Statistical Estimation and Inference for Large-Scale Categorical Data

by

Chengcheng Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2022

Doctoral Committee:

Associate Professor Gongjun Xu, Chair
Assistant Professor Yunxiao Chen
Professor Naiysin Wang
Assistant Professor Zhenke Wu

lccvic@umich.edu
ORCID ID: 0000-0002-9539-7455

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor Dr. Gongjun Xu. He is the best advisor one could ever hope for. Without his constant support, guidance and encouragement, I could not have possibly completed this dissertation. Gongjun is an exemplar researcher who is always sharp-minded and productive; an inspirational role model who is always hard-working and passionate about research; an amazing advisor, who always offers insightful comments and hands-on help; a caring mentor and a close friend, who is always supportive and encouraging. Since I first met him, he has been clear that he would always keep his door open whenever I have any questions to discuss or research updates to share. I am deeply indebted to Gongjun for devoting so much of his precious time and energy to guiding and mentoring me throughout my graduate study and for bringing the best part of me.

I am also very grateful to Dr. Yunxiao Chen, Dr. Naisyin Wang and Dr. Zhenke Wu for serving on my doctoral dissertation committee and for providing me with invaluable comments. I first met Yunxiao two years ago at the London School of Economies and Politics (LSE). Since then, I had been working with him very closely on various research projects. He always amazed me by his sharpness on research, his

ability of distilling the essence of a problem rapidly, and his broad knowledge and deep insights about many topics. His guidance and help are very valuable to me on my research journey as a graduate student. I first met Naisyin on a reading group on data differential privacy. I feel very lucky to be able to work with her later on the same topic. She is a very knowledgeable advisor, and a caring and supportive mentor. She would always give very detailed and helpful comments each time we met, from which I benefited a lot. I got to know Zhenke in the first year of my graduate study and I am very thankful to him for serving on the committee for both my doctoral preliminary exam and my final dissertation. I am always amazed by his sharpness in thinking and expertise in a wide variety of topics, and I learned a lot from his insightful comments and ideas each time we met.

My thanks also go to other faculty members and staffs in the department of statistics, for their help and support over the past few years. I am especially grateful to Dr. Liza Levina and Dr. Ji Zhu for all the care and support they offered when serving as the Ph.D. program director. I would also like to thank Dr. Xuming He. I am truly impressed by his charming leadership style and learned a lot from the interