligned} &\min_{(L,S)\in\mathcal{M}_1} h_N(S+L) + \gamma_N \|\mathbf{O}(S)\|_1 + \delta_N \|L\|_*, \\ &\text{subject to } L \text{ is positive semidefinite and } S \text{ is symmetric.} \end{aligned} \tag{3.45}$$

To prove this, we will show that $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ satisfies the first order condition

$$\mathbf{0}_{J\times J} \in \partial_S H|_{(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})} \text{ and } \mathbf{0}_{J\times J} \in \partial_L H|_{(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})}, \tag{3.46}$$

where the function H is the objective function

$$H(S, L) = h_N(S+L) + \gamma_N \|\mathbf{O}(S)\|_1 + \delta_N \|L\|_* \tag{3.47}$$

and $\partial_S H$ and $\partial_L H$ denotes the sub-differentials of H. See Rockafellar (2015) for more details of sub-differentials of convex functions. We first derive an explicit expression of the first order condition. The sub-differential with respect to S is defined as

$$\partial_S H|_{(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})} = \{\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) + \gamma_N(\text{sign}(\mathbf{O}(S^*))) + \gamma_N W : \|W\|_\infty \le 1 \text{ and } W \in \mathcal{S}^{*\perp}\}, \tag{3.48}$$

where $\mathcal{S}^{*\perp}$ is the orthogonal complement space of $\mathcal{S}^*$ in the space of symmetric matrices. According to Example 2 of Watson (1992), the sub-differential with respect to L is

$$\partial_L H|_{(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})} = \{\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) + \delta_N \hat{U}_{1,\mathcal{M}_2} \hat{U}_{1,\mathcal{M}_2}^\top + \delta_N \hat{U}_{2,\mathcal{M}_2} W \hat{U}_{2,\mathcal{M}_2}^\top : \|W\|_2 \le 1\}, \tag{3.49}$$

where $\hat{U}_{2,\mathcal{M}_2}$ is a $J \times (J-K)$ matrix satisfying $\hat{U}_{1,\mathcal{M}_2}^\top \hat{U}_{2,\mathcal{M}_2} = \mathbf{0}_{K\times(J-K)}$, $\hat{U}_{2,\mathcal{M}_2}^\top \hat{U}_{2,\mathcal{M}_2} = I_{J-K}$ and $\|\hat{U}_{2,\mathcal{M}_2} - U_2^*\|_\infty \le \gamma_N^{1-\eta}$. For all (S, L), if

$$\mathbf{P}_{\mathcal{S}^*} S = \mathbf{0}_{J\times J},\ \mathbf{P}_{\mathcal{S}^{*\perp}} S = \mathbf{0}_{J\times J},\ \mathbf{P}_{T_{\hat{L}_{\mathcal{M}_2}}} \mathfrak{L} L = \mathbf{0}_{J\times J} \text{ and } \mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}} \mathfrak{L})^\perp} L = \mathbf{0}_{J\times J},$$

then $S = \mathbf{0}_{J\times J}$ and $L = \mathbf{0}_{J\times J}$. Consequently, to prove (3.46), it is sufficient to show that

$$\mathbf{P}_{\mathcal{S}^*}\partial_S H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})} = \{\mathbf{0}_{J\times J}\} \text{ and } \mathbf{P}_{T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L}}\partial_L H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})} = \{\mathbf{0}_{J\times J}\}, \tag{3.50}$$

and

$$\mathbf{0}_{J\times J} \in \mathbf{P}_{\mathcal{S}^{*\perp}}\partial_S H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})} \text{ and } \mathbf{0}_{J\times J} \in \mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L})^\perp}\partial_L H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})}. \tag{3.51}$$

According to the definition of $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$, it is the solution to the optimization (3.26). In addition, according to the discussion in Section 3.8.1.1, $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ does not lie on the boundary of $\mathcal{M}_2$. Therefore, it satisfies the first order condition of (3.26), which turns out to be (3.50). Thus, to prove (3.46) it is sufficient to show (3.51). The next lemma establishes an equivalent expression for (3.51).

Lemma 5. (3.51) *is equivalent to*

$$\|\mathbf{P}_{\mathcal{S}^{*\perp}}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2})\|_\infty \leq_P \gamma_N \text{ and } \|\mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L})^\perp}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2})\|_2 \leq_P \delta_N. \tag{3.52}$$

We proceed to prove (3.52). Taking gradient on both side of (3.32), we have

$$\nabla h_N(S^* + L^* + \Delta) = \mathcal{I}^*(\Delta) + \nabla R_N(\Delta). \tag{3.53}$$

We plug $\Delta = \hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2} - S^* - L^*$ into the above equation,

$$\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) = \mathcal{I}^*(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2} - S^* - L^*) + \nabla R_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2} - S^* - L^*). \tag{3.54}$$

According to Lemma 4,

$$\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2} - S^* - L^* = \mathbf{A}\mathbf{F}^{-1}(\gamma_N \text{sign}(\mathbf{O}(S^*)), \delta_N U_1^* U_1^{*T}) + o_P(\gamma_N),$$

where $\mathbf{A}$ is the adding operator of two matrices $\mathbf{A}(A, B) = A + B$. Combining this with (3.34), (3.54), and notice that $\delta_N = \rho\gamma_N$, we have

$$\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) = \gamma_N \mathcal{I}^* \mathbf{A}\mathbf{F}^{-1}(\text{sign}(\mathbf{O}(S^*)), \rho U_1^* U_1^{*T}) + o_P(\gamma_N).$$

and consequently,

$$\begin{aligned}
\mathbf{P}_{\mathcal{S}^{*\perp}}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) &= \gamma_N \mathbf{P}_{\mathcal{S}^{*\perp}}\mathcal{I}^*\mathbf{A}\mathbf{F}^{-1}(\text{sign}(\mathbf{O}(S^*)), \rho U_1^* U_1^{*T}) + o_P(\gamma_N), \\
\mathbf{P}_{T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L}^\perp}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) &= \gamma_N \mathbf{P}_{\mathcal{T}_{\mathcal{L}^*}\mathfrak{L}^\perp}\mathcal{I}^*\mathbf{A}\mathbf{F}^{-1}(\text{sign}(\mathbf{O}(S^*)), \rho U_1^* U_1^{*T}) + o_P(\gamma_N).
\end{aligned} \tag{3.55}$$

We complete the proof for $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ to be a solution of (3.45) by noticing that (3.52) is a direct application of Assumption A6 and the above equation. We proceed to the proof of the uniqueness of the solution to (3.45). Because the objective function $H(S, L)$ is a convex function, it is sufficient to show the uniqueness of the solution in a neighborhood of $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$. We choose a small neighborhood as follows.

$$\mathcal{N} = \Big\{(S, L) : \|S - \hat{S}_{\mathcal{M}_2}\|_\infty < e^{-N},\ S \text{ is symmetric, } L \text{ has the eigendecomposition } L = \begin{bmatrix} U_1, U_2 \end{bmatrix} \begin{bmatrix} D_1 & \mathbf{0}_{J\times(J-K)} \\ \mathbf{0}_{(J-K)\times J} & D_2 \end{bmatrix} \begin{bmatrix} U_1, U_2 \end{bmatrix}^\top, \|U_1 - \hat{U}_{1,\mathcal{M}_2}\|_\infty < e^{-N}, \|U_2 - \hat{U}_{2,\mathcal{M}_2}\|_\infty < e^{-N}, \|D_1 - \hat{D}_{1,\mathcal{M}_2}\|_\infty < e^{-N}, \|D_2\|_\infty < e^{-N}\Big\}. \tag{3.56}$$

If on the contrary $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}) \in \mathcal{N}$ is a solution to (3.45) and is different from $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$, the next lemma establishes that $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}) \in \mathcal{M}_2$.

Lemma 6. *For all $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}) \in \mathcal{N}$, if $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})$ is a solution to* (3.45)*, then*

$$(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}) \in \mathcal{M}_2$$

.

This contradicts the uniqueness of solution to (3.26) established in Lemma 4.

3.8.2 Appendix B: Proof of the Supporting Lemmas

Proof of Lemma 1. We consider the first order condition for the optimization problem (3.30). Notice that $\mathcal{S}^*$ and $\hat{\mathcal{D}}$ are linear spaces, so the first order condition becomes

$$\mathbf{0}_{J\times J} \in \mathbf{P}_{\mathcal{S}^*} \partial_S H|_{(S,L)} \text{ and } \mathbf{0}_{J\times J} \in \mathbf{P}_{\hat{\mathcal{D}}} \partial_L H|_{(S,L)},$$

where H is defined in(3.47). We will show that there is a unique $(S, L) \in \mathcal{S}^* \times \hat{\mathcal{D}}$ satisfying the first order condition. Because of the convexity of the optimization problem (3.30), it is sufficient to show that with a probability converging to 1 there is a unique $(S, L) \in \mathcal{B}$ satisfying the first order condition, where

$$\mathcal{B} = \{(S, L) \in \mathcal{S}^* \times \hat{\mathcal{D}} : \|S - S^*\|_\infty \le \gamma_N^{1-\eta}, \text{ and } \|L - L^*\|_\infty \le \gamma_N^{1-\eta}\}.$$

We simplify the first order condition for $(S, L) \in \mathcal{B}$. For the L_1 penalty term, if $\|S-S^*\|_\infty \leq \gamma_N^{1-\eta}$ and $S \in \mathcal{S}^*$, then $\|\mathbf{O}(S)\|_1$ is smooth and

$$\mathbf{P}_{\mathcal{S}^*}\partial_S\|\mathbf{O}(S)\|_1 = \text{sign}(\mathbf{O}(S^*)) \text{ for } S \in \mathcal{S}^*. \tag{3.57}$$

Similarly, for $L \in \hat{\mathcal{D}}$ and $\|L - L^*\|_\infty \leq \gamma_N^{1-\eta}$, $\|L\|_*$ is smooth over the linear space $\hat{\mathcal{D}}$ and

$$\mathbf{P}_{\hat{\mathcal{D}}}\partial_L\|L\|_* = \hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top \text{ for } L \in \hat{\mathcal{D}}. \tag{3.58}$$

Combining (3.57) and (3.58) with the ∇h_N term, we arrive at an equivalent form of the first order condition,

$$\begin{aligned}\mathbf{P}_{\mathcal{S}^*}\nabla h_N(S+L) + \gamma_N\text{sign}(\mathbf{O}(S^*)) &= \mathbf{0}_{J\times J},\\ \mathbf{P}_{\hat{\mathcal{D}}}\nabla h_N(S+L) + \delta_N\hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top &= \mathbf{0}_{J\times J}.\end{aligned}$$

We will show the existence and uniqueness of the solution to the above equations using contraction mapping theorem. We first construct the contraction operator. Let $(S, L) = (S^* + \Delta_S, L^* + \Delta_L)$. We plug (3.53) into the above equation, and arrive at an equivalent equation

$$\begin{aligned}\mathbf{P}_{\mathcal{S}^*}\mathcal{I}^*(\Delta_S+\Delta_L) + \mathbf{P}_{\mathcal{S}^*}\nabla R_N(\Delta_S+\Delta_L) + \gamma_N\text{sign}(\mathbf{O}(S^*)) &= \mathbf{0}_{J\times J},\\ \mathbf{P}_{\hat{\mathcal{D}}}\mathcal{I}^*(\Delta_S+\Delta_L) + \mathbf{P}_{\hat{\mathcal{D}}}\nabla R_N(\Delta_S+\Delta_L) + \delta_N\hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top &= \mathbf{0}_{J\times J}.\end{aligned} \tag{3.59}$$

We define an operator $\tilde{\mathbf{F}}_{\hat{\mathcal{D}}} : \mathcal{S}^* \times (\hat{\mathcal{D}} - L^*) \to \mathcal{S}^* \times \hat{\mathcal{D}}$ that is similar to $\tilde{\mathbf{F}}$,

$$\tilde{\mathbf{F}}_{\hat{\mathcal{D}}}(\Delta_S, \Delta_L) = \Big(\mathbf{P}_{\mathcal{S}^*}\mathcal{I}^*(\Delta_S+\Delta_L), \mathbf{P}_{\hat{\mathcal{D}}}\mathcal{I}^*(\Delta_S+\Delta_L)\Big),$$

where the set $\hat{\mathcal{D}} - L^* = \{L - L^* : L \in \hat{\mathcal{D}}\}$. We further transform equation (3.59) to

$$\begin{aligned}&\tilde{\mathbf{F}}_{\hat{\mathcal{D}}}(\Delta_S, \Delta_L) + (\mathbf{P}_{\mathcal{S}^*}\nabla R_N(\Delta_S+\Delta_L), \mathbf{P}_{\hat{\mathcal{D}}}\nabla R_N(\Delta_S+\Delta_L)) + (\gamma_N\text{sign}(\mathbf{O}(S^*)), \delta_N\hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top)\\ =&(\mathbf{0}_{J\times J}, \mathbf{0}_{J\times J}).\end{aligned} \tag{3.60}$$

Notice that the projection $\mathbf{P}_{\hat{\mathcal{D}}}$ is uniquely determined by the matrix $\hat{U}_{1,\mathcal{M}_2}$. The next lemma states that the mapping $\hat{U}_{1,\mathcal{M}_2} \to \mathbf{P}_{\hat{\mathcal{D}}}$ is Lipschitz.

Lemma 7. *We write the eigendecomposition $L = UDU^\top = U_1D_1U_1^\top$, and define the corresponding linear spaces $T_L\mathfrak{L}$ and $\mathcal{D}$ as*

$$\mathcal{D} = \{U_1D_1'U_1^\top : D_1' \text{ is a } K \times K \text{ diagonal matrix}\}, \tag{3.61}$$

and

$$T_L\mathfrak{L} = \{U_1 Y + Y^\top U_1^\top : Y \text{ is a } K \times J \text{ matrix}\}. \tag{3.62}$$

Then, the mappings $U_1 \to \mathbf{P}_{\mathcal{L}}$ *and* $U_1 \to \mathbf{P}_{\mathcal{D}}$ *are Lipschitz in* U_1*. That is, for all* $J \times J$ *symmetric matrix* M*, there exists a constant* κ *such that*

$$\|\mathbf{P}_{T_L\mathfrak{L}} M - \mathbf{P}_{T_{L^*}\mathfrak{L}} M\|_\infty \le \kappa \|U_1 - U_1^*\|_\infty \|M\|_\infty, \text{ and}$$
$$\|\mathbf{P}_{\mathcal{D}} M - \mathbf{P}_{\mathcal{D}^*} M\|_\infty \le \kappa \|U_1 - U_1^*\|_\infty \|M\|_\infty.$$

According to the above lemma, Assumption A5 that $\tilde{\mathbf{F}}_{\mathcal{D}^*}$ is invertible over $\mathcal{S}^* \times \mathcal{D}^*$, and the fact that $\|\hat{U}_{1,\mathcal{M}_2} - U_1^*\|_\infty \le \gamma_N^{1-\eta}$, we know that $\tilde{\mathbf{F}}_{\hat{\mathcal{D}}}$ is also invertible over $\mathcal{S}^* \times (\hat{\mathcal{D}} - L^*)$ and is Lipschitz in $\hat{U}_{1,\mathcal{M}_2}$ for sufficiently large N. We apply $\tilde{\mathbf{F}}_{\hat{\mathcal{D}}}^{-1}$ on both sides of (3.60) and transform it to a fixed point problem,

$$(\Delta_S, \Delta_L) = \mathbf{C}(\Delta_S, \Delta_L), \tag{3.63}$$

where the operator $\mathbf{C}$ is defined by

$$\begin{aligned}&\mathbf{C}(\Delta_S, \Delta_L)\\ &= -\tilde{\mathbf{F}}_{\hat{\mathcal{D}}}^{-1}\Big((\mathbf{P}_{\mathcal{S}^*}\nabla R_N(\Delta_S + \Delta_L), \mathbf{P}_{\hat{\mathcal{D}}}\nabla R_N(\Delta_S + \Delta_L)) + (\gamma_N \text{sign}(\mathbf{O}(S^*)), \delta_N \hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top)\Big).\end{aligned} \tag{3.64}$$

Define the set $\mathcal{B}^* = \mathcal{B} - (S^*, L^*) = \{(S - S^*, L - L^*) : (S, L) \in \mathcal{B}\}$. We will show that with a probability converging to 1, $\mathbf{C}$ is a contraction mapping over $\mathcal{B}^*$. First, according to (3.34) and the definition of set $\mathcal{B}$, it is easy to check that with probability converging to 1, $\mathbf{C}(\Delta_S, \Delta_L) \in \mathcal{B}^*$ for all $(S, L) \in \mathcal{B}^*$, so $\mathbf{C}(\mathcal{B}^*) \subset \mathcal{B}^*$. Next, according to (3.35) and the boundedness of $\tilde{\mathbf{F}}_{\hat{\mathcal{D}}}^{-1}$, we know that $\mathbf{C}(\Delta_S, \Delta_L)$ is Lipschitz in (Δ_S, Δ_L) with a probability converging to 1. To see the size of the Lipschitz constant, according to (3.35) we know that $\nabla R_N(\Delta_S + \Delta_L)$ is Lipschitz with respect to (Δ_S, Δ_L) with the Lipschitz constant of order $O_P(\gamma_N^{1-\eta})$. Therefore, the Lipschitz constant for $\mathbf{C}$ is also of order $O_P(\gamma_N^{1-\eta})$. Consequently, $\mathbf{C}$ is a contraction mapping over the complete metric space $\mathcal{B}^*$ with a probability converging to 1. According to Banach fixed point theorem, (3.63) has a unique solution in $\mathcal{B}^*$ with a probability converging to 1. This concludes our proof. □

Proof of Lemma 2. According to Assumption A4, $\mathcal{S}^* \cap T_{L^*}\mathfrak{L} = \mathbf{0}_{J\times J}$. Then, for all $L \in$

$T_{L^*}\mathfrak{L}$, we have

$$\|L - \mathbf{P}_{\mathcal{S}^*} L\|_F > 0,$$

where the norm $\|\cdot\|_F$ is the Frobenius norm. Because the set $\{L : \|L\|_F = 1\}$ is compact, $\inf_{\|L\|_F=1, L\in T_{L^*}\mathfrak{L}} \|L - \mathbf{P}_{\mathcal{S}^*} L\|_F > 0$. As a result,

$$\|L - \mathbf{P}_{\mathcal{S}^*} L\|_F \geq \varepsilon \|L\|_F, \tag{3.65}$$

where $\varepsilon = \inf_{\|L\|_F=1, L\in T_{L^*}\mathfrak{L}} \|L - \mathbf{P}_{\mathcal{S}^*} L\|_F$. We proceed to a lower bound for $\|L+S\|_F$. For $L \in T_{L^*}\mathfrak{L}$ and $S \in \mathcal{S}^*$ we have

$$\|S + L\|_F = \|\mathbf{P}_{\mathcal{S}^{*\top}} L\|_F + \|\mathbf{P}_{\mathcal{S}^*} L + S\|_F \geq \|\mathbf{P}_{\mathcal{S}^{*\perp}} L\|_F = \|L - \mathbf{P}_{\mathcal{S}^*} L\|_F \geq \varepsilon \|L\|_F.$$

The last inequality in the above display is due to (3.65). We complete the proof by noticing all norms are equivalent for finite dimensional space. □

Proof of Lemma 3. Taking $Y = D_1^*(\hat{U}_{1,\mathcal{M}_2} - U_1^*)^\top + \frac{1}{2}(\hat{D}_{1,\hat{\mathcal{D}}} - D_1^*)U_1^{*T}$, we have $\Delta_L = U_1^* Y + Y^\top U_1^{*T}$. Therefore, $\Delta_L \in T_{L^*}\mathfrak{L}$ and (i) is proved. Now, we write

$$\Delta_{U_1} = \hat{U}_{1,\mathcal{M}_2} - U_1^* \text{ and } \Delta_{\hat{D}_{1,\mathcal{M}_2}} - D_1^*,$$

then we have

$$\hat{L}_{\hat{\mathcal{D}}} - L^* = (U_1^* + \Delta_{U_1})(D_1^* + \Delta_{D_1})(U_1^* + \Delta_{U_1})^\top - U_1^* D_1^* U_1^{*\top} = \Delta_L + O(\|\Delta_{U_1}\|_\infty^2 + \|\Delta_{D_1}\|_\infty^2).$$

According to Lemma 1 and (3.29) we have $O(\|\Delta_{U_1}\|_\infty^2 + \|\Delta_{D_1}\|_\infty^2) \leq \kappa(\gamma_N^{1-\eta})^2$. Thus, (iii) is proved. To prove (ii), we need the following eigenvalue perturbation result.

Lemma 8 (Eigenvalue perturbation)**.** *Under Assumption A2, for all $J \times K$ matrix U_1 such that $U_1^\top U_1 = I_K$, and $\|U_1 - U^*\|_\infty = \gamma_N^{1-\eta}$, and all $K \times K$ diagonal matrix D_1 such that $\|D_1 - D_1^*\|_\infty \leq \kappa\gamma_N^{1-\eta}$, there exists a positive constant ε independent with U_1 and D_1 (possibly depending on κ) satisfying*

$$\|U_1 D_1 U_1^\top - L^*\|_\infty \geq \varepsilon\gamma_N^{1-\eta}.$$

As a direct application of the above lemma, we have

$$\|\hat{L}_{\hat{\mathcal{D}}} - L^*\|_\infty \geq \varepsilon\gamma_N^{1-\eta}. \tag{3.66}$$

Combing (iii) with (3.66), we have (ii) proved. □

Proof of Lemma 4. Assume that on the contrary, (3.26) has two solutions $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ and $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})$. Similar to $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$, $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})$ also satisfy (3.27), (3.28) and (3.29) if we replace $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$ by $(\hat{S}_{\mathcal{M}_2}, \hat{L}_{\mathcal{M}_2})$, and it is also an interior point of $\mathcal{M}_2$. Thus, it satisfies the first order condition of (3.26). That is,

$$\begin{aligned}\mathbf{P}_{\mathcal{S}^*}\nabla h_N(\hat{S}_{\mathcal{M}_2}+\hat{L}_{\mathcal{M}_2})+\gamma_N\text{sign}(\mathbf{O}(S^*)) &= \mathbf{0}_{J\times J}\\ \mathbf{P}_{T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L}}\nabla h_N(\hat{S}_{\mathcal{M}_2}+\hat{L}_{\mathcal{M}_2})+\delta_N\hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top &= \mathbf{0}_{J\times J}.\end{aligned}$$

We define an operator $\mathbf{F}_L : \mathcal{S}^* \times T_L\mathfrak{L} \to \mathcal{S}^* \times T_{L1}\mathfrak{L}$ in a similar way as that of $\mathbf{F}$,

$$\mathbf{F}_L(S, L') = (\mathbf{P}_{\mathcal{S}^*}\{\mathcal{I}^*(S+L')\}, \mathbf{P}_{T_L\mathfrak{L}}\{\mathcal{I}^*(S+L')\}). \tag{3.67}$$

With similar arguments as those above (3.63), we know that $\mathbf{F}_{\hat{L}_{\mathcal{M}_2}}$ is invertible under Assumption A3. Moreover, similar to (3.63), we have

$$\begin{aligned}&(\hat{S}_{\mathcal{M}_2}-S^*, \hat{L}_{\mathcal{M}_2}-L^*)\\ =&\mathbf{F}^{-1}_{\hat{L}_{\mathcal{M}_2}}\Big((\mathbf{P}_{\mathcal{S}^*}\nabla R_N(\Delta_{\hat{S}_{\mathcal{M}_2}}+\Delta_{\hat{L}_{\mathcal{M}_2}}), \mathbf{P}_{T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L}}\nabla R_N(\Delta_{\hat{S}_{\mathcal{M}_2}}+\Delta_{\hat{L}_{\mathcal{M}_2}}))+(\gamma_N\text{sign}(\mathbf{O}(S^*)), \delta_N\hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top)\Big).\end{aligned} \tag{3.68}$$

Similarly for $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})$, we have

$$\begin{aligned}&(\tilde{S}_{\mathcal{M}_2}-S^*, \tilde{L}_{\mathcal{M}_2}-L^*)\\ =&\mathbf{F}^{-1}_{\tilde{L}_{\mathcal{M}_2}}\Big((\mathbf{P}_{\mathcal{S}^*}\nabla R_N(\Delta_{\tilde{S}_{\mathcal{M}_2}}+\Delta_{\tilde{L}_{\mathcal{M}_2}}), \mathbf{P}_{T_{\tilde{L}_{\mathcal{M}_2}}\mathfrak{L}}\nabla R_N(\Delta_{\tilde{S}_{\mathcal{M}_2}}+\Delta_{\tilde{L}_{\mathcal{M}_2}}))+(\gamma_N\text{sign}(\mathbf{O}(S^*)), \delta_N\tilde{U}_{1,\mathcal{M}_2}\tilde{U}_{1,\mathcal{M}_2}^\top)\Big).\end{aligned} \tag{3.69}$$

Similar to the definition (3.64), for $(S, L) \in \mathcal{M}_2$ we define

$$\begin{aligned}&\mathbf{C}_{(S,L)}(\Delta_S, \Delta_L)\\ =&-\mathbf{F}_L^{-1}\Big((\mathbf{P}_{\mathcal{S}^*}\nabla R_N(\Delta_S+\Delta_L), \mathbf{P}_{T_L\mathfrak{L}}\nabla R_N(\Delta_S+\Delta_L))+(\gamma_N\text{sign}(\mathbf{O}(S^*)), \delta_N U_1U_1^\top\Big),\end{aligned}$$

where L has the eigendecomposition $L = U_1D_1U_1^\top$, $\|U_1 - U_1^*\|_\infty \le \gamma_N^{1-\eta}$ and $\|D_1 - D_1^*\|_\infty \le \gamma_N^{1-2\eta}$. The operator $\mathbf{C}_{(S,L)}$ is well defined, because for $L \in \mathcal{M}_2$ the eigendeposition of L is uniquely determined given (U_1, D_1) is in the set $\{(U_1, D_1) : \|U_1 - U_1^*\|_\infty \le \gamma_N^{1-\eta}$ and $\|D_1 - D_1^*\|_\infty \le \gamma_N^{1-2\eta}\}$. See more results on eigenvalue perturbation in (Parlett, 1980, Chapter 4). Now, we take difference between (3.68) and (3.69),

$$(\tilde{S}_{\mathcal{M}_2}-\hat{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}-\hat{L}_{\mathcal{M}_2}) = \mathbf{C}_{\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}}(\Delta_{\tilde{S}_{\mathcal{M}_2}}, \Delta_{\tilde{L}_{\mathcal{M}_2}}) - \mathbf{C}_{\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2}}(\Delta_{\hat{S}_{\mathcal{M}_2}}, \Delta_{\hat{L}_{\mathcal{M}_2}}). \tag{3.70}$$

We provide an upper bound for the norm of right-hand-side of the above equation. By means of triangular inequality, we split the right-hand-side of the above display into two terms.

$$\begin{aligned} & \|\mathbf{C}_{\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}}(\Delta_{\tilde{S}_{\mathcal{M}_2}},\Delta_{\tilde{L}_{\mathcal{M}_2}}) - \mathbf{C}_{\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2}}(\Delta_{\hat{S}_{\mathcal{M}_2}},\Delta_{\hat{L}_{\mathcal{M}_2}})\|_\infty \\ \leq \; & \|\mathbf{C}_{\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}}(\Delta_{\tilde{S}_{\mathcal{M}_2}},\Delta_{\tilde{L}_{\mathcal{M}_2}}) - \mathbf{C}_{\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}}(\Delta_{\hat{S}_{\mathcal{M}_2}},\Delta_{\hat{L}_{\mathcal{M}_2}})\|_\infty \qquad (3.71) \\ & + \|\mathbf{C}_{\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}}(\Delta_{\hat{S}_{\mathcal{M}_2}},\Delta_{\hat{L}_{\mathcal{M}_2}}) - \mathbf{C}_{\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2}}(\Delta_{\hat{S}_{\mathcal{M}_2}},\Delta_{\hat{L}_{\mathcal{M}_2}})\|_\infty \qquad (3.72) \end{aligned}$$

We present upper bounds for (3.71) and (3.72) separately. For (3.71), using similar arguments as those in the Proof of Lemma 1, we have that with probability converging to 1, $\mathbf{C}_{\tilde{S}_{\mathcal{M}_2},\tilde{L}_{\mathcal{M}_2}}(\cdot,\cdot)$ is a Lipschitz operator with an $O(\gamma_N^{1-\eta})$ Lipschitz constant. That is,

$$(3.71) \leq_P \kappa\gamma_N^{1-\eta} \times \max\Big(\|\hat{S}_{\mathcal{M}_2} - \tilde{S}_{\mathcal{M}_2}\|_\infty, \|\hat{L}_{\mathcal{M}_2} - \tilde{L}_{\mathcal{M}_2}\|_\infty\Big).$$

We proceed to an upper bound of (3.72). We notice that with a probability converging to 1, $\mathbf{C}_{(S,L)}(\Delta_{\hat{S}_{\mathcal{M}_2}},\Delta_{\hat{L}_{\mathcal{M}_2}})$ is Lipschitz in (S,L). Moreover, according to (3.34), $\|\Delta_{\hat{S}_{\mathcal{M}_2}}\|_\infty \leq \gamma_N^{1-2\eta}$ and $\|\Delta_{\hat{L}_{\mathcal{M}_2}}\|_\infty \leq \gamma_N^{1-\eta}$, we have

$$\|\nabla R_N(\Delta_{\tilde{S}_{\mathcal{M}_2}} + \Delta_{\tilde{L}_{\mathcal{M}_2}})\|_\infty \leq O_P(\frac{1}{\sqrt{N}}). \qquad (3.73)$$

Combining the above display with the fact that $\mathbf{F}_L$ and $\mathbf{P}_{T_L\mathfrak{L}}$ are locally Lipschitz in L, we have that

$$(3.72) \leq_P \kappa\gamma_N \times \max\Big(\|\hat{S}_{\mathcal{M}_2} - \tilde{S}_{\mathcal{M}_2}\|_\infty, \|\hat{L}_{\mathcal{M}_2} - \tilde{L}_{\mathcal{M}_2}\|_\infty\Big).$$

We combine the upper bounds for (3.71) and (3.72) with the equation (3.70), and arrive at

$$\max\Big(\|\hat{S}_{\mathcal{M}_2}-\tilde{S}_{\mathcal{M}_2}\|_\infty, \|\hat{L}_{\mathcal{M}_2}-\tilde{L}_{\mathcal{M}_2}\|_\infty\Big) \leq_P 2\kappa\gamma_N^{1-\eta} \times \max\Big(\|\hat{S}_{\mathcal{M}_2}-\tilde{S}_{\mathcal{M}_2}\|_\infty, \|\hat{L}_{\mathcal{M}_2}-\tilde{L}_{\mathcal{M}_2}\|_\infty\Big).$$

Consequently,

$$\hat{S}_{\mathcal{M}_2} =_P \tilde{S}_{\mathcal{M}_2} \text{ and } \hat{L}_{\mathcal{M}_2} =_P \tilde{L}_{\mathcal{M}_2}.$$

We proceed to prove (3.44). According to (3.68) and (3.73), we have

$$(\hat{S}_{\mathcal{M}_2} - S^*, \hat{L}_{\mathcal{M}_2} - L^*) = \gamma_N \mathbf{F}^{-1}_{\hat{L}_{\mathcal{M}_2}} \mathbf{q}_{\hat{U}_{1,\mathcal{M}_2}} + o_P(\gamma_N)$$

where $\mathbf{q}_{\hat{U}_{1,\mathcal{M}_2}} = (\text{sign}(\mathbf{O}(S^*)), \rho\hat{U}_{1,\mathcal{M}_2}\hat{U}_{1,\mathcal{M}_2}^\top)$. Because both $\mathbf{F}_{\hat{L}_{\mathcal{M}_2}}$ and $\mathbf{q}_{\hat{U}_{1,\mathcal{M}_2}}$ are continuous in $\hat{U}_{1,\mathcal{M}_2}$, and $\|\hat{U}_{1,\mathcal{M}_2} - U_1^*\| \leq \gamma_N^{1-\eta}$, we have

$$(\hat{S}_{\mathcal{M}_2} - S^*, \hat{L}_{\mathcal{M}_2} - L^*) = \gamma_N \mathbf{F}^{-1}\mathbf{q}_{U_1^*} + o_P(\gamma_N).$$

□

Proof of Lemma 5. According to (3.48), we have

$$\mathbf{P}_{\mathcal{S}^{*\perp}}\partial_S H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})} = \{\mathbf{P}_{\mathcal{S}^{*\perp}}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) + \gamma_N \mathbf{P}_{\mathcal{S}^{*\perp}}(\text{sign}(\mathbf{O}(S^*))) + \gamma_N \mathbf{P}_{\mathcal{S}^{*\perp}}W : \|W\|_\infty \leq 1 \text{ and } W \in \mathcal{S}^{*\perp}\}.$$

Notice that $\mathbf{P}_{\mathcal{S}^{*\perp}}(\text{sign}(\mathbf{O}(S^*))) = \mathbf{0}_{J\times J}$ and $\mathbf{P}_{\mathcal{S}^{*\perp}}W = W$ for $W \in \mathcal{S}^{*\perp}$. Therefore, we have

$$\begin{aligned} lll\mathbf{0}_{J\times J} \in \mathbf{P}_{\mathcal{S}^{*\perp}}\partial_S H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})} &\iff \exists W \in \mathcal{S}^* \text{ such that } \|W\|_\infty \leq 1 \\ &\quad \text{and } \mathbf{P}_{\mathcal{S}^{*\perp}}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) = -\gamma_N W. \\ &\iff \|\mathbf{P}_{\mathcal{S}^{*\perp}}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2})\|_\infty \leq \gamma_N. \end{aligned}$$

Similarly, according to (3.49), we have

$$\begin{aligned} &\mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L})^\perp}\partial_L H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})} \\ =&\{\mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L})^\perp}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) + \delta_N \hat{U}_{2,\mathcal{M}_2}W\hat{U}_{2,\mathcal{M}_2}^\top : \|W\|_2 \leq 1\}. \end{aligned}$$

Consequently,

$$\begin{aligned} &\mathbf{0}_{J\times J} \in \mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L})^\perp}\partial_L H|_{(\hat{S}_{\mathcal{M}_2},\hat{L}_{\mathcal{M}_2})} \\ \iff &\exists W \text{ such that } \|W\|_2 \leq 1 \text{ and } \mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L})^\perp}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) = -\hat{U}_{2,\mathcal{M}_2}W\hat{U}_{2,\mathcal{M}_2}^\top. \\ \iff &\|\mathbf{P}_{(T_{\hat{L}_{\mathcal{M}_2}}\mathfrak{L})^\perp}\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2})\|_2 \leq \delta_N. \end{aligned}$$

These two equivalent expressions concludes our proof. □

Proof of Lemma 6. Let

$$\tilde{S} = \tilde{S}_{\mathcal{M}_2} + \tilde{S}_{\mathcal{S}^{*\perp}},$$

where $\tilde{S}_{\mathcal{M}_2} \in \mathcal{S}^*$ and $\tilde{S}_{\mathcal{S}^{*\perp}} \in \mathcal{S}^{*\perp}$, and

$$\tilde{L} = \tilde{L}_{\mathcal{M}_2} + \tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top},$$

where $\tilde{L}_{\mathcal{M}_2} = \tilde{U}_{1,\mathcal{M}_2}\tilde{D}_{1,\mathcal{M}_2}\tilde{U}_{1,\mathcal{M}_2}^\top$ and $\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top} = \tilde{U}_{2,\mathcal{M}_2}\tilde{D}_{2,\mathcal{M}_2}\tilde{U}_{2,\mathcal{M}_2}^\top$. Also $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}) \in \mathcal{M}_2$ and $\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top} \in (T_{\tilde{L}}\mathfrak{L})^\top$. Similar to (3.48) and (3.49) we have the sub-differentials of $H(S, L)$ at $(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})$,

$$\begin{aligned}
&\partial_S H|_{(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})} \\
=&\{\nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2}) + \gamma_N \text{sign}(\mathbf{O}(S^*)) + \gamma_N W_1 : \|W_1\|_\infty \leq 1, \text{ and } W_1 \in \mathcal{S}^{*\perp}\},
\end{aligned}$$

and

$$\partial_L H|_{(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})} = \{\nabla h_N(\hat{S}_{\mathcal{M}_2} + \hat{L}_{\mathcal{M}_2}) + \delta_N \tilde{U}_{1,\mathcal{M}_2}\tilde{U}_{1,\mathcal{M}_2}^\top + \delta_N \tilde{U}_{2,\mathcal{M}_2} W_2 \tilde{U}_{2,\mathcal{M}_2}^\top : \|W_2\|_2 \leq 1\}.$$

Let $W_1 = \text{sign}(\tilde{S}_{\mathcal{S}^{*\perp}})$ and $W_2 = \text{sign}(\tilde{D}_{2,\mathcal{M}_2})$ in the above expressions for sub-differentials. According to the definition of sub-differential, we have

$$\begin{aligned}
&H(\tilde{S}, \tilde{L}) - H(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}) \\
\geq&\nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2}) \cdot (\tilde{S}_{\mathcal{S}^{*\perp}} + \tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}) + \gamma_N \|\tilde{S}_{\mathcal{S}^{*\perp}}\|_1 + \delta_N \|\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}\|_*.
\end{aligned}$$

Because $\tilde{S}_{\mathcal{S}^{*\perp}} \in \mathcal{S}^{*\perp}$ and $\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top} \in (T_{\tilde{L}}\mathfrak{L})^\top$, we further expand the above inequality,

$$\begin{aligned}
&H(\tilde{S}, \tilde{L}) - H(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2}) \\
\geq&\mathbf{P}_{\mathcal{S}^{*\perp}} \nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2}) \cdot \tilde{S}_{\mathcal{S}^{*\perp}} + \mathbf{P}_{(T_{\tilde{L}}\mathfrak{L})^\perp} \nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2}) \cdot \tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top} \qquad (3.74) \\
&+ \gamma_N(\|\tilde{S}_{\mathcal{S}^{*\perp}}\|_1 + \rho\|\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}\|_*).
\end{aligned}$$

We provide a lower bound for the right-hand-side of (3.74). According to (3.55) and (3.56) and the definition of $\mathbf{F}^\perp$, we have

$$\begin{aligned}
&\Big(\mathbf{P}_{\mathcal{S}^{*\perp}} \nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2}), \mathbf{P}_{(T_{\tilde{L}}\mathfrak{L})^\top} \nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2})\Big) \\
=&\gamma_N \mathbf{F}^\perp \mathbf{F}^{-1} \Big(\text{sign}(\mathbf{O}(S^*)), \rho U_1^* U_1^{*T}\Big) + o_P(\gamma_N).
\end{aligned}$$

According to Assumption A6 and the above expression, we have

$$\begin{aligned}
|\mathbf{P}_{\mathcal{S}^{*\perp}} \nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2}) \cdot \tilde{S}_{\mathcal{S}^{*\perp}}| &<_P \gamma_N \|\tilde{S}_{\mathcal{S}^{*\perp}}\|_\infty \\
|\mathbf{P}_{(T_{\tilde{L}}\mathfrak{L})^\perp} \nabla h_N(\tilde{S}_{\mathcal{M}_2} + \tilde{L}_{\mathcal{M}_2}) \cdot \tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}| &<_P \delta_N \|\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}\|_2. \qquad (3.75)
\end{aligned}$$

We proceed to the L_1 penalty term. It has a lower bound

$$\|\tilde{S}_{\mathcal{S}^{*\perp}}\|_1 \geq \|\tilde{S}_{\mathcal{S}^{*\perp}}\|_\infty. \tag{3.76}$$

For the nuclear norm term, we have

$$\|\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}\|_* = \|\tilde{D}_{2,\mathcal{M}_2}\|_* \geq \|\tilde{D}_{2,\mathcal{M}_2}\|_\infty = \|\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}\|_2 \tag{3.77}$$

The first equality in the above display is due to the definition of $\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}$. The inequality and second equality hold because $\tilde{D}_{2,\mathcal{M}_2}$ is a diagonal matrix and its nuclear norm is the same as L_1 norm, and its spectral norm is the same as its maximum norm. We combine (3.74), (3.75), (3.76) and (3.77), then

$$H(\tilde{S}, \tilde{L}) >_P H(\tilde{S}_{\mathcal{M}_2}, \tilde{L}_{\mathcal{M}_2})$$

with probability converging to 1, as long as $\|\tilde{S}_{\mathcal{S}^{*\perp}}\|_\infty > 0$ or $\|\tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top}\|* > 0$. Because $(\tilde{S}, \tilde{L})$ is a solution to (3.45), the above statement implies $\tilde{S}_{\mathcal{S}^{*\perp}} = \tilde{L}_{(T_{\tilde{L}}\mathfrak{L})^\top} = \mathbf{0}_{J\times J}$. Therefore, $\tilde{S} = \tilde{S}_{\mathcal{M}_2}$, $\tilde{L} = \tilde{L}_{\mathcal{M}_2}$, and $(\tilde{S}, \tilde{L}) \in \mathcal{M}_2$, □

Proof of Lemma 7. We first investigate the linear space $T_L\mathfrak{L}$,

$$T_L\mathfrak{L} = \Big\{ [U_1, U_2] \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & \mathbf{0}_{(J-K)\times(J-K)} \end{bmatrix} [U_1, U_2]^\top : \\ Y_{11} \text{ is a } K \times K \text{ symmetric matrix, and } Y_{12} = Y_{21}^\top \text{ is a } J \times (J-K) \text{ matrix.} \Big\} \tag{3.78}$$

For any symmetric matrix M, let $N = U^\top MU$. Then N is symmetric and $M = UNU^\top$. We write M as

$$M = [U_1, U_2] \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} [U_1, U_2]^\top.$$

Therefore,

$$\mathbf{P}_{T_L\mathfrak{L}} M = [U_1, U_2] \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & \mathbf{0}_{(J-K)\times(J-K)} \end{bmatrix} [U_1, U_2]^\top,$$

which is Lipschitz in $[U_1, U_2]$. Because U_2 is orthogonal to U_1, we could choose U_2 such that U_2 is also Lipschitz in U_1. As a result,the operator $\mathbf{P}_{T_L\mathfrak{L}}$ is Lipschitz in U_1. Similarly for

$\mathbf{P}_{\mathcal{D}}$, we have

$$\mathcal{D} = \{U_1 D_1 U_1^\top : D_1 \text{ is a } K \times K \text{ diagonal matrix}\}.$$

For the same symmetric matrix M and N discussed before, we have

$$\mathbf{P}_{\mathcal{D}} M = [U_1, U_2] \begin{bmatrix} diag(N_{11}) & \mathbf{0}_{(J-K)\times K} \\ \mathbf{0}_{J\times(J-K)} & \mathbf{0}_{(J-K)\times(J-K)} \end{bmatrix} [U_1, U_2]^\top,$$

where $diag(N_{11})$ is the diagonal components of N_{11}. Therefore, the above display is also Lipschitz continuous in U_1. □

Proof of Lemma 8. Lemma 8 is an direct application of Parlett (1980) Theorem 4.5.1, Theorem 11.7.1 and Assumption A2. We omit the details. □

Chapter 4

Data-Driven Learning of Q-Matrix

4.1 Introduction

Psychological and educational tests are often conducted to investigate multiple latent traits or skills by making use of dichotomous-response or categorical-response items. A key element in such tests is the relationship between the items and the latent traits. It is conventional to pre-specify the relationship by experts' prior knowledge of the items and of the latent traits. The correct specification of latent traits associated with each item is crucial both for the model parameter calibration and for the measurement of individuals. In particular, there is a large body of literature in confirmatory factor analysis on verifying the prespecified loading structure (see e.g. Kline, 2015) and the effect of Q-matrix misspecification has also been investigated in the literature of DCMs (e.g. Rupp and Templin, 2008a).

Learning the Q-matrix based on the data is related to the problem of finding simple loading structure in exploratory factor analysis. Exploratory factor analysis is a statistical method used to uncover the underlying structure of a set of items. It is commonly used by researchers when developing a scale (a scale is a collection of items used to measure a particular research topic) and serves to identify a set of latent constructs underlying a battery of measured variables (e.g. Fabrigar et al., 1999). In what follows, we briefly discuss how the Q-matrix is implicitly constructed in exploratory factor analysis, although the name "Q-matrix" is not used. For simplicity, we only consider the M2PL model, in

which the marginal distribution $f(\boldsymbol{\theta})$ is a standard normal distribution. The marginal likelihood becomes

$$L(A, \mathbf{d}) = \prod_{i=1}^{N} \int \prod_{j=1}^{J} \frac{1}{\sqrt{(2\pi)^K}} \frac{e^{(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}) y_{ij}}}{1 + e^{d_j + \mathbf{a}_j^\top \boldsymbol{\theta}}} e^{-\frac{1}{2}\|\boldsymbol{\theta}\|^2} d\boldsymbol{\theta}.$$

Let

$$(\hat{A}, \hat{\mathbf{d}}) = \arg\max_{A, \mathbf{d}} L(A, \mathbf{d})$$

be a maximal likelihood estimate of the model parameters. It is not difficult to observe that due to the rotational invariancy of the standard normal distribution, for any $K \times K$ orthonormal matrix R,

$$L(\hat{A}, \hat{\mathbf{d}}) = L(\hat{A}R, \hat{\mathbf{d}}).$$

Therefore, for any given solution with two or more factors, there exists an infinite number of alternative orientations of the factors in the multidimensional space that will explain the data equally well. In other words, the loading matrix A is not identifiable. This rotational indeterminacy of the loading matrix A also appears in the proposed latent and undirected graphical measurement models in Chapter 3 and the varimax rotation is applied in the real data analysis in Section 3.7. Rotational methods are proposed in exploratory factor analysis to find a rotation such that the resulting loading matrix has as many entries close to zero as possible, such as the varimax rotation used in Section 3.7. The idea of simple loading structure comes from Thurstone (1947) and Cattell (1978) and we refer to Browne (2001) for a review of popular rotational methods. Once a simple loading matrix $\hat{A}$ is found, exploratory factor analysis implicitly reconstructs the Q-matrix by hard thresholding $\hat{A}$ using a prespecified threshold such as 0.1 or 0.3 (Osborne, 2015; Fabrigar et al., 1999). The latent factors are interpreted based on the thresholded $\hat{A}$. There is not enough theoretical justification on such a procedure and the procedure is very sensitive to the prespecified threshold.

Furthermore, the procedure discussed above is limited to exploratory factor analysis, for which rotational invariancy exists. What if other measurement models are considered? For example, diagnostic classification models and multidimensional IRT models when $\boldsymbol{\theta}$ is not normally distributed. Therefore, a rigorous treatment is needed for the problem of Q-matrix estimation. To tackle this problem, we provide a statistical framework, under which finding

the Q-matrix becomes a latent variable selection problem. It has several advantages. First, the problem of finding Q-matrix now becomes a statistical problem. Under this framework, there is an underlying true Q^* and the statistical question now becomes if we are able to come up with some estimator $\hat{Q}$ such that $\hat{Q} = Q^*$ with high probability as sample size N grows to infinity. Second, such a $\hat{Q}$ can be obtained based on a computationally tractable and efficient method that is to be discussed in a sequel.

4.2 Learning Q-matrix as a Latent Variable Section Problem

In this section, we will formulate the problem of estimating Q-matrix as a latent variable selection problem. We focus on the log-linear models for binary data, but the idea can be easily extended to categorical data and to other latent variable models. Under the log-linear models, there is a one-to-one correspondence between the Q-matrix and the zero patterns of the item parameters $\boldsymbol{\beta} = (\beta_{k_1 \cdots k_T, j} : \{k_1, \cdots, k_T\} \subseteq \{1, ..., K\}, j = 1, 2, ..., J)$. More precisely, for a specific entry of Q,

$$q_{jk} = 1 \text{ if and only if } \exists\ \beta_{k_1 \cdots k_T} \neq 0 \text{ and } k \in \{k_1, \cdots, k_T\}.$$

For specific IRT models or DCMs, the relationship could be even simpler, due to their reduced forms. For example, for the M2PL model,

$$q_{jk} = 1 \text{ if and only if } a_{jk} = 0,$$

where $A = (a_{jk})$ is the factor loading matrix. For the DINA model,

$$\beta_{k_1 \cdots k_T, j} \neq 0$$

if and only if q_{jk_1}, ..., q_{jk_T} are all the nonzero entries of $\mathbf{q}_j = (q_{j1}, ..., q_{jK})$.

Following the discussion above, each Q-matrix corresponds to a zero pattern of the regression coefficient $\boldsymbol{\beta}$. Therefore, learning the Q-matrix is equivalent to identifying the zero regression coefficients. There is a vast literature on variable and model selection, most of which are developed for linear and generalized linear models. Technically speaking, the log linear model (2.2) is a generalized linear mixed model with $\theta_1, ..., \theta_K$ and their interactions being the random covariates and $\boldsymbol{\beta}$ being the regression coefficients. We would employ

variable selection methods for the Q-matrix estimation. Notice that the current setup is different from the regular regression setting in that the covariates θ_k's are not directly observed. Therefore the variables to be selected are all latent. In what follows, we apply this idea to specific models, including the DINA, DINO and M2PL models. Extensions to more general situations will be discussed.

4.3 Learning Q-matrix for the DINA and DINO Models

Due to the latent nature of the attribute profiles, when and whether the Q-matrix and other models parameters can be estimated consistently by the observed data is a challenging problem. Furthermore, theoretical results on the identifiability usually do not imply practically feasible estimation procedures. In this section, we first propose sufficient conditions under which the Q-matrix can be recovered for the DINA and DINO models. These sufficient conditions provide a guideline for designing diagnostic tests. Then an implementable estimation procedure is constructed for learning the Q-matrix.

4.3.1 Identifiability of the Q-matrix

We consider two matrices Q and Q' that are identical if we appropriately rearrange the orders of their columns. Each column in the Q-matrix corresponds to an attribute. Reordering the columns corresponds to relabeling the attributes and it does not change the model. Upon estimating the Q-matrix, the data do not contain information about the specific meaning of each attribute. Therefore, one cannot differentiate Q and Q' solely based on data if they are identical up to a column permutation. For this sake, we present the following equivalent relation.

Definition 1. *We write $Q \sim Q'$ if and only if Q and Q' have identical column vectors that could be arranged in different orders; otherwise, we write $Q \nsim Q'$.*

Definition 2. *We say that Q is identifiable if there exits an estimator $\hat{Q}$ such that*

$$\lim_{N\to\infty} P(\hat{Q} \sim Q) = 1.$$

We point out that the identifiability of the Q-matrix requires conditions. In what follows, we provide an example, for which the Q-matrix is not identifiable based on the data.

Example 3. *Let* $\mathbf{s} = (0,0)^\top$ *and* $\mathbf{g} = (0,0)^\top$. *In addition, we let*

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

$$\mathbf{p} = (p_{00}, p_{10}, p_{01}, p_{11})^\top = (0.25, 0.25, 0.25, 0.25)^\top$$

and

$$\mathbf{p}' = (p'_{00}, p'_{10}, p'_{01}, p'_{11})^\top = (0.50, 0.25, 0.00, 0.25)^\top$$

Then the DINA with parameter $(\mathbf{s}, \mathbf{g}, Q, \mathbf{p})$ *and the DINA model with parameter* $\mathbf{s}, \mathbf{g}, Q', \mathbf{p}'$ *have the same response probability:*

$$P(\mathbf{Y} = (0,0)^\top) = 0.5, \quad P(\mathbf{Y} = (1,0)^\top) = 0.25,$$
$$P(\mathbf{Y} = (0,1)^\top) = 0, \quad P(\mathbf{Y} = (1,1)^\top) = 0.25.$$

In other words, the Q-matrix is not identifiable in this situation.

Given a response vector $\mathbf{y} = (y_1, \cdots, y_J)^\top$, the likelihood function of a diagnostic classification model can be written as

$$L(\boldsymbol{\beta}, \mathbf{p}, Q) = \sum_{\boldsymbol{\theta} \in \{0,1\}^K} p_{\boldsymbol{\theta}} \prod_{j=1}^{J} P(Y_j = 1 | \boldsymbol{\beta}, \boldsymbol{\theta}, Q)^{y_j} (1 - P(Y_j = 1 | \boldsymbol{\beta}, \boldsymbol{\theta}, Q))^{1-y_j}.$$

Definition 3 (Definition 11.2.2 in Casella and Berger (2002))**.** *For a given* Q*, we say that the model parameters* $\boldsymbol{\beta}$ *and* $\mathbf{p}$ *are identifiable if distinct values of* $(\boldsymbol{\beta}, \mathbf{p})$ *yield different distributions of* $\mathbf{Y}$*, i.e., there is no* $(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{p}}) \neq (\boldsymbol{\beta}, \mathbf{p})$ *such that* $L(\boldsymbol{\beta}, \mathbf{p}, Q) \equiv L(\tilde{\boldsymbol{\beta}}, \tilde{\mathbf{p}}, Q)$ *for all* $\mathbf{Y} \in \{0,1\}^J$.

Let $\hat{Q}$ be a consistent estimator. Notice that the Q-matrix is a discrete parameter. The uncertainty of $\hat{Q}$ in estimating Q is not captured by its standard deviation or confidence interval type of statistics. It is more natural to consider the probability $P(\hat{Q} \nsim Q)$ that is usually very difficult to compute. Nonetheless, it is believed that $P(\hat{Q} \nsim Q)$ decays exponentially fast as the sample size (total number of subjects) tends to infinity. We do

not pursue along this direction. The parameters $\boldsymbol{\beta}$ and $\mathbf{p}$ are both continuous parameters. As long as they are identifiable, the analysis falls into routine inference framework. That is, the maximum likelihood is asymptotically normal centered around the true value and its covariance matrix is the inverse of the Fisher information matrix. In what follows, we present some technical conditions that will be referred to in the subsequent sections. Let $\mathbf{s} = (s_1, ..., s_J)^\top$ and $\mathbf{g} = (g_1, ..., g_J)^\top$ be the vectors of the slipping and guessing parameters.

A1 $\boldsymbol{\theta}_1$,...,$\boldsymbol{\theta}_N$ are independently and identically distributed random vectors following distribution $P(\boldsymbol{\theta}_i = \boldsymbol{\theta}) = p_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \{0,1\}^K$. The population is fully diversified meaning that $p_{\boldsymbol{\theta}} > 0$ for all $\boldsymbol{\theta} \in \{0,1\}^K$.

A2 All items have discriminating power meaning that $1 - s_j > g_j$ for all j.

A3 The true matrix Q_0 is complete meaning that $\{\mathbf{e}_i : i = 1, ..., k\} \subset R_Q$, where R_Q is the set of row vectors of Q and $\mathbf{e}_i$ is a row vector such that the i-th element is one and the rest are zero.

A4 Each attribute is required by at least two items, that is, $\sum_{j=1}^{J} q_{jk} \geq 2$ for all k.

The completeness of the Q-matrix requires that for each attribute there exists at least one item requiring only that attribute. If Q is complete, then we can rearrange row and column orders (corresponding to reordering the items and attributes) such that it takes the following form

$$Q = \begin{pmatrix} \mathcal{I}_K \\ \cdots \end{pmatrix}, \tag{4.1}$$

where matrix $\mathcal{I}_K$ is the $K \times K$ identity matrix. Completeness is an important assumption throughout the subsequent discussion. Without loss of generality, we assume that the rows and columns of the Q-matrix have been rearranged such it takes the above form.

Remark 3. *One of the main objectives of cognitive diagnosis is to identify subjects' attribute profiles. It has been established that completeness is the sufficient and necessary condition for a set of items to consistently identify all types of attribute profiles for the DINA model when the slipping and the guessing parameters are both zero. It is usually recommended to*

use a complete Q-matrix. More discussions regarding this issue can be found in Chiu et al. (2009).

We start the discussion by citing the main result of Liu et al. (2013).

Theorem 3 (Theorem 4.2, Liu et al. (2013))**.** *For the DINA model, if the guessing parameters g_j's are known, under Condition A1, A2, and A3, the Q-matrix is identifiable.*

The first result generalizes Theorem 3 to the DINO model with a known slipping parameter. In addition, we provide sufficient and necessary conditions for the identifiability of the slipping and guessing parameters.

Theorem 4. *For the DINO model with known slipping parameters, under Conditions A1, A2, and A3, the Q-matrix is identifiable; the guessing parameters g_j and the attribute population* $\mathbf{p}$ *are identifiable if and only if Condition A4 holds.*

Furthermore, under the setting of Theorem 3, the slipping parameters s_j and the attribute population parameter $\mathbf{p}$ *are identifiable if and only if Condition A4 holds.*

Theorems 3 and 4 require the knowledge of the slipping parameter (the DINO model) or the guessing parameter (the DINA model). They are applicable under certain situations. In the educational testing context, some testing problems are difficult to guess, for instance, the guessing probability of "$879 \times 234 =$?" is basically zero; for multiple choice problems, if all the choices look "equally correct," then the guessing probability may be set to be one over the number of choices.

We further extend the results to the situation when neither the slipping nor the guessing parameters is known, for which additional conditions are required.

A5 Each attribute of the Q-matrix is associated to at least three items, that is, $\sum_{j=1}^{J} q_{jk} \geq 3$ for all k.

A6 Q has two complete submatrices, that is, for each attribute, there exists at least two items requiring only that attribute. If so, we can appropriately arrange the columns

and rows such that

$$Q = \begin{pmatrix} \mathcal{I}_K \\ \mathcal{I}_K \\ Q_1 \end{pmatrix}. \tag{4.2}$$

Theorem 5. *Under the DINA and DINO models, if Conditions A1, 2, 5, and 6 hold, then Q is identifiable, i.e., one can construct an estimator $\hat{Q}$ such that for all $(\mathbf{s}, \mathbf{g}, \mathbf{p})$*

$$\lim_{N\to\infty} P(\hat{Q} \sim Q) = 1.$$

Theorem 6. *Suppose that Conditions A1, 2, 5, and 6 hold. Then* $\mathbf{s}$, $\mathbf{g}$, *and* $\mathbf{p}$ *are all identifiable.*

Theorems 5 and 6 state the identifiability results of Q and other model parameters. They are nontrivial generalizations of Theorems 3 and 4. As we mentioned in the previous section, given that $\mathbf{s}$, $\mathbf{g}$, and $\mathbf{p}$ are identifiable, their estimation falls into routine analysis. The asymptotic distribution of the maximum likelihood estimator and generalized estimating equation estimators are all asymptotically multivariate normal centered around the true values and their variances can be estimated either by the Fisher information inverse or by the sandwich variance estimators.

The identifiability results only state the existence of a consistent estimator. We present the following corollary that the maximum likelihood estimator is consistent under the conditions required by the above theorems. The maximum likelihood estimator (MLE) takes the following form

$$\hat{Q}_{MLE} = \arg\sup_{Q} \sup_{\mathbf{s},\mathbf{g},\mathbf{p}} L_N(\mathbf{s}, \mathbf{g}, \mathbf{p}, Q), \tag{4.3}$$

where

$$L_N(\mathbf{s}, \mathbf{g}, \mathbf{p}, Q) = \prod_{i=1}^{N} \sum_{\boldsymbol{\theta}\in\{0,1\}^K} p_{\boldsymbol{\theta}} \prod_{j=1}^{J} P(Y_{ij} = 1|\mathbf{s}, \mathbf{g}, \boldsymbol{\theta}, Q)^{y_{ij}} (1 - P(Y_{ij} = 1|\mathbf{s}, \mathbf{g}, \boldsymbol{\theta}, Q))^{1-y_{ij}}.$$

Corollary 1. *Under the conditions of Theorem 5, $\hat{Q}_{MLE}$ is consistent. Moreover, the maximum likelihood estimator of* $\mathbf{s}, \mathbf{g}, \mathbf{p}$

$$(\hat{\mathbf{s}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}) = \arg\sup_{\mathbf{s},\mathbf{g},\mathbf{p}} L_N(\mathbf{s}, \mathbf{g}, \mathbf{p}, Q_{MLE})$$

are asymptotically normal with mean centered at the true parameters and variance being the inverse Fisher information matrix.

Proof of Corollary 1. Based on the results and proofs of Theorems 5 and 6, this corollary is straightforward to develop by means of Taylor expansion of the likelihood. We therefore omit the details. □

To compute the maximum likelihood estimator $\hat{Q}_{MLE}$, one needs to evaluate the profiled likelihood, $\sup_{\mathbf{s},\mathbf{g},\mathbf{p}} L_N(\mathbf{s},\mathbf{g},\mathbf{p},Q)$, for all possible J by K matrices with binary entries. The computation of $\hat{Q}_{MLE}$ induces a substantial computational overhead and is practically impossible to carry out. In the following section, we present a computationally feasible estimator via the regularized maximum likelihood estimator.

Remark 4. *The identifiability results are developed under the situation when there is no information about Q at all. In practice, partial information about the Q-matrix are usually available. For instance, a submatrix for some items (rows) is known and the rest needs to be estimated. This happens when new items are to be calibrated based on existing ones. Sometimes, the submatrix is known for some attributes (columns) and that corresponding to other attributes needs to learn. Under such circumstances, the Q-matrix is much easier to estimate and the conditions are much weaker than those in condition Theorem 5.*

4.3.2 Q-matrix Estimation via a Regularized Likelihood

By making use of the latent variable selection framework in Section 4.2, we construct a computationally feasible estimator via the regularized maximum likelihood, for which there is a large body of literature (Tibshirani, 1996, 1997; Fan and Li, 2001). The applicability of this estimator is not limited to the DINA and DINO models and it can be applied to basically all diagnostic classification models in use. A short list of such models includes DINA-type models (such as the DINA and HO-DINA models), RUM-type models (like the reduced NC-RUM, and C-RUM), and the G-DINA (de la Torre, 2011).

We consider the log-linear model parametrization and the Q-matrix is unknown. For simplicity, we write the item response function

$$c_{j,\boldsymbol{\theta}} = P(Y_j = 1|\boldsymbol{\theta},\boldsymbol{\beta}),$$

which depends on the latent vector $\boldsymbol{\theta}$ and the item parameters $\boldsymbol{\beta}$. The complete data $(\mathbf{y}_i, \boldsymbol{\theta}_i : i = 1, ..., N)$ log-likelihood can be written as

$$l_{com}(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p}; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N) = \sum_{i=1}^{N} \log(p_{\boldsymbol{\theta}_i}) + \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} \log c_{j,\boldsymbol{\theta}_i} + (1 - y_{ij}) \log(1 - c_{j,\boldsymbol{\theta}_i}).$$

And the observed data $(\mathbf{y}_i : i = 1, ..., N)$ log-likelihood is

$$l(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p}) = \sum_{i=1}^{N} \log \left\{ \sum_{\boldsymbol{\theta} \in \{0,1\}^K} p_{\boldsymbol{\theta}} \prod_{j=1}^{J} c_{j,\boldsymbol{\theta}}^{y_{ij}} (1 - c_{j,\boldsymbol{\theta}})^{1-y_{ij}} \right\}.$$

A regularized maximum likelihood estimator of the β-coefficients and $\mathbf{p}$ is given by

$$(\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_J, \hat{\mathbf{p}}) = \arg \max_{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p}} l(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p}) - N \sum_{j=1}^{J} p_{\lambda_j}(\boldsymbol{\beta}_j) \tag{4.4}$$

where p_{λ_j} is some penalty function and λ_j is the regularization parameter. In this paper, we choose p_λ to be either the L_1 penalty or the SCAD penalty (Fan and Li, 2001). In particular, to apply the L_1 penalty, we let

$$p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{k=1}^{K} |\beta_k|$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)$; to apply the SCAD penalty, we let

$$p_\lambda(\boldsymbol{\beta}) = \sum_{k=1}^{K} p_\lambda^S(\beta_k).$$

The function $p_\lambda^S(x)$ is defined as $p_\lambda^S(0) = 0$ and

$$\frac{dp_\lambda^S}{dx}(x) = \lambda \left\{ I(x \leq \lambda) + \frac{\max(0, a\lambda - x)}{(a-1)\lambda} \right\}$$

for $x > 0$; for $x < 0$, the function is $p_\lambda^S(x) = p_\lambda^S(-x)$. There is an additional "a" parameter that is chosen to be $a = 3.7$ as suggested by Fan and Li (2001).

On the consistency of the regularized estimator. A natural issue is whether the consistency results developed in the previous section can be applied to the regularized estimator. The consistency results for the regularized estimator can be established by means of the techniques developed in the literature (Zhao and Yu, 2006; Fan and Lv, 2011;

Fan and Li, 2001). Therefore, we only provide an outline and omit the details. First of all, the parameter dimension is fixed and the sample size becomes large. The regularization parameter is chosen such that $\lambda_j \to 0$ and $\sqrt{N}\lambda_j \to \infty$ as $N \to \infty$. For the DINA (or DINO) model, let Q_1 and Q_2 be two matrices. If $Q_1 \nsim Q_2$, the consistency results in the previous section ensure that the two families of distributions under different Q's are separated. Thus, with probability tending to one, the true matrix Q is the global maximizer of the profiled likelihood. Since $\lambda_j = o(1)$ and the penalty term is of order $o(N)$, the results in the previous section suggests that the regularized likelihood has been to obtained with in ϵ distance from the true value, that is, the consistency results localize the regularized estimator to a small neighborhood of their true values. The oracle properties of the L_1 regularized estimator and SCAD regularized estimator are developed for maximizing the penalized likelihood function locally around the true model parameters (Zhao and Yu, 2006; Fan and Lv, 2011; Fan and Li, 2001). Thus, combining the global results (Q-matrix identifiability) and the local results (oracle condition for the local penalized likelihood maximizer), we obtain that the regularized estimators admit the oracle property in estimating the Q-matrix under the identifiability conditions in the previous section. We mention that for the L_1 regularized estimator irrepresentable condition is needed concerning the Fisher information matrix to ensure the oracle condition (Zhao and Yu, 2006).

4.3.3 Computation via Expectation-Maximization Algorithm

The advantage of the regularized maximum likelihood estimation for the Q-matrix lies in computation. As mentioned previously, the computation of $\hat{Q}_{MLE}$ in (4.3) requires evaluation of the profiled likelihood for all possible Q-matrices and there are $2^{J\times K}$ such matrices. This is computationally impossible even for some practically small J and K. The computation of (4.4) can be done by combining the expectation-maximization (EM) algorithm and the coordinate descent algorithm. In particular, we view $\boldsymbol{\theta}$ as the missing data following the unknown prior distribution $\mathbf{p}$. The EM algorithm consists of two steps.

E-step Let $(\boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)})$ be the parameter values at the tth iteration. The E-step computes function

$$\begin{aligned} & H(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p} | \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)}) \\ = \; & E[l_{com}(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p}; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N) | \mathbf{y}_i, i = 1, ..., N, \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)}] \end{aligned}$$

where the above expectation is taken with respect to $\boldsymbol{\theta}_i$, $i = 1, ..., N$, under the posterior distribution $P(\ \cdot\ | \mathbf{y}_i, i = 1, ..., N, \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)})$. The E-step is a closed form computation. First, the complete data log-likelihood function is additive. Furthermore, under the posterior distribution $\boldsymbol{\theta}_1$,..., $\boldsymbol{\theta}_N$ are jointly independent. Therefore, one only needs to evaluate

$$E[y_{ij} \log c_{j,\boldsymbol{\theta}_i} + (1 - y_{ij}) \log(1 - c_{j,\boldsymbol{\theta}_i}) | \mathbf{y}_i, \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)}]$$

and

$$E[\log(p_{\boldsymbol{\theta}_i}) | \mathbf{y}_i, \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)}]$$

for each $i = 1, ..., N$ and $j = 1, ..., J$. Notice that $\boldsymbol{\theta}$ is a discrete random variable taking values in $\{0, 1\}^K$. Therefore, the posterior distribution of each $\boldsymbol{\theta}_i$ can be computed exactly and the complexity of the above conditional expectation is 2^K that is manageable for K as large as 10 that is a very high dimension for diagnostic classification models in practice. Therefore the overall computational complexity of the E-step is $O(NJ2^K)$.

The M-step consists of maximizing the H-function with the penalty term

$$\max_{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p}} H(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_J, \mathbf{p} | \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)}) - N \sum_{j=1}^{J} p_{\lambda_j}(\boldsymbol{\beta}_j).$$

Before applying the coordinate descent algorithm, we further reduce the dimension. The objective function that is associated with $\boldsymbol{\beta}$ can be written as

$$\sum_{j=1}^{J} \Big\{ \sum_{i=1}^{N} E[y_{ij} \log c_{j,\boldsymbol{\theta}_i} + (1 - y_{ij}) \log(1 - c_{j,\boldsymbol{\theta}_i}) | \mathbf{y}_i, \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)}] - N p_{\lambda_j}(\boldsymbol{\beta}_j) \Big\}.$$

For each j, the term

$$\sum_{i=1}^{N} E[y_{ij} \log c_{j,\boldsymbol{\theta}_i} + (1 - y_{ij}) \log(1 - c_{j,\boldsymbol{\theta}_i}) | \mathbf{y}_i, \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)}] - N p_{\lambda_j}(\boldsymbol{\beta}_j)$$

consists only of $\boldsymbol{\beta}_j$. Thus, $\boldsymbol{\beta}_j$ can be optimized independently. Each $\boldsymbol{\beta}_j$ has 2^K coordinate and we apply the coordinate descent algorithm (developed for generalized linear models) to maximize the above function for each j. For details about this algorithm, see Friedman et al. (2010). Furthermore, $p_{\boldsymbol{\theta}}$ is updated by $\sum_{i=1}^{N} P(\boldsymbol{\theta}_i = \boldsymbol{\theta}|\mathbf{y}_i, \boldsymbol{\beta}_1^{(t)}, ..., \boldsymbol{\beta}_J^{(t)}, \mathbf{p}^{(t)})/N$.

The EM algorithm guarantees a monotone increasing objective function. However, there is no guarantee that the algorithm converges to the global maximum. We empirically found that the algorithm sometimes does stop at a local maximum, especially when λ is large. Therefore, we suggest applying the algorithm with different starting points and select the best.

4.3.4 Further Discussions

It is suggested by the theories that the regularization parameter λ be chosen such that $\lambda \to 0$ and $\sqrt{N}\lambda \to \infty$ that is a wide range. For specific diagnostic classification models, we may have more specific choices of λ. For the DINA and the DINO model, each row of the Q-matrix, corresponding the attribute requirement of one item, maps to two non-zero coefficients. Therefore, we may choose λ_j for each item differently such that the resulted coefficients $\boldsymbol{\beta}_j$ has exactly two non-zero elements.

4.3.5 Simulation Study 1: Independent Attributes

Attribute profiles are generated from the uniform distribution

$$p_{\boldsymbol{\theta}} = 2^{-K}.$$

We consider the cases that $K = 3$ and 4 and $J = 18$ items. The following Q-matrices are adopted

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

These two matrices are chosen such that the identifiability conditions are satisfied. The slipping and guessing parameters are set to be 0.2. All other conditions are also satisfied. For each Q, we consider sample sizes $N = 1000$, 2000, and 4000. For each particular Q, 100 independent data sets are generated to evaluate the performance.

L_1 regularized estimator. The simulation results are summarized in Tables 4.1, 4.2, and 4.3. According to Table 4.1, for both $K = 3$ and 4, our method estimates the Q-matrix almost without error when the sample size is as large as 2000 or more. In addition, the higher the dimension is the more difficult the problem is. Furthermore, for the cases when the estimator misses the Q-matrix, $\hat{Q}$ differs from the true by only one or two rows. We look closer into the estimators in Table 4.2 that reports the proportion of entries correctly

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$K = 3$	38	62	81	19	98	2	100	0
$K = 4$	20	80	48	52	77	23	99	1

Table 4.1: Numbers of correctly estimated Q-matrices among 100 simulations with sample size 500, 1000, 2000, and 4000 for the L_1 regularized estimator.

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$K = 3$	98.1%	99.6%	100.0%	100.0%
$K = 4$	97.7%	98.9%	99.6%	100.0%

Table 4.2: Proportion of entries correctly specified by $\hat{Q}$ for the L_1 regularized estimator

	Q_1					Q_2			
Sample size	500	1000	2000	4000		500	1000	2000	4000
$\hat{Q}_{1:15} = Q_{1:15}$	100	100	100	100	$\hat{Q}_{1:14} = Q_{1:14}$	98	100	100	100
$\hat{Q}_{1:15} \neq Q_{1:15}$	0	0	0	0	$\hat{Q}_{1:14} \neq Q_{1:14}$	2	0	0	0
$\hat{Q}_{16:18} = Q_{16:18}$	38	81	98	100	$\hat{Q}_{15:18} = Q_{15:18}$	20	48	77	99
$\hat{Q}_{16:18} \neq Q_{16:18}$	62	19	2	0	$\hat{Q}_{15:18} \neq Q_{15:18}$	80	52	23	1

Table 4.3: Numbers of correctly estimated $Q_{1:15}$ and $Q_{16:18}$ for Q_1 and numbers of correctly estimated $Q_{1:14}$ and $Q_{15:18}$ for Q_2 among 100 simulations with solutions for the L_1 regularized estimator

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$K = 3$	98	2	100	0	100	0	100	0
$K = 4$	30	70	96	4	100	0	100	0

Table 4.4: Numbers of correctly estimated Q-matrices among 100 simulations with sample size 500, 1000, 2000, and 4000 for the SCAD estimator.

	$K = 500$	$K = 1000$	$K = 2000$	$K = 4000$
$K = 3$	99.9%	100%	100.0%	100.0%
$K = 4$	97.6%	99.9%	100.0%	100.0%

Table 4.5: Proportion of entries correctly specified by $\hat{Q}$ ($CR(\hat{Q})$) for the SCAD estimator.

specified by $\hat{Q}$

$$CR(\hat{Q}) = \max_{Q' \sim Q} \left\{ \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} 1_{\{\hat{q}_{jk} = q'_{jk}\}} \right\}.$$

We empirically found that the row vectors of Q_1 and Q_2 that require three attributes or four attributes are much more difficult to estimate than others. This phenomenon is reflected by Table 4.3, in which the notation $Q_{I_1:I_2}$ represents the submatrix of Q containing row I_1 to row I_2. In fact, for all simulations in this study, most misspecifications are due to the misspecification of the submatrices of Q_1 and Q_2 that the corresponding items require three attributes or more.

SCAD estimator. Under the same setting, we investigate the SCAD estimator. The results are summarized in Tables 4.4 and 4.5. The SCAD estimator performs better than the L_1 regularized estimator. As the sample size is 1000, the SCAD estimator yields better estimates upon comparing Table 4.1 and Table 4.4.

4.3.6 Simulation Study 2: Dependent Attributes

We consider the situation that the attribute profile $\boldsymbol{\theta}$ follows a nonuniform distribution. For each participant, we generate $\xi = (\xi_1, \cdots, \xi_K)$ that is a multivariate normal distribution $N(\mathbf{0}, \Sigma)$, where the covariance matrix Σ has unit variance and has a common correlation $\rho = 0.05$, 0.15 and 0.25. Then the attribute profile $\boldsymbol{\theta}$ is given by

$$\theta_k = \begin{cases} 1 & \text{if } \xi_k \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

We consider $K = 3$ and Q_1 be the Q-matrix. Table 4.6 shows the probability distribution $p_{\boldsymbol{\theta}}$. The slipping and the guessing parameters remain 0.2. The rest of the setting is the same as that of Study 1.

L_1 regularized estimator. The simulation results are summarized in Table 4.7 to 4.9. Based on Table 4.7, the estimation accuracy is improved when the sample size increases. We also observe that the proposed algorithm performs better when ρ increases. A heuristic interpretation is as follows. The row vector of Q tends to be more difficult to estimate when numbers of subjects who are capable and who are not capable to answer are not balanced. The row vector $(1, 1, 1)$ or $(1, 1, 1, 1)$ is the most difficult to estimate because only subjects with attribute profile $(1, 1, 1)$ or $(1, 1, 1, 1)$ are able to solve them and all other subjects are not. According to Table 4.6, as ρ increases, the proportion of subjects with attribute profile $(1, 1, 1)$ or $(1, 1, 1, 1)$ increases, which explains the improvement of the performance. In fact, similar to the situation that $\boldsymbol{\theta}$ follows a uniform distribution, for most simulations in which the $\hat{Q}$ misses the true, $\hat{Q}$ differs from the true at the row vectors whose true value is $(1, 1, 1)$. The results are shown in Table 4.8.

SCAD estimator. Under the same simulation setting, the results of the SCAD estimator are summarized in Tables 4.10 and 4.11. Its performance is empirically better than that of the L_1 regularized estimator. When the sample size is as small as 500, it has a very high probability estimating all the entries of the Q-matrix correctly.

Class	(0,0,0)	(1,0,0)	(0,1,0)	(1,1,0)	(0,0,1)	(1,0,1)	(0,1,1)	(1,1,1)
$\rho = 0.05$	0.137	0.121	0.121	0.121	0.121	0.121	0.121	0.137
$\rho = 0.15$	0.161	0.113	0.113	0.113	0.113	0.113	0.113	0.161
$\rho = 0.25$	0.185	0.105	0.105	0.105	0.105	0.105	0.105	0.185

Table 4.6: The distribution of the latent attributes of the three-dimensional DINA model for $\rho = 0.05, 0.15$ and 0.25

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$\rho = 0.05$	54	46	87	13	99	1	100	0
$\rho = 0.15$	67	33	93	7	100	0	100	0
$\rho = 0.25$	76	24	95	5	100	0	100	0

Table 4.7: Numbers of correctly estimated Q-matrices among 100 simulations for sample sizes 500, 1000, 2000, and 4000 for the L_1 regularized estimator.

		$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$\rho = 0.05$	$\hat{Q}_{1:15} = Q_{1:15}$	100	100	100	100
	$\hat{Q}_{1:15} \neq Q_{1:15}$	0	0	0	0
	$\hat{Q}_{16:18} = Q_{16:18}$	54	87	99	100
	$\hat{Q}_{16:18} \neq Q_{16:18}$	46	13	1	0
$\rho = 0.15$	$\hat{Q}_{1:15} = Q_{1:15}$	100	100	100	100
	$\hat{Q}_{1:15} \neq Q_{1:15}$	0	0	0	0
	$\hat{Q}_{16:18} = Q_{16:18}$	67	93	100	100
	$\hat{Q}_{16:18} \neq Q_{16:18}$	33	7	0	0
$\rho = 0.25$	$\hat{Q}_{1:15} = Q_{1:15}$	100	100	100	100
	$\hat{Q}_{1:15} \neq Q_{1:15}$	0	0	0	0
	$\hat{Q}_{16:18} = Q_{16:18}$	76	95	100	100
	$\hat{Q}_{16:18} \neq Q_{16:18}$	24	5	0	0

Table 4.8: Numbers of correctly estimated $Q_{1:15}$ and $Q_{16:18}$ for Q_1 for the L_1 regularized estimator.

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$\rho = 0.05$	98.5%	99.7%	100.0%	100.0%
$\rho = 0.15$	99.2%	99.8%	100.0%	100.0%
$\rho = 0.25$	99.4%	99.9%	100.0%	100.0%

Table 4.9: Proportion of entries correctly specified by $\hat{Q}$ for the L_1 regularized estimator.

	$N = 500$		$N = 1000$		$N = 2000$		$N = 4000$	
	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$	$\hat{Q} = Q$	$\hat{Q} \neq Q$
$\rho = 0.05$	97	3	100	0	100	0	100	0
$\rho = 0.15$	98	2	100	0	100	0	100	0
$\rho = 0.25$	99	1	100	0	100	0	100	0

Table 4.10: Numbers of correctly estimated Q-matrices among 100 simulations for sample sizes 500, 1000, 2000, and 4000 under the SCAD penalty

	$N = 500$	$N = 1000$	$N = 2000$	$N = 4000$
$\rho = 0.05$	99.7%	100.0%	100.0%	100.0%
$\rho = 0.15$	100.0%	100.0%	100.0%	100.0%
$\rho = 0.25$	100.0%	100.0%	100.0%	100.0%

Table 4.11: Proportion of entries correctly specified by $\hat{Q}$ ($CR(\hat{Q})$) under the SCAD penalty.

4.3.7 Real Data Example: Social Anxiety Disorder Data

The social anxiety disorder data is a subset of the National Epidemiological Survey on Alcohol and Related Conditions (NESARC) (Grant et al., 2003). We consider participants' binary responses (Yes/No) to thirteen diagnostic questions for social anxiety disorder. The questions are designed by the Diagnostic and Statistical Manual of Mental Disorders, 4th ed (American Psychiatric Association, 1994) and are displayed in Table 4.12. Incomplete cases are removed from the data set. The sample size is 5226. To understand the latent structure of social phobia, we fit the compensatory DINO model for $K = 2$, 3, and 4.

We first consider the L_1 penalty and fit the two-dimensional DINO model. The estimates $\hat{Q}$, $\hat{\mathbf{s}}$, and $\hat{\mathbf{g}}$ are summarized as Case $K = 2$ of Table 4.13. In addition, the correlation between the two attributes is 0.47. We further explore the latent structure by considering the three-dimensional DINO model. For the result, $\hat{Q}$, $\hat{\mathbf{s}}$, and $\hat{\mathbf{g}}$ are summarized as Case $K = 3$ of Table 4.13. A similar latent structure as $\hat{Q}$ in Case $K = 3$ of Table 4.13 is found in Iza et al. (2014) through exploratory factor analysis based on an item response theory model, where

ID	Have you ever had a strong fear or avoidance of
1	speaking in front of other people?
2	taking part/ speaking in class?
3	taking part/ speaking at a meeting?
4	performing in front of other people?
5	being interviewed?
6	writing when someone watches?
7	taking an important exam?
8	speaking to an authority figure?
9	eating/drinking in front of other people?
10	having conversations with people you don't know well?
11	going to parties/social gatherings?
12	dating?
13	being in a small group situation?

Table 4.12: The content of 13 items for the social anxiety disorder data.

the item-attribute structure is prespecified. In their study, the three (continuous) factors are interpreted as "public performance", "close scrutiny", and "interaction", which correspond roughly to those in Case $K = 3$ of Table 4.13. Finally, the four-dimensional DINO model is considered. The results are summarized as Case $K = 4$ in Table 4.13. According to the corresponding $\hat{Q}$, the third item group (items 9 - 13) in the three-dimensional model splits into two attributes. Furthermore, item 6 "writing when someone watches" becomes associated with attribute three. For all three models, it is noted that the estimated slipping parameters are relatively large for some items (such as items 6, 9, 12 and 13), which is usually not the case in educational application. Similar phenomena are observed when we analyze other self-reported psychiatric data sets. Such a result may be due to the nature of slipping parameters in this type of applications, where the slipping parameter of a particular item is the probability that the subject doesn't admit the existence of the corresponding symptom, given he or she has the symptom. Such a probability doesn't have to be small. On the other hand, a slipping parameter in educational applications is usually small since

	$K=2$				$K=3$					$K=4$					
ID	$\hat{Q}$		$\hat{s}$	$\hat{g}$	$\hat{Q}$			$\hat{s}$	$\hat{g}$	$\hat{Q}$				$\hat{s}$	$\hat{g}$
1	1	0	0.05	0.54	1	0	0	0.05	0.49	1	0	0	0	0.05	0.49
2	1	0	0.09	0.27	1	0	0	0.11	0.21	1	0	0	0	0.11	0.21
3	1	0	0.13	0.15	1	0	0	0.16	0.09	1	0	0	0	0.16	0.09
4	1	0	0.12	0.30	1	0	0	0.15	0.25	1	0	0	0	0.14	0.25
5	1	0	0.46	0.07	0	1	0	0.29	0.09	0	1	0	0	0.29	0.08
6	0	1	0.66	0.07	0	1	0	0.68	0.06	0	0	1	0	0.56	0.08
7	1	0	0.42	0.22	0	1	0	0.26	0.21	0	1	0	0	0.27	0.20
8	0	1	0.34	0.16	0	1	0	0.30	0.09	0	1	0	0	0.30	0.08
9	0	1	0.68	0.02	0	0	1	0.68	0.02	0	0	1	0	0.58	0.02
10	0	1	0.13	0.21	0	0	1	0.13	0.20	0	0	1	1	0.14	0.16
11	0	1	0.20	0.12	0	0	1	0.17	0.10	0	0	0	1	0.21	0.08
12	0	1	0.59	0.05	0	0	1	0.59	0.05	0	0	1	0	0.47	0.06
13	0	1	0.67	0.01	0	0	1	0.68	0.01	0	0	1	0	0.57	0.01

Table 4.13: The estimated Q-matrix based on L_1 regularization and the slipping and guessing parameters for the two, three and four dimensional DINO model for the social anxiety disorder data.

it represents the probability of responding incorrectly due to carelessness given the subject is capable to answer correctly. Furthermore, we estimate the Q-matrix via SCAD. The estimates are summarized in Tables 4.14 that are similar to those of the L_1 penalty.

	$K=2$				$K=3$					$K=4$					
ID	$\hat{\mathbf{Q}}$		$\hat{s}$	$\hat{g}$	$\hat{Q}$			$\hat{s}$	$\hat{g}$	$\hat{Q}$				$\hat{s}$	$\hat{g}$
1	1	0	0.05	0.54	1	0	0	0.05	0.49	1	0	0	0	0.05	0.49
2	1	0	0.09	0.27	1	0	0	0.11	0.21	1	0	0	0	0.11	0.21
3	1	0	0.13	0.14	1	0	0	0.16	0.09	1	0	0	0	0.16	0.09
4	1	0	0.12	0.30	1	0	0	0.15	0.25	1	0	0	0	0.15	0.25
5	1	0	0.46	0.07	0	1	0	0.29	0.09	0	1	0	0	0.30	0.08
6	0	1	0.65	0.07	0	1	0	0.68	0.06	0	0	1	0	0.55	0.07
7	1	1	0.43	0.20	0	1	0	0.26	0.21	0	1	0	0	0.27	0.20
8	0	1	0.33	0.16	0	1	0	0.30	0.09	0	1	0	0	0.31	0.08
9	0	1	0.68	0.02	0	0	1	0.68	0.02	0	0	0	1	0.68	0.02
10	0	1	0.13	0.21	0	0	1	0.13	0.20	0	0	0	1	0.11	0.19
11	0	1	0.20	0.13	0	0	1	0.17	0.10	0	0	0	1	0.16	0.09
12	0	1	0.59	0.05	0	0	1	0.59	0.05	0	0	1	0	0.44	0.05
13	0	1	0.67	0.01	0	0	1	0.68	0.01	0	0	1	0	0.55	0.01

Table 4.14: The estimated Q-matrix based on SCAD regularization and the slipping and guessing parameters for the two, three and four dimensional DINO model for the social anxiety disorder data.

4.4 Learning Q-matrix for the M2PL Model

4.4.1 Latent Variable Selection via L_1 Regularized Regression

A regularized estimator. We consider to learn the design matrix Q in the M2PL model. Following the discussion in Section 4.2, it is equivalent to find the support of the factor loading matrix A:

$$Q = \left(1_{\{a_{jk}\neq 0\}}\right).$$

This analysis relies on the belief that each latent factor is only associated with a few items, which comes from the idea of simple loading structures proposed by Thurstone (1947) and Cattell (1978). Under this belief, the true Q-matrix has many zero entries and therefore the factor loading matrix A is sparse. Such sparse pattern is important information for the model selection.

Suppose that the responses of N examinees have been collected. The latent traits $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N$ are independently and identically distributed following the prior distribution $N(0, \Sigma)$ whose density is denoted by $f(\boldsymbol{\theta})$. For this moment, we assume Σ is known and the situation that Σ is unknown will be discussed in a sequel. Given $\boldsymbol{\theta}_i$, $\mathbf{Y}_i$ is assumed to follow the M2PL model (2.3). Then, the complete data log-likelihood for the M2PL model is

$$l_{com}(A, \mathbf{d}; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N) = \sum_{i=1}^{N} \log f(\boldsymbol{\theta}_i) + \sum_{i=1}^{N}\sum_{j=1}^{J}(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i)y_{ij} - \log(1 + e^{d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i}). \quad (4.5)$$

Furthermore, the log-likelihood of the observe responses is given by

$$l(A, \mathbf{d}) = \sum_{i=1}^{N} \log \int f(\boldsymbol{\theta}) \prod_{j=1}^{J} \frac{e^{(d_j + \mathbf{a}_j^\top \boldsymbol{\theta})y_{ij}}}{1 + e^{d_j + \mathbf{a}_j^\top \boldsymbol{\theta}}} d\boldsymbol{\theta}. \quad (4.6)$$

In the exploratory factor analysis, one maximizes the log-likelihood function and obtains the maximum likelihood estimator

$$(\tilde{A}, \tilde{\mathbf{d}}) = \arg\max_{A, \mathbf{d}} \quad l(A, \mathbf{d}).$$

The maximum likelihood estimator does not directly serve the purpose of variable selection. For variable selection, we further consider the L_1 regularized estimator

$$(\hat{A}^\eta, \hat{\mathbf{d}}^\eta) = \arg\max_{A, \mathbf{d}} \Big\{ l(A, \mathbf{d}) - N\eta \|A\|_1 \Big\}, \quad (4.7)$$

where $\eta > 0$ and

$$\|A\|_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}|a_{jk}|.$$

The regularization parameter η controls sparsity. By choosing $\eta = 0$, the L_1 regularized estimator $(\hat{A}^\eta, \hat{\mathbf{d}}^\eta)$ recovers the maximum likelihood estimator that almost surely contains all nonzero estimates of the a-coefficients and it corresponds to no sparsity. On the other hand, by choosing η sufficiently large (for instance, $\eta = \infty$), the corresponding estimate of the discrimination parameters are $\hat{A}^\infty = 0$. In this case, any nonzero discrimination parameter a_{jk} would make the penalized log-likelihood negative infinity. Thus, $\eta = \infty$ corresponds to complete sparsity. Generally speaking, the regularization parameter η controls the sparsity and large values of η lead to more sparse estimates of A. Ideally, we hope to find an appropriate $\eta \in (0, +\infty)$, under which the zero patterns of $\hat{A}^\eta$ are consistent with the true loading structure.

Choice of regularization parameter η. We apply the Bayesian information criterion to choose the sparsity parameter η. In particular, each choice of η results in a selected model that corresponds to a BIC value. Then, we choose the parameter η that leads to the smallest BIC value. More precisely, let

$$\mathcal{M}^\eta = \Big\{(A, \mathbf{d}) : a_{jk} = 0 \text{ if } \hat{a}^\eta_{jk} = 0\Big\}$$

be the selected model corresponding to $\hat{A}^\eta$. For each model $\mathcal{M}^\eta$, the Bayesian information criterion is defined as

$$\text{BIC}(\mathcal{M}^\eta) = -2 \max_{(A,\mathbf{d})\in\mathcal{M}^\eta} l(A, \mathbf{d}) + |\mathcal{M}^\eta| \log N, \tag{4.8}$$

where the above maximized likelihood is subject to the constraint that A has the same zero pattern as $\hat{A}^\eta$. $|\mathcal{M}^\eta|$ counts the number of free parameters in $\mathcal{M}^\eta$:

$$|\mathcal{M}^\eta| = \sum_{j=1}^{J}\sum_{k=1}^{K} 1_{\{\hat{a}^\eta_{jk}\neq 0\}} + J,$$

where the first term is counting the number of free parameters in A and the second term J is the number of parameters in $\mathbf{d}$. The regularization parameter η is chosen to be the one

admitting the smallest BIC value, that is,

$$\eta_* = \arg\min_{\eta} \mathrm{BIC}(\mathcal{M}^{\eta}).$$

Our estimate of the Q-matrix is

$$\hat{Q} = \left(1_{\{\hat{a}_{jk}^{\eta_*} \neq 0\}}\right).$$

Remark 5. *To guarantee parameter identifiability, some constraints need to be imposed on the item parameters. As summarized by Béguin and Glas (2001), there are typically two ways to impose constraints. One is to set $a_{jk} = 0$ for $j = 1, \cdots, K-1$ and $k = j+1, \cdots, K$ (Fraser and McDonald, 1988), which is similar to the constraint of Jöreskog (1969). The other is to set $a_{jj} = 1$ and $a_{jk} = 0$ for $j = 1, \cdots, K$, $k = 1, \cdots, K$, and $j \neq k$. Note that for the former constraint, rotating the parameter space is usually necessary for the interpretation of the factor patterns (Bolt and Lall, 2003; Cai, 2010).*

We adopt a similar approach as the second. In particular, each of the first K items is associated with only one trait, that is $a_{ii} \neq 0$ and $a_{ij} = 0$, for $1 \leq i \neq j \leq K$. This corresponds to the fact that a sub-matrix of Q is known to be the identity matrix (after appropriate reordering of the rows and the columns), but the coefficients a_{ii}'s are not necessarily unity. We further restrict the variances of $\boldsymbol{\theta}_i$ be unity.

In practice, one may impose different constraints on A or Q to ensure identifiability. In the simulation study and the real data analysis, we experimented two different sets of constraints and found that the results are similar. The second constraint is as follows. We identify K items (e.g. the first K items) and let $a_{ii} \neq 0$ for $i = 1, ..., K$. Unlike the first constraint, we do not force a_{ij} $(i \neq j)$ to be zero. Rather, we impose L_1 penalties on them. Thus, the penalty includes all elements in A except for a_{ii} for $i = 1, ..., K$.

In practice, the constraint on Q relies on the priori knowledge of the items and the entire study. It is usually formulated to meet specific needs. For instance, if we want to define a factor (a skill or a mental disorder) by an item or multiple items, then these items are naturally included in the constraints. We would like to raise a warning for readers that inappropriate constraints on Q (equivalent, identifying wrong items for each trait) may lead to misleading or noninterpretable results. We recommend trying different constraints, checking if the results are consistent, and selecting the most sensible one.

On the correlation among the traits $\boldsymbol{\theta}$. The regularized estimator is introduced assuming that the covariance matrix of $\boldsymbol{\theta}$ is known. In case that Σ is unknown, we suggest two treatments. One is to consider Σ as an additional parameter in the specification of the log-likelihood function (4.6) and maximize it together with A. This approach typically induces additional computation. In the subsequent analysis, we consider a second approach that estimates Σ through an exploratory analysis (without regularization on the parameter A) under the constraints in Remark 5, with which Σ can be uniquely identified. Then, we rescale the variances in $\hat{\Sigma}$ to be unity, treat it as the true, and proceed to the regularized estimator (4.4).

4.4.2 Computation via Expectation-Maximization and Coordinate Descent Algorithm

In this section, we proceed to the computation of the estimators in (4.7) for a given sparsity parameter η. Notice that implementation in maximizing the regularized likelihood is not straightforward. We apply the expectation-maximization (EM) algorithm (Dempster et al., 1977) that is developed to compute the maximum likelihood estimator or posterior mode in presence of missing data. The EM algorithm is an iterative algorithm. Each iteration consists of two steps. The E-step computes the expected log-likelihood with respect to the posterior distribution of the missing data and the M-step maximizes the expected log-likelihood computed from the E-step. Adapted to our particular problem, the E-step is not in a closed form and we compute the expected log-likelihood via numerical approximation. The M-step is a convex optimization problem, and we use coordinate descent that is developed for the computation of L_1 regularized estimators for generalized linear models (Friedman et al., 2010). More detailed computation scheme is described as follows.

E-step. Let $(A^{(t)}, \mathbf{d}^{(t)})$ be the parameter values at the tth iteration. In order to evolve to the $(t+1)$th iteration, one first computes the expected complete data log-likelihood with respect to the posterior distribution

$$H(A, \mathbf{d}|A^{(t)}, \mathbf{d}^{(t)}) = E[l_{com}(A, \mathbf{d}; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)|A^{(t)}, \mathbf{d}^{(t)}],$$

where $l_{com}(A, \mathbf{d}; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$ is defined as in (4.5). The above expectation $E[\cdot|A^{(t)}, \mathbf{d}^{(t)}]$ is taken with respect to $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$ under the posterior distribution

$$p(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N|A^{(t)}, \mathbf{d}^{(t)}, \mathbf{y}_1, ..., \mathbf{y}_N) \propto \exp\{l_{com}(A^{(t)}, \mathbf{d}^{(t)}; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)\}. \tag{4.9}$$

The posterior expectation in the definition of the H-function is not in a closed form. We evaluate H numerically as follows. First, we write

$$H(A, \mathbf{d}|A^{(t)}, \mathbf{d}^{(t)}) = \sum_{j=1}^{J} H_j(\mathbf{a}_j, d_j|A^{(t)}, \mathbf{d}^{(t)}) + C,$$

where

$$\begin{aligned} &H_j(\mathbf{a}_j, d_j|A^{(t)}, \mathbf{d}^{(t)}) \\ =&\sum_{i=1}^{N} E\left[(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i) y_{ij} - \log(1 + e^{d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i}) \Big| A^{(t)}, \mathbf{d}^{(t)}\right] \end{aligned} \tag{4.10}$$

and $C = \sum_{i=1}^{N} E\left[\log f(\boldsymbol{\theta}_i)|A^{(t)}, \mathbf{d}^{(t)}\right]$ is a constant that does not depend on A and $\mathbf{d}$. As $H(A, \mathbf{d}|A^{(t)}, \mathbf{d}^{(t)})$ is maximized with respect to $A, \mathbf{d}$ in the M-step, we do not evaluate C. The $\boldsymbol{\theta}_i$'s are independent under the posterior distribution that is given by

$$p(\boldsymbol{\theta}_i|A^{(t)}, \mathbf{d}^{(t)}) \propto \prod_{j=1}^{J} \frac{\exp\{(d_j^{(t)} + (\mathbf{a}_j^{(t)})^\top \boldsymbol{\theta}_i) y_{ij}\}}{1 + \exp\{d_j^{(t)} + (\mathbf{a}_j^{(t)})^\top \boldsymbol{\theta}_i\}} f(\boldsymbol{\theta}_i).$$

We approximate the integration in (4.10) by a summation. More precisely, we consider grid points $\mathcal{G} \subseteq [-4, 4]^K$ and approximate the posterior distribution by

$$\hat{p}(\boldsymbol{\theta}_i|A^{(t)}, \mathbf{d}^{(t)}) \propto \begin{cases} \prod_{j=1}^{J} \frac{\exp\{(d_j^{(t)} + (\mathbf{a}_j^{(t)})^\top \boldsymbol{\theta}_i) y_{ij}\}}{1 + \exp\{d_j^{(t)} + (\mathbf{a}_j^{(t)})^\top \boldsymbol{\theta}_i\}} f(\boldsymbol{\theta}_i) & \text{if } \boldsymbol{\theta}_i \in \mathcal{G} \\ 0 & \text{otherwise,} \end{cases}$$

and $\sum_{\boldsymbol{\theta} \in \mathcal{G}} \hat{p}(\boldsymbol{\theta}|A^{(t)}, \mathbf{d}^{(t)}) = 1$. Thus, H_j is approximated by

$$\begin{aligned} &\hat{H}_j(\mathbf{a}_j, d_j|A^{(t)}, \mathbf{d}^{(t)}) \\ =&\sum_{i=1}^{N} \sum_{\boldsymbol{\theta}_i \in \mathcal{G}} \left[(d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i) y_{ij} - \log(1 + e^{d_j + \mathbf{a}_j^\top \boldsymbol{\theta}_i})\right] \hat{p}(\boldsymbol{\theta}_i|A^{(t)}, \mathbf{d}^{(t)}). \end{aligned}$$

Thus, the H-function is approximated by

$$\hat{H}(A, \mathbf{d}|A^{(t)}, \mathbf{d}^{(t)}) = \sum_{j=1}^{J} \hat{H}_j(\mathbf{a}_j, d_j|A^{(t)}, \mathbf{d}^{(t)}).$$

We choose $\mathcal{G}$ to be $\mathcal{S} \times \cdots \times \mathcal{S}$, where $\mathcal{S}$ is the set of $M = 21$ (for $K = 3$) and 11 (for $K = 4$) grid points on the interval $[-4, 4]$.

M-step. With the H-function computed in the E-step, we further perform the M-step, that is,

$$(A^{(t+1)}, \mathbf{d}^{(t+1)}) = \arg\max_{A,\mathbf{d}} \{\hat{H}(A, \mathbf{d}|A^{(t)}, \mathbf{d}^{(t)}) - N\eta\|A\|_1\}. \tag{4.11}$$

Notice that the function $\hat{H}$ factorizes to the sum of $\hat{H}_j$'s. Each $\hat{H}_j$ is a function only of $\mathbf{a}_j$ and d_j. Then, the above maximization can be reduce to maximizing each $\hat{H}_j$ separately, that is,

$$(\mathbf{a}_j^{(t+1)}, d_j^{(t+1)}) = \arg\max_{\mathbf{a}_j, d_j} \{\hat{H}_j(\mathbf{a}_j, d_j|\mathbf{a}_j^{(t)}, d_j^{(t)}) - N\eta\|\mathbf{a}_j\|_1\}. \tag{4.12}$$

The above maximization is of a much lower dimension than that of (4.11). It is straightforward to verify that $A^{(t+1)} = (\mathbf{a}_j^{(t+1)} : 1 \le j \le J)$ and $\mathbf{d}^{(t+1)} = (d_j^{(t+1)} : 1 \le j \le J)$. We point out that the optimization in (4.12) is a convex optimization problem and we solve it by using the coordinate descent algorithm developed by Friedman et al. (2010). The EM algorithm evolves according to (4.11) until convergence.

4.4.3 Simulation Study

In this section, we perform simulations to illustrate the performance of the proposed method under various settings. As the main objective of the study is the Q-matrix, we mainly consider the correct estimation rate of the Q-matrix that is defined as

$$CR = \frac{1}{K(J-K)} \sum_{K+1 \le j < J, 1 \le k \le K} I(\hat{q}_{jk} = q_{jk}^*) \tag{4.13}$$

where $\hat{Q} = (\hat{q}_{jk})$ is an estimate and $Q^* = (q_{jk}^*)$ is the true matrix. In what follows, we investigate the correct estimation rates of the L_1 regularized estimator with the regularization parameter η chosen according to the Bayesian information criterion under various model settings. For the estimate of the A-matrix, we consider the mean squared error for each entry.

In this study, we consider the M2PL model for $K = 3$ and 4 respectively. For $K = 3$, we consider two different A-matrices given as in Tables 4.15 and 4.16, denoted by A_1 and A_2. We choose the A-matrices so that they contain some single-, double-, and triple-attribute items. The difference between these two matrices is that the coefficients in A_1 are larger and there are more single-trait items. Thus, A_1 is considered as easier to estimate. For

$K = 4$, the matrix A_3 is given in Table 4.17. Furthermore, the latent traits $\boldsymbol{\theta}$ have variance one and a common correlation $\rho = 0.1$ and Σ is considered as unknown. For each A-matrix, we generate 50 independent data sets of sample size $N = 2000$ to evaluate the Frequentist properties of our estimator.

Table 4.15: Loading matrix A_1 used in the simulation study.

	Item									
Latent trait	1	2	3	4	5	6	7	8	9	10
1	1.9	1.7	1.5	1.3	1.1	0.9	0.7	0.5	0.3	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Item									
Latent trait	11	12	13	14	15	16	17	18	19	20
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	1.7	1.5	1.3	1.1	0.9	0.7	0.5	0.3	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.9	1.7
	Item									
Latent trait	21	22	23	24	25	26	27	28	29	30
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.7
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	1.3	1.5
3	1.5	1.3	1.1	0.9	0.7	0.5	0.3	0.0	0.0	0.0
	Item									
Latent trait	31	32	33	34	35	36	37	38	39	40
1	0.9	1.1	1.3	0.0	0.0	0.0	0.3	0.5	0.7	0.9
2	0.0	0.0	0.0	0.5	0.7	0.9	1.1	1.3	1.5	1.7
3	0.7	0.9	1.1	1.3	1.5	1.7	1.9	1.9	1.1	0.5

The parameters are estimated via the algorithm described as in Section 4.4.2 with the sparsity parameter η chosen according to BIC. To ensure identifiability, we consider the following two sets of constraints on the parameters.

1 We designate one item for each factor and this item is associated with only that factor.

That is, we set sub-Q-matrix corresponding to the K items to be identity. This is the first constraint specified in Remark 5.

2 We designate one item for each factor. This item is associated with that factor for sure and may also associated with others. That is, we set the diagonal elements of the sub-Q-matrix corresponding to the K items to be ones and off diagonal elements have no constraint. Technically, the L_1 penalty includes all coefficients except for $(a_{1,1}, a_{10,2}, a_{19,3})$ in the case of A_1. Notice that this constraint is much weaker than the first one, nevertheless still ensures identifiability as long as it is correctly specified (due to the regularization on other coefficients).

We treat the covariance Σ as unknown and estimate it via a constrained exploratory analysis as mentioned previously. To illustrate the performance, we investigate the correct estimation rates in (4.13) from different aspects. First, Figure 4.5 shows the histograms of the correct estimation rates over the 50 independent data sets for A_1 under constraints 1 and 2. The overall rates are well over 95%. We also consider the mean squared error for each a_{ij} and there are $40 \times 3 = 120$ MSE's in total whose histograms are also shown in Figure 4.5.

As we mentioned, the regularization parameter is chosen to minimize the BIC value, denoted by η_*. Its correct estimation rate is denoted by MR_*. As the true Q-matrix is known, we can further choose η to maximize the correct estimation rates that is denoted by MR_0. The first plot in Figure 4.6 shows the scatter plot of the pair (MR_*, MR_0) for all 50 data sets. BIC is a reasonable criterion to select η in terms of maximizing the correct estimation rate.

Furthermore, we investigate a data set that is randomly chosen from the 50 simulated data sets to illustrate the performance of BIC in selecting the regularization parameters. We standardize the BIC values as a function of η by some linear transformations such that it sits well in the same plot as the mis-estimation rate that is the compliment of the correct estimation rate. The second plot in Figure 4.6 shows BIC and the mis-estimation rate as a function of η in the same plot. The BIC and mis-estimation curves both decrease first and then increase. The decreasing slope of the BIC curve is induced by the $\log N$ penalty. The

minima of both curves coincide suggesting that BIC is a good criterion for selecting η.

The correct estimation rates of Q and MSE of A_2 are given as in Figure 4.3. The correct estimation rates is lower (still mostly over 90%) because the magnitude of the coefficients are smaller. The corresponding results for the 4-dimensional case A_3 is given in Figure 4.4. The results are similar.

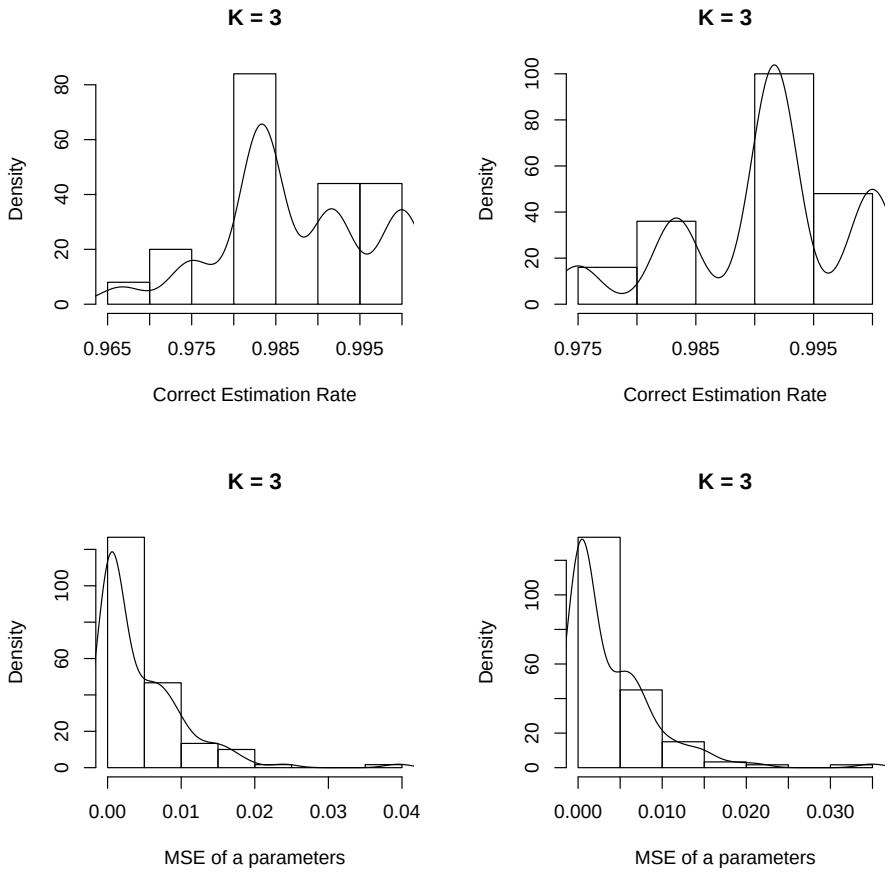

Figure 4.1: Histograms of the correct rates for Q (row 1) and MSE of the estimate of the a parameters for A_1 (row 2) under constraint 1 (left column) and constraint 2 (right column).

4.4.4 Real Data Analysis: EPQ-R Data

The study analyzes 824 females' responses to the revised Eysenck Personality Questionnaire short scales. There are in total 36 items. Based on Eysenck's theory on personality, there

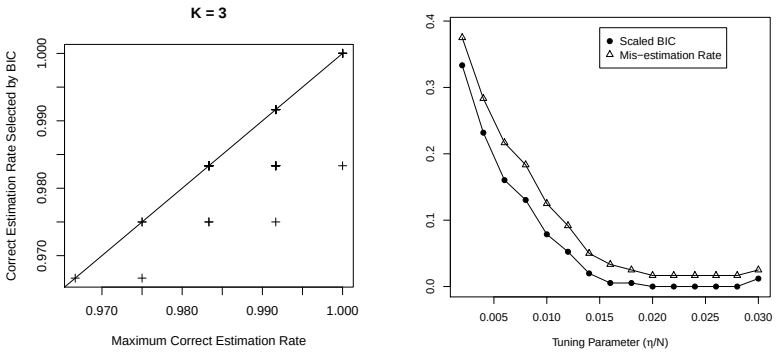

Figure 4.2: Left: comparing the correct estimation rates selected by BIC and the optimal rates. Right: mis-estimation rates and BIC against η.

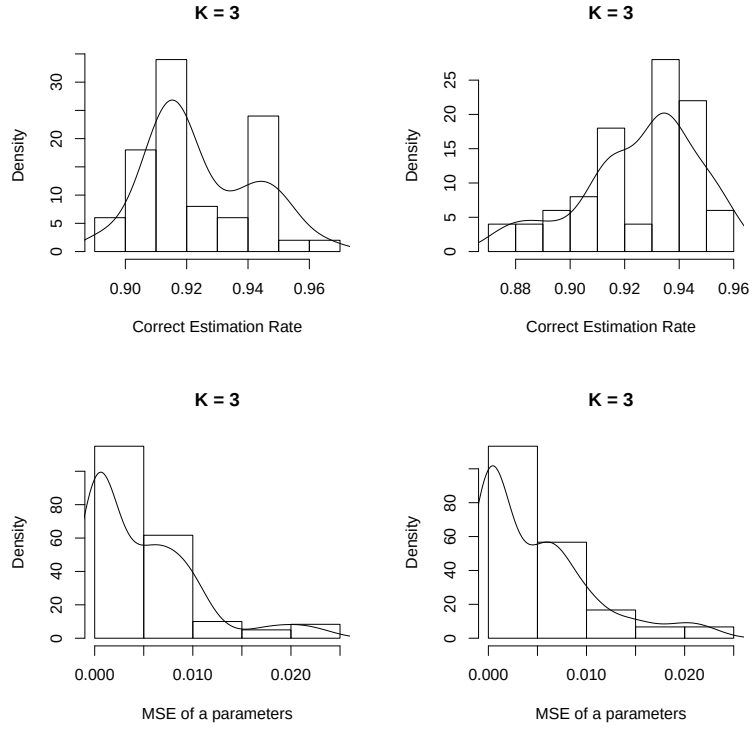

Figure 4.3: Histograms of the correct rates for Q (row 1) and MSE of the estimate of the a parameters for A_2 (row 2) under constraint 1 (left column) and constraint 2 (right column).

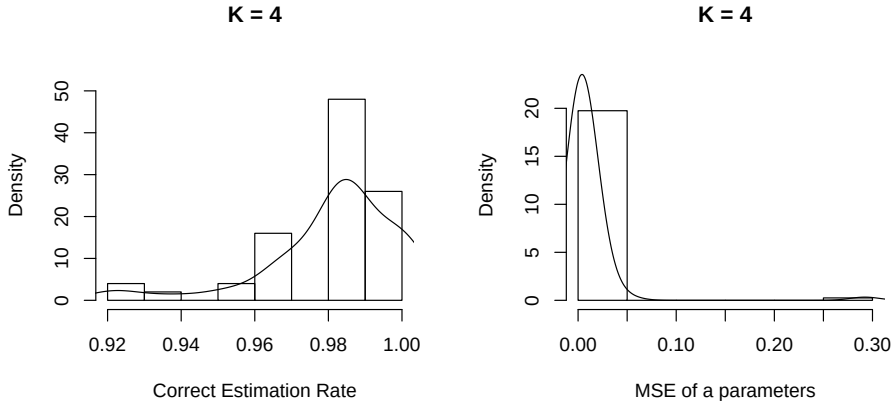

Figure 4.4: Histograms of the correct rates for Q (row 1) and MSE of the estimate of the a parameters for A_3 (row 2) under constraint 1 (left column) and constraint 2 (right column).

are three factors: Psychoticism (P), Extraversion (E), and Neuroticism (N) scales. In the pre-specified Q-matrix of the confirmatory analysis, each item is associated with only one factor. In particular, items 1-12 are associated with "Psychoticism"; items 13 - 24 to "Extraversion"; items 25 - 36 to "Neuroticism." The specific questions can be found in Appendix 2 of Eysenck et al. (1985) and we reorder the items according to Table 9 of Eysenck et al. (1985). Furthermore, the data set has been preprocessed so that the negatively worded items have already been reversely scored. Thus, "yes" to a question is coded as "0" if the question has been reversed.

In the analysis, we impose two sets of different constraints on Q to ensure identifiability. They eventually lead to similar results.

1. We designate two items for each factor and these two items are associated with only that factor. In particular, for "Psychoticism", we select items 1 and 2 and set rows 1 and 2 of Q to be $(1, 0, 0)$; for "Extraversion", we set rows 13 and 14 of Q to be $(0, 1, 0)$; for "Neuroticism", we set rows 25 and 26 of Q to be $(0, 0, 1)$.

2. We designate two items for each factor. These two items are associated with that factor for sure but may also associated with others. More specifically, for "Psychoticism", we select items 1 and 2 and set rows 1 and 2 of Q to be $(1, ?, ?)$; for "Extraversion", we set rows 13 and 14 of Q to be $(?, 1, ?)$; for "Neuroticism", we set rows 25 and 26

of Q to be $(?, ?, 1)$. The question mark "?" means that this entry is to be estimated. Technically, we do not penalize the coefficients $(a_{1,1}, a_{2,1}, a_{12,2}, a_{13,2}, a_{25,3}, a_{26,3})$ and penalize all other a_{ij}'s.

We have also experimented with constraints on other items and the results are similar and we only report the results of the above selection. For the covariance matrix of $\boldsymbol{\theta}$, we estimate it by fitting a confirmatory model stated at the beginning of this section and treat it as known. The rescaled estimate (to variance one) is

$$\hat{\Sigma} = \begin{pmatrix} 1.00 & 0.11 & -0.03 \\ 0.11 & 1.00 & -0.25 \\ -0.03 & -0.25 & 1.00 \end{pmatrix}.$$

For each set of constraints, we compute BIC for the regularization parameter for $\eta \in [0.00, 0.04]$. The plots of the BIC values against η are showed in Figure 4.7. For constraint 1, the BIC selects $\lambda = 0.030$ and the coefficients are showed in Table 4.18. For constraint 2, the BIC selects $\lambda = 0.032$ and the estimated coefficients are showed in Table 4.19.

The nonzero patterns of the a-coefficients in both tables are very similar and so it is for their estimated values. In fact, Constraint 1 is inconsistent with Table 4.19 on items 1, 13, and 25 that are forced to be single-trait items. But, the results on other unconstrained items are similar. This also illustrates the robustness of the current method. According to the A-matrix, most items remain associated with a single trait. There are some associated with more than one traits. We examined those items and found that most of them are very sensible. For instance, item 6 "Do you take much notice of what people think?" (a reverse question) is also related to "Neuroticism" that is characterized by anxiety, fear, worry, etc. and its wording is similar to those of items 32 "Are you a worrier?" and 34 "Do you worry too long after an embarrassing experience?"; for item 9 "Do you enjoy co-operating with others?" (a reverse question, originally designed for "Psychoticism"), there is a good reason to believe that it is associated with "Extraversion"; item 27 "Are you an irritable person?" is associated with both "Psychoticism" (characterized by aggressiveness and interpersonal hostility) and "Neuroticism".

We further compare the above results to a classical method as follows. We first fit

an exploratory model on the data under the constraint 1 without the L_1 penalty. The estimated coefficients are given in Table 4.20. We then set a_{ij} to zero if $|a_{ij}| < \varepsilon_0$. Table 4.21 shows the Q-matrix for $\varepsilon_0 = 0.5$ that yields the closest results to ours. We see that its basic pattern is similar to that of Table 4.19, but the L_1 regularized estimator does keep some low magnitude coefficients nonzero, such as items 7, 25, 30, and 34.

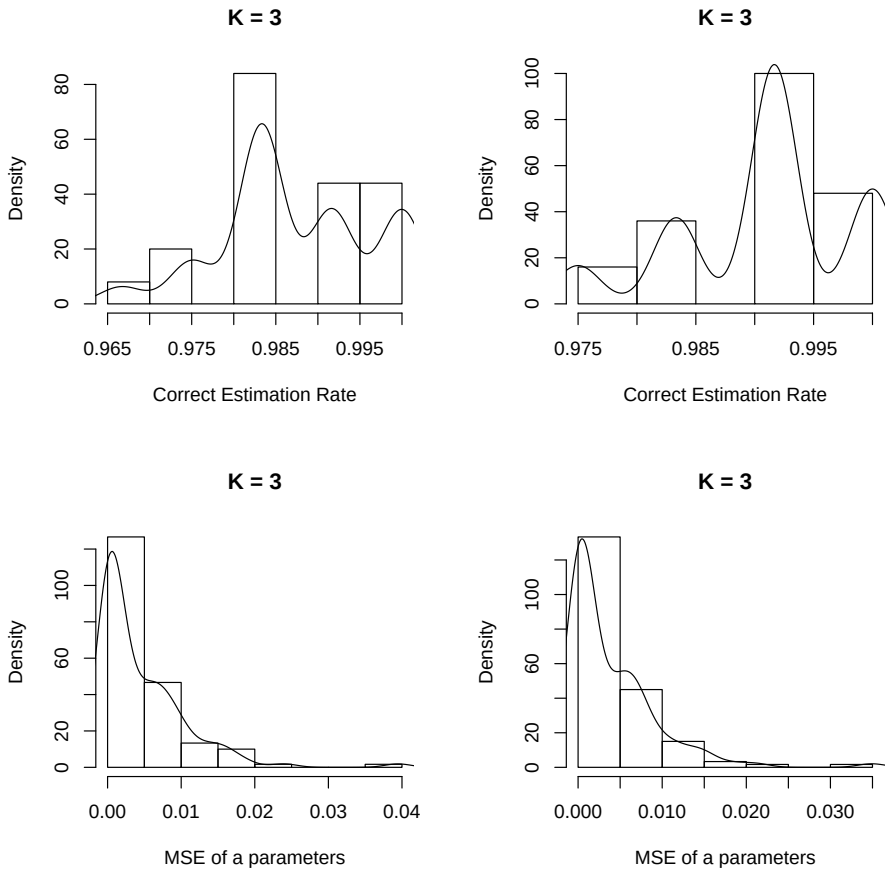

Figure 4.5: Histograms of the correct rates for Q (row 1) and MSE of the estimate of the a parameters for A_1 (row 2) under constraint 1 (left column) and constraint 2 (right column).

4.4.5 Extension to MNRM Model

In the above analysis, we focus on the M2PL model, which is one of the most popular multidimensional IRT models for binary response data. In fact, this method can be generalized

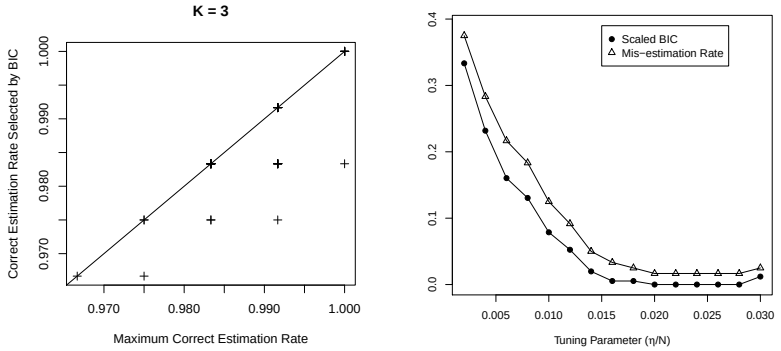

Figure 4.6: Left: comparing the correct estimation rates selected by BIC and the optimal rates. Right: mis-estimation rates and BIC against η.

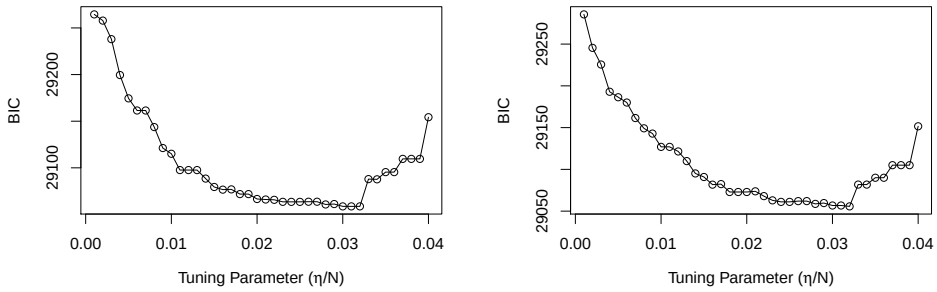

Figure 4.7: The BIC values on the solution path for the EPQ-R data for constraint 1 (left) and constraint 2(right).

to other models, such as the MGPCM model for polytomous responses and the MNRM model for categorical responses. In what follows, we briefly discuss its extension to the Q-matrix estimation under the MNRM model. For the MNRM model (2.4),

$$\log\left(\frac{P(Y_j = y|\boldsymbol{\theta})}{P(Y_j = 0|\boldsymbol{\theta})}\right) = (\mathbf{a}_j^y)^\top \boldsymbol{\theta} + d_j^y,$$

where $\mathbf{a}_j^y = (a_{j1}^y, ..., a_{jK}^y)^\top$ is a K-dimensional vector containing the slope parameters of item j and the response category y of this item. Similar to the M2PL, there is a relationship between the Q-matrix and the zero patterns of the factor loadings. Specifically,

$$\|\mathbf{a}_{jk}\| = 0 \text{ if and only if } q_{jk} = 0,$$

where $\mathbf{a}_{jk} = (a_{jk}^1, ...a_{jk}^{c_j})^\top$ is the vector of slope parameters corresponding to jth item and kth dimension of the latent vector. It slightly differs from the M2PL model that $q_{jk} = 0$ no longer corresponds to a single entry of A being zero, but a vector being a zero vector. In other words, the entries of A naturally fall into groups ($\mathbf{a}_{jk}$s) and the Q-matrix being sparse implies that the matrix A is group-wise sparse. To impose the group-wise sparse structure, the group Lasso regularization tends to perform better than the L_1 regularization, because the L_1 regularization tends to make decisions based on the strength of individual loadings, not a group of loadings. The group Lasso regularization on A is

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\sqrt{c_j}\|\mathbf{a}_{jk}\|,$$

where the $\sqrt{c_j}$ terms account for the varying group sizes and the grouping effect comes from treating $\mathbf{a}_{jk}$ as a group in $\|\mathbf{a}_{jk}\|$.

Let the log-likelihood of the observe responses be $l(A, \mathbf{d})$, where $A = (a_{jk}^y : y = 1, ..., c_j, j = 1, ..., J, k = 1, ..., K)$ and $\mathbf{d} = (d_j^y : y = 1, ..., c_j, j = 1, ..., J)$. The group Lasso regularized estimator is obtained as follows:

$$(\hat{A}^\eta, \hat{\mathbf{d}}^\eta) = \arg\max_{A,\mathbf{d}} \Big\{ l(A, \mathbf{d}) - N\eta \sum_{j=1}^{J}\sum_{k=1}^{K}\sqrt{c_j}\|\mathbf{a}_{jk}\| \Big\},$$

where η is the tuning parameter that controls the group-wise sparsity. The computation is essentially the same as described in the subsection 4.4.2, except that in the M-step the coordinate decent will be replaced by block-wise coordinate descent algorithm. We refer to Yuan and Lin (2006) for more details of group Lasso regularization.

Table 4.16: Loading matrix A_2 used in the simulation study.

	Item									
Latent trait	1	2	3	4	5	6	7	8	9	10
1	1.2	1.0	0.8	0.6	0.4	0.2	0.1	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	1.0	0.8
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Item									
Latent trait	11	12	13	14	15	16	17	18	19	20
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.6	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	1.2	1.0	0.8	0.6	0.4	0.2
	Item									
Latent trait	21	22	23	24	25	26	27	28	29	30
1	0.0	0.8	0.6	0.4	0.2	0.8	0.6	0.4	0.2	0.8
2	0.0	0.2	0.8	0.6	0.4	0.2	0.0	0.0	0.0	0.0
3	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.8	0.6	0.4
	Item									
Latent trait	31	32	33	34	35	36	37	38	39	40
1	0.6	0.0	0.0	0.0	0.0	0.0	0.8	0.6	0.4	0.2
2	0.0	0.6	0.4	0.2	0.8	0.6	0.2	0.4	0.6	0.8
3	0.2	0.8	0.6	0.4	0.2	0.1	0.8	0.6	0.4	0.2

Table 4.17: Loading matrix A_3 used in the simulation study.

	Item									
Latent trait	1	2	3	4	5	6	7	8	9	10
1	1.5	1	0.5	0.0	0	0.0	0.0	0	0.0	0.0
2	0.0	0	0.0	1.5	1	0.5	0.0	0	0.0	0.0
3	0.0	0	0.0	0.0	0	0.0	1.5	1	0.5	0.0
4	0.0	0	0.0	0.0	0	0.0	0.0	0	0.0	1.5
	Item									
Latent trait	11	12	13	14	15	16	17	18	19	20
1	0	0.0	0.5	0.5	0.5	0.0	0	0.0	0.5	0.0
2	0	0.0	1.0	0.0	0.0	1.0	1	0.0	1.0	1.5
3	0	0.0	0.0	1.5	0.0	1.5	0	1.5	1.5	1.0
4	1	0.5	0.0	0.0	0.5	0.0	1	1.5	0.0	0.5

Table 4.18: The model selected by BIC for constraint 1.

	A			b		A			b		A			b
1	1.50	0.00	0.00	-2.58	13	0.00	1.89	0.00	1.14	25	0.00	0.00	1.50	1.10
2	1.44	0.00	0.00	0.55	14	0.00	2.55	0.00	2.01	26	0.00	0.00	1.10	0.70
3	1.23	0.00	0.00	-2.42	15	0.00	1.57	0.00	1.53	27	0.39	0.00	1.29	-0.83
4	0.74	0.00	0.00	-0.89	16	0.00	1.71	0.00	3.10	28	0.00	0.00	1.27	1.24
5	1.09	0.00	0.00	-3.07	17	0.00	1.40	0.00	0.61	29	0.00	0.00	1.53	0.08
6	0.87	0.00	-0.60	-1.23	18	0.00	2.56	0.00	-1.43	30	0.00	0.32	1.28	-0.04
7	0.96	0.27	-0.37	-2.49	19	0.00	1.76	0.00	3.20	31	0.00	-0.27	2.03	-1.12
8	1.13	0.00	0.00	-2.26	20	0.00	2.19	0.00	-0.60	32	-0.66	0.00	2.27	0.94
9	1.20	-0.63	0.00	-2.89	21	0.00	1.23	0.00	1.18	33	0.00	0.00	1.84	-1.46
10	0.87	0.00	0.00	-1.96	22	0.00	2.34	0.00	0.92	34	-0.66	-0.21	1.44	0.69
11	1.45	0.00	0.00	-3.09	23	0.53	2.62	0.00	0.68	35	0.00	0.00	2.05	-1.15
12	1.28	0.00	0.00	-0.16	24	0.00	2.21	0.00	1.29	36	0.00	0.00	1.19	-0.99

Table 4.19: The model selected by BIC for constraint 2.

	A			b		A			b		A			b
1	1.54	0.00	0.51	-2.68	13	0.00	2.04	0.37	1.17	25	0.00	0.28	1.63	1.13
2	1.44	0.00	0.00	0.56	14	0.00	2.54	0.00	2.01	26	0.00	0.00	1.13	0.71
3	1.21	0.00	0.00	-2.41	15	0.00	1.56	0.00	1.53	27	0.39	0.00	1.32	-0.84
4	0.74	0.00	0.00	-0.89	16	0.00	1.71	0.00	3.11	28	0.00	0.00	1.27	1.24
5	1.06	0.00	0.00	-3.05	17	0.00	1.40	0.00	0.61	29	0.00	0.00	1.54	0.08
6	0.87	0.00	-0.58	-1.23	18	0.00	2.51	0.00	-1.42	30	0.00	0.32	1.27	-0.04
7	0.96	0.28	-0.32	-2.49	19	0.00	1.75	0.00	3.20	31	0.00	-0.26	1.98	-1.11
8	1.13	0.00	0.00	-2.26	20	0.00	2.16	0.00	-0.60	32	-0.66	0.00	2.22	0.93
9	1.20	-0.64	0.00	-2.88	21	0.00	1.23	0.00	1.19	33	0.00	0.00	1.84	-1.46
10	0.88	0.00	0.00	-1.96	22	0.00	2.32	0.00	0.92	34	-0.66	-0.20	1.42	0.69
11	1.45	0.00	0.00	-3.08	23	0.54	2.62	0.00	0.69	35	0.00	0.00	2.01	-1.13
12	1.27	0.00	0.00	-0.16	24	0.00	2.21	0.00	1.30	36	0.00	0.00	1.19	-1.00

Table 4.20: The estimated A-matrix of the exploratory analysis.

	A				A				A		
1	1.61	0.00	0.00	13	0.00	2.00	0.00	25	0.00	0.00	1.64
2	1.34	0.00	0.00	14	0.00	2.57	0.00	26	0.00	0.00	1.14
3	1.29	-0.37	0.04	15	0.32	1.53	-0.16	27	0.21	-0.19	1.30
4	0.74	0.04	-0.09	16	0.08	1.70	-0.47	28	-0.45	-0.12	1.37
5	1.15	-0.14	0.28	17	0.14	1.38	-0.14	29	-0.13	-0.15	1.58
6	1.10	-0.24	-0.81	18	0.72	2.50	-0.32	30	-0.38	0.23	1.34
7	1.09	0.25	-0.53	19	-0.34	1.88	-0.14	31	-0.54	-0.47	2.12
8	1.27	-0.45	-0.27	20	0.48	2.11	-0.27	32	-1.12	-0.22	2.41
9	1.15	-0.68	0.11	21	0.02	1.24	-0.16	33	-0.04	-0.30	1.86
10	1.01	0.10	-0.36	22	0.23	2.36	-0.07	34	-1.00	-0.30	1.62
11	1.46	0.15	0.04	23	0.83	2.53	-0.38	35	-0.35	-0.28	2.07
12	1.22	-0.06	-0.19	24	0.20	2.16	-0.23	36	-0.02	-0.25	1.21

Table 4.21: The Q-matrix from a hard-thresholding method with threshold 0.5.

	Q				Q				Q		
1	1	0	0	13	0	1	0	25	0	0	1
2	1	0	0	14	0	1	0	26	0	0	1
3	1	0	0	15	0	1	0	27	0	0	1
4	1	0	0	16	0	1	0	28	0	0	1
5	1	0	0	17	0	1	0	29	0	0	1
6	1	0	1	18	1	1	0	30	0	0	1
7	1	0	1	19	0	1	0	31	1	0	1
8	1	0	0	20	0	1	0	32	1	0	1
9	1	1	0	21	0	1	0	33	0	0	1
10	1	0	0	22	0	1	0	34	1	0	1
11	1	0	0	23	1	1	0	35	0	0	1
12	1	0	0	24	0	1	0	36	0	0	1

4.5 Appendix of Chapter 4

4.5.1 Appendix A: Some Technical Constructions

Proposition 1 in Section 2.4 suggests that the DINA and the DINO model are mathematically the same but with different parameterizations. Therefore, all the theoretical results we developed for the DINA model can be directly translated to the DINO model based on Proposition 1. Therefore, the rest of the technical proofs are all for the DINA model. In the rest of this subsection, we present some technical construction for the subsequent proof.

T-matrix for the DINA model. For notational convenience, we will write

$$c = 1 - s$$

that is the correct response probability for capable students ("c" for correct). Then,

$$\mathbf{c} = \mathbf{1} - \mathbf{s}$$

is the corresponding parameter vector.

The T-matrix serves as a connection between the observed response distribution and the model structure. We first specify each row vector of the T-matrix for a general conjunctive diagnostic model.

For each item j, we have

$$P(Y_j = 1|Q, \mathbf{p}, \mathbf{c}, \mathbf{g}) = \sum_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}} c_{j,\boldsymbol{\theta}} = \sum_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}} P(Y_j = 1|Q, \boldsymbol{\theta}, \mathbf{c}, \mathbf{g}), \tag{4.14}$$

We create a row vector $B_{\mathbf{c},\mathbf{g},Q}(j)$ of length 2^K containing the probabilities $c_{j,\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}$'s and arrange those elements in an appropriate order, then we write (4.14) in the form of a matrix product

$$\sum_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}} c_{j,\boldsymbol{\theta}} = B_{\mathbf{c},\mathbf{g},Q}(j)\mathbf{p},$$

where $\mathbf{p}$ is the column vector containing the probabilities $p_{\boldsymbol{\theta}}$. For each pair of items, we may establish that the probability of responding positively to both items j_1 and j_2 is

$$P(Y_{j_1} = 1, Y_{j_2} = 1|Q, \mathbf{p}, \mathbf{c}, \mathbf{g}) = \sum_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}} c_{j_1,\boldsymbol{\theta}} c_{j_2,\boldsymbol{\theta}} = B_{\mathbf{c},\mathbf{g},Q,}(j_1, j_2)\mathbf{p}.$$

where $B_{\mathbf{c},\mathbf{g},Q}(j_1, j_2)$ is defined as a row vector containing the probabilities $c_{j_1,\boldsymbol{\theta}} c_{j_2,\boldsymbol{\theta}}$ for each $\boldsymbol{\theta}$. Note that each element of $B_{\mathbf{c},\mathbf{g},Q}(j_1, j_2)$ is the product of the corresponding elements of $B_{\mathbf{c},\mathbf{g},Q}(j_1)$ and $B_{\mathbf{c},\mathbf{g},Q}(j_2)$. With a completely analogous construction, for items $j_1, \cdots, j_l$, we can write the probability of responding positively to all items as

$$P(Y_{j_1} = 1, \ldots, Y_{j_l} = 1 | Q, \mathbf{p}, \mathbf{c}, \mathbf{g}) = B_{\mathbf{c},\mathbf{g},Q}(j_1, \ldots, j_l)\mathbf{p},$$

Note that $B_{\mathbf{c},\mathbf{g},Q}(j_1, \ldots, j_l)$ is the element-by-element product of $B_{\mathbf{c},\mathbf{g},Q}(j_1)$,...,$B_{\mathbf{c},\mathbf{g},Q}(j_l)$.

The T-matrix for the DINA model has 2^K columns and 2^J rows. Each of the first $2^J - 1$ row vectors of the T-matrix is one of the vectors $B_{\mathbf{c},\mathbf{g},Q}(j_1, ..., j_l)$. The last row of the T-matrix is taken as $\mathbf{1}^\top$. The T-matrix can be written as

$$T_{\mathbf{c},\mathbf{g}}(Q) = \begin{pmatrix} B_{\mathbf{c},\mathbf{g},Q}(1) \\ \vdots \\ B_{\mathbf{c},\mathbf{g},Q}(J) \\ B_{\mathbf{c},\mathbf{g},Q}(1,2) \\ \vdots \\ B_{\mathbf{c},\mathbf{g},Q}(1, ..., J) \\ \mathbf{1}^\top \end{pmatrix}. \tag{4.15}$$

Response $\boldsymbol{\gamma}$-vector. We further define $\boldsymbol{\gamma}$ to be the vector containing the probabilities of the empirical distribution corresponding to those in $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$, e.g., the first element of $\boldsymbol{\gamma}$ is $\frac{1}{N}\sum_{i=1}^{N} I(Y_{i1} = 1)$ and the $(J+1)$-th element is $\frac{1}{N}\sum_{i=1}^{N} I(Y_{i1} = 1 \; and \; Y_{i2} = 1)$, i.e.,

$$\boldsymbol{\gamma} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} I(Y_{i1} = 1) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} I(Y_{iJ} = 1) \\ \frac{1}{N}\sum_{i=1}^{N} I(Y_{i1} = 1 \; and \; Y_{i2} = 1) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} I(Y_{i1} = 1, \; Y_{i2} = 1, \cdots, \; and \; Y_{iJ} = 1) \\ 1 \end{pmatrix}. \tag{4.16}$$

An objective function. Under the true Q-matrix Q, let $(\mathbf{c}, \mathbf{g}, \mathbf{p})$ be the true model parameters. By the the law of large number, we have that

$$\boldsymbol{\gamma} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} I(Y_{i1}=1) \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} I(Y_{iJ}=1) \\ \frac{1}{N}\sum_{i=1}^{N} I(Y_{i1}=1 \; and \; Y_{i2}=1) \\ \vdots \end{pmatrix}$$

$$\to \begin{pmatrix} P(Y_{i1}=1|Q,\mathbf{c},\mathbf{g},\mathbf{p}) \\ \vdots \\ P(Y_{iJ}=1|Q,\mathbf{c},\mathbf{g},\mathbf{p}) \\ P(Y_{i1}=1 \; and \; Y_{i2}=1|Q,\mathbf{c},\mathbf{g},\mathbf{p}) \\ \vdots \end{pmatrix} = T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$$

almost surely as $N \to \infty$. For each Q, we define

$$S(Q) = \inf_{\mathbf{c},\mathbf{g},\mathbf{p}} |T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p} - \boldsymbol{\gamma}|^2, \tag{4.17}$$

where the minimization is subject to the natural constraints that $c_j, g_j, p_{\boldsymbol{\theta}} \in (0,1)$ and $\sum_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}} = 1$. Here $|\cdot|$ means the Euclidian norm. Thanks to the law of large numbers, $S(Q) \to 0$ as $N \to \infty$. The estimator

$$\tilde{Q} = \text{argmin}_Q S(Q)$$

is consistent meaning that

$$P(\tilde{Q} \sim Q) \to 1$$

if and only if the vector $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p} \neq T_{\mathbf{c}',\mathbf{g}'}(Q')\mathbf{p}'$ for $Q' \nsim Q$ and all possible $\mathbf{c}'$, $\mathbf{g}'$ and $\mathbf{p}'$.

4.5.2 Appendix B: Proof of Theorems

The following proposition provides a connection between the likelihood function and the T-matrix, which makes it possible to use the T-matrix to show the model identifiability.

Proposition 2. *Under the DINA and DINO models, for two sets of parameters* $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ *and* $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$*,*

$$L(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}, Q) = L(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}}, Q)$$

for all $\mathbf{Y}$ *if and only if the following equation holds:*

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}. \tag{4.18}$$

The following proposition provides a relationship between T-matrices of different model parameters.

Proposition 3. *There exists an invertible matrix* $D_{\mathbf{g}^*}$ *depending only on* $\mathbf{g}^* = (g_1^*, ..., g_J^*)$*, such that*

$$D_{\mathbf{g}^*} T_{\mathbf{c},\mathbf{g}}(Q) = T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q).$$

Thus, (4.18) is equivalent to $T_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*}(Q)\hat{\mathbf{p}}$ for some $\mathbf{g}^*$. This is a very important technique that will be used repeatedly in the subsequent development. We now cite a proposition.

Proposition 4 (Proposition 6.6 in Liu et al. (2013))**.** *For the DINA model, with* $(Q, \mathbf{c}, \mathbf{g}, \mathbf{p})$ *under Condition A1-3, when* $Q' \nsim Q$*,* $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$ *is not in the column space of* $T_{\mathbf{c}',\mathbf{g}}(Q')$ *for all* $\mathbf{c}'$*, that is,* $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p} \neq T_{\mathbf{c}',\mathbf{g}}(Q')\mathbf{p}'$ *for all* $\mathbf{c}'$ *and* $\mathbf{p}'$*. In addition,* $T_{\mathbf{c},\mathbf{g}}(Q)$ *is of full column rank.*

The following proposition provides the first step result.

Proposition 5. *Under the DINA and DINO models, with* Q*,* $\mathbf{s}$*, and* $\mathbf{g}$ *being known, the population proportion parameter* $\mathbf{p}$ *is identifiable if and only if* Q *is complete.*

Proof of Proposition 5. When Q is complete, the matrix $T_{\mathbf{c},\mathbf{g}}(Q)$ has full column rank from Proposition 4. Thus, $\mathbf{p}$ is identifiable by Proposition 2.

Consider the case where the Q is incomplete. Without loss of generality, we assume $\mathbf{e}_1 = (1, 0, \cdots, 0)$ is not in the set of row vectors of Q. Then in the T-matrix $T_{\mathbf{c},\mathbf{g}}(Q)$, the columns corresponding to attribute profiles $\mathbf{0}$ and $\mathbf{e}_1$ are the same. Therefore, by Proposition 2, we can always find two different set of estimates of $\mathbf{p_0}$ and $\mathbf{p_{e_1}}$ such that equation (4.18) holds and therefore $\mathbf{p} = (p_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \{0, 1\}^K)$ is nonidentifiable. □

Proof of Theorem 4. The identifiability of the Q-matrix for the DINO model is an application of Theorem 3 and Proposition 1. In what follows, we focus on the identifiability of the model parameters $\mathbf{c}$ and $\mathbf{p}$ under the DINA model.

We only need to show that when $\mathbf{g}$ is known, for two sets of parameters $(\hat{\mathbf{c}}, \mathbf{g}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \mathbf{g}, \bar{\mathbf{p}})$, $L(\hat{\mathbf{c}}, \mathbf{g}, \hat{\mathbf{p}}, Q) = L(\bar{\mathbf{c}}, \mathbf{g}, \bar{\mathbf{p}}, Q)$ holds if and only if A4 satisfied. By Propositions 2 and 3, two sets of parameters $(\hat{\mathbf{c}}, \mathbf{g}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \mathbf{g}, \bar{\mathbf{p}})$ yield identical likelihood if and only if

$$T_{\hat{\mathbf{c}}-\mathbf{g},\mathbf{0}}(Q)\hat{\mathbf{p}} = D_{\mathbf{g}} T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}} = D_{\mathbf{g}} T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}} = T_{\bar{\mathbf{c}}-\mathbf{g},\mathbf{0}}(Q)\bar{\mathbf{p}}. \tag{4.19}$$

Thus under the assumption that $c_j > g_j$, we only need to consider that $\mathbf{g} = \mathbf{0}$.

Sufficiency of A4. For notational convenience, we write $B_Q(j_1, ..., j_l) = B_{\mathbf{c},\mathbf{g},Q}(j_1, ..., j_l)$ when $\mathbf{c} = \mathbf{1}$ and $\mathbf{g} = \mathbf{0}$. For each item $j \in 1, \cdots, J$, condition A4 implies that there exist items $j_1, ..., j_l$ (different from j) such that

$$B_Q(j, j_1, ..., j_l) = B_Q(j_1, ..., j_l),$$

that is, the attributes required by item j are a subset of the attributes required by items $j_1, ..., j_l$.

Let a and a_* be the row vectors in $D_{\mathbf{g}}$ corresponding to item combinations $j_1, ..., j_l$ and $j, j_1, ..., j_l$; see (4.19) for the definition of $D_{\mathbf{g}}$. If $(\hat{\mathbf{c}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{p}})$ satisfy by (4.19), then

$$\frac{{a_*}^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}}{a^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}} = \frac{{a_*}^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}}{a^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}}.$$

On the other hand, we have that

$$\begin{aligned}
\frac{{a_*}^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}}{a^\top T_{\hat{\mathbf{c}},\mathbf{g}}(Q)\hat{\mathbf{p}}} &= \frac{B_{\hat{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j, j_1, ..., j_l)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j_1, ..., j_l)\hat{\mathbf{p}}} = \hat{c}_j - g_j, \\
\frac{{a_*}^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}}{a^\top T_{\bar{\mathbf{c}},\mathbf{g}}(Q)\bar{\mathbf{p}}} &= \frac{B_{\bar{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j, j_1, ..., j_l)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\mathbf{g},\mathbf{0};Q}(j_1, ..., j_l)\bar{\mathbf{p}}} = \bar{c}_j - g_j.
\end{aligned}$$

Therefore, $\hat{c}_j = \bar{c}_j$ for all $j = 1, \cdots, J$, which gives the identifiability of the slipping parameter. According to Proposition 5, the completeness of the Q-matrix ensures that the identifiability of $\mathbf{p}$, therefore we have the sufficiency of A4.

Necessity of A4. We reach the conclusion by contradiction. (4.19) suggests that it is sufficient to show the necessity for $\mathbf{g} = \mathbf{0}$. Without loss of generality, suppose that the first attribute only appears once in the first column of the Q-matrix, i.e., the Q-matrix takes the following form:

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q_1 \end{pmatrix}. \tag{4.20}$$

We construct $\bar{\mathbf{c}}$ and $\bar{\mathbf{p}}$ different from $\hat{\mathbf{c}}$ and $\hat{\mathbf{p}}$ such that $T_{\hat{\mathbf{c}},\mathbf{0}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\mathbf{0}}(Q)\bar{\mathbf{p}}$. We write $\hat{\mathbf{c}} = (\hat{c}_1, \cdots, \hat{c}_J)$ and $\hat{\mathbf{p}} = \{\hat{p}_{(b,a)} : b \in \{0,1\}, a \in \{0,1\}^{K-1}\}$. For some x close to 1, define

$$\bar{\mathbf{c}} = (\bar{c}_1, \bar{c}_2, \cdots, \bar{c}_J) = (x\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_J)$$

and

$$\bar{\mathbf{p}} = \{\bar{p}_{(b,a)} : \bar{p}_{(1,a)} = \hat{p}_{(1,a)}/x \text{ and } \bar{p}_{(0,a)} = \hat{p}_{(0,a)} + \hat{p}_{(1,a)}(1 - 1/x), \text{ for all } a \in \{0,1\}^{K-1}\}.$$

Notice that the parameters related to the first item have been changed. Consider the rows in the T-matrix related to the first item. Keeping in mind that $\mathbf{g} = \mathbf{0}$, we have that

$$\hat{c}_1 \sum_{a\in\{0,1\}^{K-1}} \hat{p}_{(1,a)} + g_1 \sum_{a\in\{0,1\}^{K-1}} \hat{p}_{(0,a)} = \bar{c}_1 \sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)} + g_1 \sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)}. \tag{4.21}$$

This corresponds to $P(Y_1 = 1)$. Similar identities can be established for $P(Y_1 = Y_{j_1} = ... = Y_{j_l} = 1)$. Therefore, we have constructed $(\bar{\mathbf{c}}, \bar{\mathbf{p}}) \neq (\hat{\mathbf{c}}, \hat{\mathbf{p}})$ such that $T_{\bar{\mathbf{c}},\mathbf{0}}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}},\mathbf{0}}(Q)\hat{\mathbf{p}}$. Thus, $\mathbf{c}$ and $\mathbf{p}$ are not identifiable if A4 does not hold. □

Proof of Theorem 5. Consider the true Q and a candidate $Q' \nsim Q$. According to the discussion at the end of Section 4.5.1, it is sufficient to show that it is impossible to have two sets of parameters $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ such that $\hat{c}_j > \hat{g}_j$, $\bar{c}_j > \bar{g}_j$, $\hat{p}_{\boldsymbol{\theta}} > 0$, $\bar{p}_{\boldsymbol{\theta}} > 0$, and

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q')\bar{\mathbf{p}}. \tag{4.22}$$

We prove this by first assuming that there exist two such sets of parameters and then reaching a contradiction. The true matrix Q is arranged as in (4.2) such that the first $2K$ rows form two identity matrices. We try to reach a contradiction under the following two cases.

Case 1: either $Q'_{1:K}$ or $Q'_{K+1:2K}$ is incomplete. We only focus on the case when $Q'_{1:K}$ is not $\mathcal{I}_K$. We borrow an intermediate result in the proof of Proposition 6.4 in Liu et al. (2013): we can identify an item $1 \leq h \leq K$ and an item set $\mathcal{H} \subset \{1, \cdots, K\}$ $(h \notin \mathcal{H})$ such that under Q', $\mathcal{H}$ requires all attributes required by item h, that is, if someone is capable of solving all problems in $\mathcal{H}$ then he/she is able to solve problem h. We say someone "is able to" or "can" solve a problem or a set of problems if his/her ideal responses to the set of problems are all one.

For items $K+1, \cdots, 2K$, since $Q_{K+1:2K} = \mathcal{I}_K$, there exists an item set $\mathcal{B} \subset \{K+1, ..., 2K\}$ such that under Q it requires the same attributes as $\mathcal{H}$, that is, if a person is capable of solving all items in $\mathcal{B}$ if and only if they can solving all problems in $\mathcal{H}$. Since $Q_{1:K} = \mathcal{I}_K$, under Q, the attributes required by $\mathcal{H}$ and $\mathcal{B}$ are different from those of item h. Define

$$\tilde{\mathbf{g}} = (\bar{g}_1, \cdots, \bar{g}_K, \hat{g}_{K+1}, \cdots, \hat{g}_J).$$

Assumption (4.22) and Proposition 3 suggests $T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$.

Under Q' if h requires strictly fewer attributes than $\mathcal{H}$, there are three types of attributes profiles: unable to answer h (denoted by $0_h0_{\mathcal{H}}$), unable to answer $\mathcal{H}$ but able to answer h (denoted by $0_{\mathcal{H}}1_h$), and able to answer $\mathcal{H}$ (denoted by $1_{\mathcal{H}}$). We have

$$\begin{array}{rccc} & 0_h0_{\mathcal{H}} & 0_{\mathcal{H}}1_h & 1_{\mathcal{H}} \\ B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H}) = (& 0 & 0 & \prod_{j\in\mathcal{H}}(\bar{c}_j - \bar{g}_j) \quad), \\ B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(h) = (& 0 & (\bar{c}_h - \bar{g}_h) & (\bar{c}_h - \bar{g}_h) \quad), \\ B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h) = (& 0 & 0 & (\bar{c}_h - \bar{g}_h)\prod_{j\in\mathcal{H}}(\bar{c}_j - \bar{g}_j) \quad), \end{array}$$

If h and $\mathcal{H}$ require the same attributes, $0_{\mathcal{H}}1_h$ case does not exist and the above equations do not have the $0_{\mathcal{H}}1_h$ column. Under both situations, we have

$$\bar{c}_h - \bar{g}_h = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H})\bar{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h,K+1,\cdots,2K)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},K+1,\cdots,2K)\bar{\mathbf{p}}}. \tag{4.23}$$

Under Q, we have

$$\begin{aligned}
& & \boldsymbol{\theta} \neq \mathbf{1} & & \boldsymbol{\theta} = \mathbf{1} \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(K+1,\cdots,2K) &= (& 0 & & \textstyle\prod_{j=K+1}^{2K}(\hat{c}_j-\hat{g}_j)), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},K+1,\cdots,2K) &= (& 0 & & \textstyle\prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j=K+1}^{2K}(\hat{c}_j-\hat{g}_j)), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,K+1,\cdots,2K) &= (& 0 & (\hat{c}_h-\bar{g}_h)\textstyle\prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j=K+1}^{2K}(\hat{c}_j-\hat{g}_j)).
\end{aligned}$$

This gives

$$\hat{c}_h-\bar{g}_h=\frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,K+1,\cdots,2K)\hat{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},K+1,\cdots,2K)\hat{\mathbf{p}}}. \tag{4.24}$$

$T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}}=T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$ allows to equate the right-hand sides of (4.23) and (4.24) which yields

$$\hat{c}_h=\bar{c}_h. \tag{4.25}$$

Now under Q', with a similarly argument, we have

$$\bar{c}_h-\bar{g}_h=\frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h,\mathcal{B})\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},\mathcal{B})\bar{\mathbf{p}}}. \tag{4.26}$$

Under Q, consider three types of attributes profiles: unable to answer $\mathcal{H}$ (denoted by $0_{\mathcal{H}}$), able to answer $\mathcal{H}$ but unable to answer h (denoted by $0_h1_{\mathcal{H}}$), and able to answer both $\mathcal{H}$ and h (denoted by $1_{\mathcal{H}}1_h$). We have

$$\begin{aligned}
& & 0_{\mathcal{H}} & & 0_h1_{\mathcal{H}} & & 1_{\mathcal{H}}1_h \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},\mathcal{B}) &= (0 & & \prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j) & & \prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j)), \\
B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,\mathcal{B}) &= (0 & & (\hat{g}_h-\bar{g}_h)\prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j) & & (\hat{c}_h-\bar{g}_h)\prod_{j\in\mathcal{H}}(\hat{c}_j-\bar{g}_j)\prod_{j\in\mathcal{B}}(\hat{c}_j-\hat{g}_j)).
\end{aligned}$$

Since $\hat{g}_h-\bar{g}_h<\hat{c}_h-\bar{g}_h$ and $p_{\boldsymbol{\theta}}>0$ for all $\boldsymbol{\theta}$, we have that

$$\hat{c}_h-\bar{g}_h\neq\frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},h,\mathcal{B})\hat{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q}(\mathcal{H},\mathcal{B})\hat{\mathbf{p}}}=\frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},h,\mathcal{B})\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(\mathcal{H},\mathcal{B})\bar{\mathbf{p}}}. \tag{4.27}$$

$T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}}=T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$ allows use to equate the right-hand sides of (4.26) and (4.27), which yields $\hat{c}_h>\bar{c}_h$. This contradicts (4.25).

Thus, under this case, we have that $T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}}\neq T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$ if $\hat{c}_j>\hat{g}_j$, $\bar{c}_j>\bar{g}_j$, $\hat{p}_{\boldsymbol{\theta}}>0$, $\bar{p}_{\boldsymbol{\theta}}>0$. Furthermore, if the conditions in the theorem are satisfied and $Q'_{1:K}$ or $Q'_{(K+1):2K}$ is incomplete, then we cannot find parameters $\bar{\mathbf{c}}$, $\bar{\mathbf{g}}$, and $\bar{\mathbf{p}}$ that yields the same response distribution as Q and thus Q can be differentiated from Q' by the maximum likelihood.

Case 2: both $Q'_{1:K}$ and $Q'_{K+1:2K}$ are complete, but $Q \nsim Q'$. In this case, we can always arrange the columns of Q' such that $Q'_{1:K} = \mathcal{I}_K$. Redefine

$$\tilde{\mathbf{g}} = (\bar{c}_1, \cdots, \bar{c}_K, \hat{c}_{K+1}, \cdots, \hat{c}_{2K}, 0, \cdots, 0)$$

and assumption (4.22) suggests that $T_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}}}(Q')\bar{\mathbf{p}}$.

The row vectors of T-matrices corresponding to items 1,..., $2K$ are

$$B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, 2K) = \left(\prod_{k=1}^{K}(\hat{g}_k - \bar{c}_k)\prod_{k=K+1}^{2K}(\hat{g}_k - \hat{c}_k), \mathbf{0}^\top\right)$$

and

$$B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, 2K) = \left(\prod_{k=1}^{K}(\bar{g}_k - \bar{c}_k)\prod_{k=K+1}^{2K}(\bar{g}_k - \hat{c}_k), \mathbf{0}^\top\right)$$

where only the element corresponding to zero attribute is non-zero. Therefore, for any $j \geq 2K+1$, we have

$$\hat{g}_j = \frac{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, 2K, j)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, 2K)\hat{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, 2K, j)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, 2K)\bar{\mathbf{p}}} = \bar{g}_j.$$

Once again, we redefine $\tilde{\mathbf{g}} = (\bar{g}_1, \cdots, \bar{g}_K, 0, \cdots, 0, \hat{g}_{2K+1}, \cdots, \hat{g}_J)$. By Condition A5, we have for $K+1 \leq j \leq 2K$

$$\begin{aligned}\hat{c}_j &= \frac{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K, j, (2K+1), \cdots, J)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K, (2K+1), \cdots, J)\hat{\mathbf{p}}} \\ &= \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K, j, (2K+1), \cdots, J)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K, (2K+1), \cdots, J)\bar{\mathbf{p}}} = \bar{c}_j.\end{aligned}$$

Similarly take $\tilde{\mathbf{g}} = (0, \cdots, 0, \bar{g}_{K+1}, \cdots, \bar{g}_{2K}, \hat{g}_{2K+1}, \cdots, \hat{g}_J)$. We have $\hat{c}_j = \bar{c}_j$ for $1 \leq j \leq K$.

Now take $\tilde{\mathbf{g}} = (\bar{c}_1, \cdots, \bar{c}_K, 0, \cdots, 0)$, we have for $K+1 \leq j \leq 2K$

$$\hat{g}_j = \frac{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K, j)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\tilde{\mathbf{g}},\hat{\mathbf{g}}-\tilde{\mathbf{g}},Q}(1, \cdots, K)\hat{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K, j)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\tilde{\mathbf{g}},\bar{\mathbf{g}}-\tilde{\mathbf{g}},Q'}(1, \cdots, K)\bar{\mathbf{p}}} = \bar{g}_j.$$

Similarly, for $\hat{g}_j = \bar{g}_j$ for $j = 1, ..., K$. Thus, we have $\hat{g}_j = \bar{g}_j$ for $j = 1, ..., J$. Therefore, assumption (4.22) becomes

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\hat{\mathbf{g}}}(Q')\bar{\mathbf{p}}. \tag{4.28}$$

This contradicts Proposition 4. Thus, we have reached the conclusion that

$$T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} \neq T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q')\bar{\mathbf{p}}.$$

for all $\hat{c}_j > \hat{g}_j$, $\bar{c}_j > \bar{g}_j$, $\hat{p}_{\boldsymbol{\theta}} > 0$, $\bar{p}_{\boldsymbol{\theta}} > 0$ and $Q' \nsim Q$. Thus, by maximizing the profiled likelihood, Q can be consistently estimated. □

Proof of Theorem 6. Suppose there are two sets of parameters $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ such that $L(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}) = L(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$, equivalently, $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$. We show that $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}) = (\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ if $\hat{c}_j > \hat{g}_j$, $\hat{p}_{\boldsymbol{\theta}} > 0$, $\bar{c}_j > \bar{g}_j$, and $\bar{p}_{\boldsymbol{\theta}} > 0$. Condition A5 allows us to consider the following three cases.

Case 1. There exit at least three items with Q-matrix row vector $\mathbf{e}_1$. Without loss of generality, we write the Q-matrix as (with reordering of the rows)

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{0}^\top \\ 1 & \mathbf{0}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q' \end{pmatrix}. \tag{4.29}$$

In what follows, we show that $\hat{c}_j = \bar{c}_j$ and $\hat{g}_j = \bar{g}_j$ for $j = 1, 2, 3$. By Proposition 3, $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$ suggests that $T_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}}}(Q)\bar{\mathbf{p}}$. Together with the fact that

$$\frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,2,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,2)\hat{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1)\hat{\mathbf{p}}} = \hat{c}_3 - \hat{g}_3, \tag{4.30}$$

we have that

$$\frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1)\bar{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,2,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,2)\bar{\mathbf{p}}}. \tag{4.31}$$

Expanding the above identity, we have

$$\begin{aligned} & \frac{(\bar{g}_1 - \hat{g}_1)(\bar{g}_3 - \hat{g}_3)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + (\bar{c}_1 - \hat{g}_1)(\bar{c}_3 - \hat{g}_3)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}}{(\bar{g}_1 - \hat{g}_1)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + (\bar{c}_1 - \hat{g}_1)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}} \\ = & \frac{\prod_{j=1}^{3}(\bar{g}_j - \hat{g}_j)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + \prod_{j=1}^{3}(\bar{c}_j - \hat{g}_j)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}}{(\bar{g}_1 - \hat{g}_1)(\bar{g}_2 - \hat{g}_2)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(0,a)} + (\bar{c}_1 - \hat{g}_1)(\bar{c}_2 - \hat{g}_2)\sum_{a\in\{0,1\}^{K-1}} \bar{p}_{(1,a)}}, \end{aligned} \tag{4.32}$$

which can be simplified to $(\bar{g}_1 - \hat{g}_1)(\bar{c}_1 - \hat{g}_1)(\bar{c}_2 - \bar{g}_2)(\bar{c}_3 - \bar{g}_3) = 0$. Then under the constraint that $\bar{c}_j > \bar{g}_j$, we have $\bar{g}_1 = \hat{g}_1$ or $\bar{c}_1 = \hat{g}_1$. A similar argument yields

$$\begin{cases} \bar{g}_2 = \hat{g}_2 \text{ or } \bar{c}_2 = \hat{g}_2 \\ \bar{g}_3 = \hat{g}_3 \text{ or } \bar{c}_3 = \hat{g}_3 \end{cases} \text{ and } \begin{cases} \hat{g}_1 = \bar{g}_1 \text{ or } \hat{c}_1 = \bar{g}_1 \\ \hat{g}_2 = \bar{g}_2 \text{ or } \hat{c}_2 = \bar{g}_2 \\ \hat{g}_3 = \bar{g}_3 \text{ or } \hat{c}_3 = \bar{g}_3 \end{cases}.$$

For $j = 1$, 2, or 3, if $\hat{g}_j \neq \bar{g}_j$ we have $\hat{c}_j = \bar{g}_j$ and $\bar{c}_j = \hat{g}_j$. This contradict the condition that $\hat{c}_j > \hat{g}_j$ and $\bar{c}_j > \bar{g}_j$. Thus we have $\hat{g}_j = \bar{g}_j$ for $j = 1, 2, 3$. Repeating the proof of Theorem 4, we have $\hat{c}_j = \bar{c}_j$ for $i = 1, 2, 3$.

Case 2. There exit two items with row vector $\mathbf{e}_1$. Without loss of generality, we write the Q-matrix as

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{0}^\top \\ 1 & \mathbf{v}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q' \end{pmatrix}, \quad Q_{1:4} = \begin{pmatrix} 1 & 0 & \mathbf{0}^\top \\ 1 & 0 & \mathbf{0}^\top \\ 1 & 1 & {\mathbf{v}_*}^\top \\ 0 & 1 & \mathbf{0}^\top \end{pmatrix}, \tag{4.33}$$

where $\mathbf{v}$ is a non-zero vector. Without loss of generality we assume $\mathbf{v}^\top = (1, {\mathbf{v}_*}^\top)$. Consider the sub-matrix containing the first four items. i.e., $Q_{1:4}$ in (4.33). Similar to the proof of Case 1, for $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ such that $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$, we will show

$$\begin{cases} \hat{c}_j = \bar{c}_j & j = 1, 2, 4 \\ \hat{g}_j = \bar{g}_j & j = 1, 2, 3 \end{cases}. \tag{4.34}$$

A similar argument as in Case 1 yields

$$\frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(3)\hat{\mathbf{p}}} = \hat{c}_1 - \hat{g}_1 = \frac{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(1,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0};\mathbf{Q}}(4,3)\hat{\mathbf{p}}}.$$

Together with the fact that $T_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}}}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}}-\hat{\mathbf{g}},\mathbf{0}}(Q)\hat{\mathbf{p}}$, we have

$$\frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(3)\bar{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(1,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\hat{\mathbf{g}},\bar{\mathbf{g}}-\hat{\mathbf{g}};Q}(4,3)\bar{\mathbf{p}}}.$$

This implies

$$\begin{aligned} & \frac{\tilde{g}_1\tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{c}_1\tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{g}_1\tilde{c}_4\tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_1\tilde{c}_4\tilde{c}_3\bar{\mathbf{p}}_{1,1}}{\tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{g}_4\tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{c}_4\tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_4\tilde{c}_3\bar{\mathbf{p}}_{1,1}} \\ = & \frac{\tilde{g}_1\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{c}_1\tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{g}_1\tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_1\tilde{c}_3\bar{\mathbf{p}}_{1,1}}{\tilde{g}_3\bar{\mathbf{p}}_{0,0} + \tilde{g}_3\bar{\mathbf{p}}_{1,0} + \tilde{g}_3\bar{\mathbf{p}}_{0,1} + \tilde{c}_3\bar{\mathbf{p}}_{1,1}}, \end{aligned} \tag{4.35}$$

where $\tilde{g}_j = \bar{g}_j - \hat{g}_j$ for $j = 1, 3, 4$, $\tilde{c}_j = \bar{c}_j - \hat{g}_j$ for $j = 1, 4$,

$$\tilde{c}_3 = \frac{(\bar{c}_3 - \hat{g}_3)\sum_{\mathbf{v}_* \preceq a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)} + (\bar{g}_3 - \hat{g}_3)\sum_{\mathbf{v}_* \not\preceq a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)}}{\sum_{a \in \{0,1\}^{K-2}} \bar{p}_{(1,1,a)}},$$

and $\bar{\mathbf{p}}_{i,j} = \sum_{a\in\{0,1\}^{K-2}} \bar{p}_{(i,j,a)}$ for $i,j \in \{0,1\}$. Here $\mathbf{v}_* \preceq a$ means that each element of $\mathbf{v}_*$ is less than or equals to the corresponding element of a, and $\mathbf{v}_* \npreceq a$ means that $\mathbf{v}_* \preceq a$ does not hold.

Simplifying (4.35), we obtain $\bar{\mathbf{p}}_{0,0}\bar{\mathbf{p}}_{1,1}\tilde{g}_3\tilde{c}_3(\tilde{g}_1-\tilde{c}_1) = \bar{\mathbf{p}}_{1,0}\bar{\mathbf{p}}_{0,1}\tilde{g}_3\tilde{g}_3(\tilde{g}_1-\tilde{c}_1)$. Since $\tilde{g}_1-\tilde{c}_1 \neq 0$, we have

$$\tilde{g}_3 = 0 \quad \text{or} \quad \bar{\mathbf{p}}_{0,0}\bar{\mathbf{p}}_{1,1}\tilde{c}_3 = \bar{\mathbf{p}}_{1,0}\bar{\mathbf{p}}_{0,1}\tilde{g}_3. \tag{4.36}$$

We show that $\tilde{g}_3$ has to be zero. Otherwise, we have

$$\bar{\mathbf{p}}_{0,0}\bar{\mathbf{p}}_{1,1}(\bar{c}_3^* - \hat{g}_3) = \bar{\mathbf{p}}_{1,0}\bar{\mathbf{p}}_{0,1}(\bar{g}_3 - \hat{g}_3), \tag{4.37}$$

where

$$\bar{c}_3^* = \tilde{c}_3 + \hat{g}_3 = \frac{\bar{c}_3 \sum_{v_*\preceq a\in\{0,1\}^{K-2}} \bar{p}_{(1,1,a)} + \bar{g}_3 \sum_{v_*\npreceq a\in\{0,1\}^{K-2}} \bar{p}_{(1,1,a)}}{\sum_{a\in\{0,1\}^{K-2}} \bar{p}_{(1,1,a)}}.$$

A similar argument gives that

$$\hat{\mathbf{p}}_{0,0}\hat{\mathbf{p}}_{1,1}(\hat{c}_3^* - \bar{g}_3) = \hat{\mathbf{p}}_{1,0}\hat{\mathbf{p}}_{0,1}(\hat{g}_3 - \bar{g}_3), \tag{4.38}$$

where

$$\hat{c}_3^* = \frac{\hat{c}_3 \sum_{v_*\preceq a\in\{0,1\}^{K-2}} \hat{p}_{(1,1,a)} + \hat{g}_3 \sum_{v_*\npreceq a\in\{0,1\}^{K-2}} \hat{p}_{(1,1,a)}}{\sum_{a\in\{0,1\}^{K-2}} \hat{p}_{(1,1,a)}}.$$

Equations (4.37) and (4.38) imply that $\hat{c}_3^* > \hat{g}_3 > \bar{c}_3^* > \bar{g}_3$ or $\bar{c}_3^* > \bar{g}_3 > \hat{c}_3^* > \hat{g}_3$, which conflicts with the equation that $B_{\hat{\mathbf{c}},\hat{\mathbf{g}};Q}(3)\hat{\mathbf{p}} = B_{\bar{\mathbf{c}},\bar{\mathbf{g}};Q}(3)\bar{\mathbf{p}}$, i.e.,

$$\hat{g}_3(\hat{\mathbf{p}}_{0,0} + \hat{\mathbf{p}}_{1,0} + \hat{\mathbf{p}}_{0,1}) + \hat{c}_3^*\hat{\mathbf{p}}_{1,1} = \bar{g}_3(\bar{\mathbf{p}}_{0,0} + \bar{\mathbf{p}}_{1,0} + \bar{\mathbf{p}}_{0,1}) + \bar{c}_3^*\bar{\mathbf{p}}_{1,1}.$$

To see this, notice that $\hat{\mathbf{p}}_{0,0} + \hat{\mathbf{p}}_{1,0} + \hat{\mathbf{p}}_{0,1} = 1 - \hat{\mathbf{p}}_{1,1}$, $\bar{\mathbf{p}}_{0,0} + \bar{\mathbf{p}}_{1,0} + \bar{\mathbf{p}}_{0,1} = 1 - \bar{\mathbf{p}}_{1,1}$, and $\hat{\mathbf{p}}_{1,1}, \bar{\mathbf{p}}_{1,1} \in (0,1)$. By simple algebra, the above identity cannot be achieved if either $\hat{c}_3^* > \hat{g}_3 > \bar{c}_3^* > \bar{g}_3$ or $\bar{c}_3^* > \bar{g}_3 > \hat{c}_3^* > \hat{g}_3$ is true. Therefore, we have $\tilde{g}_3 = \bar{g}_3 - \hat{g}_3 = 0$. Let $\underline{\mathbf{g}} = (0, 0, \hat{g}_3, 0, \cdots, 0)$. $T_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}}}(Q)\bar{\mathbf{p}} = T_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}}}(Q)\hat{\mathbf{p}}$ yields

$$\begin{aligned}
\bar{c}_1 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\hat{\mathbf{p}}} = \hat{c}_1,\\
\bar{c}_2 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(2,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(2,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(4,3)\hat{\mathbf{p}}} = \hat{c}_2,\\
\bar{c}_4 &= \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{g}},\bar{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,3)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,4,3)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{g}},\hat{\mathbf{g}}-\underline{\mathbf{g}};Q}(1,3)\hat{\mathbf{p}}} = \hat{c}_4.
\end{aligned}$$

Consider items 1 and 2. Let $\underline{\mathbf{c}} = (\hat{c}_1, \hat{c}_2, 0, \cdots, 0)$. $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$ yields

$$\bar{g}_1 = \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(2)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(2)\hat{\mathbf{p}}} = \hat{g}_1,$$
$$\bar{g}_2 = \frac{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\underline{\mathbf{c}},\bar{\mathbf{g}}-\underline{\mathbf{c}};Q}(1)\bar{\mathbf{p}}} = \frac{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(1,2)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\underline{\mathbf{c}},\hat{\mathbf{g}}-\underline{\mathbf{c}};Q}(1)\hat{\mathbf{p}}} = \hat{g}_2.$$

Therefore, (4.34) is true.

Case 3. There exits only one item with row vector $\mathbf{e}_1$, i.e., the Q-matrix can be written as

$$Q = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{v_2}^\top \\ 1 & \mathbf{v_3}^\top \\ \mathbf{0} & \mathcal{I}_{K-1} \\ \mathbf{0} & Q' \end{pmatrix}, \tag{4.39}$$

where $\mathbf{v_2}$ and $\mathbf{v_2}$ are non-zero vectors. Consider the sub-matrices:

$$Q_a = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{v_2}^\top \\ \mathbf{0} & \mathbf{e}_{h_2} \end{pmatrix} \quad and \quad Q_b = \begin{pmatrix} 1 & \mathbf{0}^\top \\ 1 & \mathbf{v_3}^\top \\ \mathbf{0} & \mathbf{e}_{h_3} \end{pmatrix}, \tag{4.40}$$

where $\mathbf{e}_{h_2} \preceq \mathbf{v_2}$ and $\mathbf{e}_{h_3} \preceq \mathbf{v_3}$. With a similar proof as in Case 2, we have $\hat{c}_1 = \bar{c}_1$ and $\hat{g}_2 = \bar{g}_2$,

Now combining the results in Cases 1-3, we have that for the Q-matrix taking the form of (4.2), the following holds:

$$\begin{cases} \hat{c}_j = \bar{c}_j & j = 1, \cdots, K \\ \hat{g}_j = \bar{g}_j & j = (K+1), \cdots, J \end{cases}. \tag{4.41}$$

Let

$$\mathbf{g}^* = (\hat{c}_1, \cdots, \hat{c}_K, \hat{g}_{K+1}, \cdots, \hat{g}_J).$$

For each $j \in \{(K+1), \cdots, J\}$, let $\mathcal{A}_j$ be the set of items $\{(K+1), \cdots, J\} \backslash \{j\}$, i.e., the set of all items from $K+1$ to J except the jth one. For the sub-matrix $Q_{K+1:J}$, condition A5

implies that each attribute appears at least twice. Therefore, we have

$$\hat{c}_j - \hat{g}_j = \frac{B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j, j)\hat{\mathbf{p}}}{B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j)\hat{\mathbf{p}}} = \frac{B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j, j)\bar{\mathbf{p}}}{B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{A}_j)\bar{\mathbf{p}}} = \bar{c}_j - \hat{g}_j.$$

This gives $\hat{c}_j = \bar{c}_j$ for $j = K+1, \cdots, J$. Together with (4.41), $\hat{c}_j = \bar{c}_j$ for all j.

Under condition A6, for each $k = 1, \cdots, K$, there exists an item set $\mathcal{B}_k \subset \{K+1, \cdots, J\}$ such that $\mathcal{B}_k$ requires all attributes except the kth one. Since item set $\cup_{k=1}^K \mathcal{B}_k$ requires all attributes, we obtain

$$B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\cup_{k=1}^K \mathcal{B}_k)\hat{\mathbf{p}} = \hat{p}_{\mathbf{1}} \prod_{j \in \cup_{k=1}^K \mathcal{B}_k} (\hat{c}_j - \hat{g}_j),$$

$$B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\cup_{k=1}^K \mathcal{B}_k)\bar{\mathbf{p}} = \bar{p}_{\mathbf{1}} \prod_{j \in \cup_{k=1}^K \mathcal{B}_k} (\bar{c}_j - \hat{g}_j),$$

where $\hat{p}_{\mathbf{1}}$ and $\bar{p}_{\mathbf{1}}$ are the proportions of attribute $\mathbf{1}$. By (4.41) and the equation that $B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\cup_{k=1}^K \mathcal{B}_k)\hat{\mathbf{p}} = B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\cup_{k=1}^K \mathcal{B}_k)\bar{\mathbf{p}}$, we have $\hat{p}_{\mathbf{1}} = \bar{p}_{\mathbf{1}}$. In addition, from the equation that $B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{B}_k)\hat{\mathbf{p}} = B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{B}_k)\bar{\mathbf{p}}$, we have

$$(\hat{p}_{\mathbf{1}-\mathbf{e}_k} + \hat{p}_{\mathbf{1}}) \prod_{j \in \mathcal{B}_k} (\hat{c}_j - \hat{g}_j) = (\bar{p}_{\mathbf{1}-\mathbf{e}_k} + \bar{p}_{\mathbf{1}}) \prod_{j \in \mathcal{B}_k} (\bar{c}_j - \hat{g}_j),$$

which implies that $\hat{p}_{\mathbf{1}-\mathbf{e}_k} = \bar{p}_{\mathbf{1}-\mathbf{e}_k}$, for $k = 1, \cdots, K$.

Since the kth item only requires the kth attribute. We have

$$B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{B}_k, k)\hat{\mathbf{p}} = \hat{p}_{\mathbf{1}-\mathbf{e}_k}(\hat{g}_k - \hat{c}_k) \prod_{j \in \mathcal{B}_k} (\hat{c}_j - \hat{g}_j),$$

$$B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*;Q}(\mathcal{B}_k, k)\bar{\mathbf{p}} = \bar{p}_{\mathbf{1}-\mathbf{e}_k}(\bar{g}_k - \hat{c}_k) \prod_{j \in \mathcal{B}_k} (\bar{c}_j - \hat{g}_j).$$

Identities in (4.41), $\hat{p}_{\mathbf{1}-\mathbf{e}_k} = \bar{p}_{\mathbf{1}-\mathbf{e}_k}$, and $B_{\bar{\mathbf{c}}-\mathbf{g}^*,\bar{\mathbf{g}}-\mathbf{g}^*}(\mathcal{B}_k, k)\bar{\mathbf{p}} = B_{\hat{\mathbf{c}}-\mathbf{g}^*,\hat{\mathbf{g}}-\mathbf{g}^*}(\mathcal{B}_k, k)\hat{\mathbf{p}}$ yields $\hat{g}_k = \bar{g}_k$, for $j = 1, \cdots, K$. Thus, we have $\hat{c}_j = \bar{c}_j$ and $\hat{g}_j = \bar{g}_j$ for all $j = 1, ..., J$. This further yields $\hat{\mathbf{p}} = \bar{\mathbf{p}}$ due to the full column rank of the matrix $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)$.

Therefore, for two sets of parameters $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}})$ and $(\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$ that $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$, we have $(\hat{\mathbf{c}}, \hat{\mathbf{g}}, \hat{\mathbf{p}}) = (\bar{\mathbf{c}}, \bar{\mathbf{g}}, \bar{\mathbf{p}})$. This finishes the proof of Theorem 6. □

4.5.3 Appendix C: Proof of Propositions

Proof of Proposition 2. Notice that the column vector $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$ contains the probabilities $P(Y_{j_1} = 1, ..., Y_{j_l} = 1)$ for all possible distinct combinations j_1,...,j_l. Thus, $T_{\mathbf{c},\mathbf{g}}(Q)\mathbf{p}$ com-

pletely characterizes the distribution of $\mathbf{Y}$. Two sets of parameters $T_{\hat{\mathbf{c}},\hat{\mathbf{g}}}(Q)\hat{\mathbf{p}} = T_{\bar{\mathbf{c}},\bar{\mathbf{g}}}(Q)\bar{\mathbf{p}}$ if and only if they correspond to the same distribution of $\mathbf{Y}$. This concludes the proof. □

Proof of the Proposition 3. In what follows, we construct a D matrix satisfying the condition in the proposition. We show that there exists a matrix D only depending on g^* so that $DT_{\mathbf{c},\mathbf{g}}(Q) = T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q)$. Note that each row of $DT_{\mathbf{c},\mathbf{g}}(Q)$ is just a row linear transform of $T_{\mathbf{c},\mathbf{g}}(Q)$. Then, it is sufficient to show that each row vector of $T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q)$ is a linear transform of rows of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. We prove this by induction.

First, note that

$$B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j) = B_{\mathbf{c},\mathbf{g};Q}(j) - g_j^*\mathbf{1}^\top$$

where $\mathbf{1}^\top$ is a row vector with all elements being 1. Then all row vectors of $T_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*}(Q)$ of the form $B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*,Q}(j)$ are inside the row space of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. Suppose that all the vectors of the form

$$B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_l)$$

for all $1 \leq l \leq \iota$ can be written linear combinations of the row vectors of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. Then, we consider

$$B_{\mathbf{c},\mathbf{g};Q}(j_1, ..., j_{\iota+1}) = \Upsilon_{h=1}^{\iota+1}\left(B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_h) + g_{j_h}^*\mathbf{1}^\top\right),$$

where "Υ" refers to element by element multiplication. The left hand side is just a row vector of $T_{\mathbf{c},\mathbf{g}}(Q)$. We expand the right hand side of the above display. Note that the last term is precisely

$$B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_{\iota+1}) = \Upsilon_{h=1}^{\iota+1}B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_h).$$

The rest terms are all of the form $B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_l)$ for $1 \leq l \leq \iota$ multiplied by coefficients only depending on g^*. Therefore, according to the induction assumption, we have that $B_{\mathbf{c}-\mathbf{g}^*,\mathbf{g}-\mathbf{g}^*;Q}(j_1, ..., j_{\iota+1})$ can be written as linear combinations of rows of $T_{\mathbf{c},\mathbf{g}}(Q)$ with coefficients only depending on g^*. □

Bibliography

Agresti, A. (1996). *Categorical data analysis*, volume 996. New York: John Wiley & Sons.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265.

Allen, N. L., Carlson, J. E., and Zelenak, C. A. (1999). *The 1996 NAEP Technical Report.* ERIC.

American Psychiatric Association (1994). *DSM-IV: Diagnostic and statistical manual of mental disorders (4th ed.).* Washington, DC: American Psychiatric Association.

Bach, F. R. (2008). Consistency of trace norm minimization. *The Journal of Machine Learning Research*, 9:1019–1048.

Béguin, A. A. and Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4):541–561.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*, pages 395–479.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46(4):443–459.

Bock, R. D., Gibbons, R., and Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12(3):261–280.

Bolt, D. M. and Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, 27(6):395–414.

Bouwmans, T. and Zahzah, E. H. (2014). Robust PCA via principal component pursuit: a review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika*, 64(2):153–168.

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. *Psychometrika*, 76(1):57–76.

Braeken, J., Tuerlinckx, F., and De Boeck, P. (2007). Copula functions for residual dependency. *Psychometrika*, 72(3):393–411.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36(1):111–150.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, 75(1):33–57.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.

Cattell, R. (1978). *The scientific use of factor analysis in behavioral and life sciences*. Springer Science & Business Media.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596.

Chen, W.-H. and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3):265–289.

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.

Chiu, C., Douglas, J., and Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74(4):633–665.

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

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

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)*, 39:1–38.

Embretson, S. E. and Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Eysenck, S. and Barrett, P. (2013). Re-introduction to cross-cultural studies of the EPQ. *Personality and Individual Differences*, 54(4):485–489.

Eysenck, S. B., Eysenck, H. J., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6(1):21–29.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *Information Theory, IEEE Transactions on*, 57(8):5467–5484.

Fazel, M., Hindi, H., and Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE.

Fraser, C. and McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23(2):267–269.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40.

Gentle, J. E. (2009). *Computational statistics*, volume 308. Springer.

Gibbons, R. D. and Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57(3):423–436.

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76.

Grant, B. F., Kaplan, K., Shepard, J., and Moore, T. (2003). Source and accuracy statement for wave 1 of the 2001–2002 national epidemiologic survey on alcohol and related conditions. *Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism.*

Groth-Marnat, G. (2009). *Handbook of psychological assessment.* John Wiley & Sons Inc.

Gulliksen, H. (1950). *Theory of mental tests.* John Wiley & Sons Inc.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9(2):139–150.

Haberman, S. J. (2007). The interaction model. In *Multivariate and Mixture Distribution Rasch Models*, pages 201–216. Springer.

Haley, D. C. (1952). Estimation of the dosage mortality relationship when the dose is subject to error (no. tr15). Technical report, Stanford University: Applied Mathematics and Statistics Labortory.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* PhD thesis, University of Illinois at Urbana-Champaign.

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(2):191–210.

Hoskens, M. and De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, 2(3):261.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258.

Iza, M., Wall, M., Heimberg, R., Rodebaugh, T., Schneier, F., Liu, S.-M., and Blanco, C. (2014). Latent structure of social fears and social anxiety disorders. *Psychological Medicine*, 44(02):361–370.

Jeon, M., Rijmen, F., and Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, 38(1):32–60.

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when $p << n$. *Statistica Sinica*, 20(2):595–611.

Johnson, M. S. et al. (2007). Marginal maximum likelihood estimation of item response models in R. *Journal of Statistical Software*, 20(10):1–24.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34:183–202.

Junker, B. W. (1999). Some statistical models and computational methods that may be useful for cognitively-relevant assessment. Prepared for the Committee on the Foundations of Assessment, National Research Council, November 30, 1999.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Knowles, E. S. and Condon, C. A. (2000). Does the rose still smell as sweet? item variability across test forms and revisions. *Psychological Assessment*, 12(3):245.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press.

Lee, J. M. (2009). *Manifolds and differential geometry*, volume 107. American Mathematical Society Providence.

Liu, J., Xu, G., and Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*, 19(5A):1790–1817.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

Masters, G. N. and Wright, B. D. (1997). The partial credit model. In *Handbook of modern item response theory*, pages 101–121. Springer.

McKinley, R. L. and Reckase, M. D. (1982). The use of the general rasch model with multidimensional item response data. Iowa City, IA: American College Testing.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2):159–176.

Osborne, J. W. (2015). What is rotating in exploratory factor analysis? *Practical Assessment, Research & Evaluation*, 20(2):1–7.

Parlett, B. N. (1980). *The symmetric eigenvalue problem*, volume 7. SIAM.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional Ising model selection using L1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

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

Revuelta, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, 38(7):549–562.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163.

Rockafellar, R. T. (2015). *Convex analysis*. Princeton University Press.

Rupp, A. and Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68(1):78–96.

Rupp, A. and Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6(4):219–262.

Rupp, A., Templin, J., and Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Schwarz, G. (1978). Estimating dimension of a model. *Annals of Statistics*, 6(2):461–464.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American Psychologist*, 54(2):93.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242.

Sireci, S. G., Thissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3):237–247.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 18(2):161–169.

Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology, 1904-1920*, 5(4):417–426.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):337–350.

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

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

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. *Test Theory for a New Generation of Tests*, pages 79–97.

Thissen, D. and Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51(4):567–577.

Thurstone, L. (1947). *Multiple Factor Analysis*. University of Chicago Press: Chicago Multiple Factor Analysis.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11(1):1–13.

Tuerlinckx, F. and De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6(2):181.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2):287–307.

von Davier, M. (2014a). 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. (2014b). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*, 2014(2):1–13.

von Davier, M. and Yamamoto, K. (2004). Partially observed mixtures of irt models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, 28(6):389–406.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

Wang, W.-C. and Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, 29(2):126–149.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45.

Wilson, M. and Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60(2):181–198.

Yao, L. and Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, 30(6):469–492.

Yatsenko, D., Josić, K., Ecker, A. S., Froudarakis, E., Cotton, R. J., and Tolias, A. S. (2015). Improved estimation and interpretation of correlations in neural circuits. *PLoS Comput Biol*, 11(3):e1004083.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3):187–213.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zhou, Z., Li, X., Wright, J., Candes, E., and Ma, Y. (2010). Stable principal component pursuit. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1518–1522. IEEE.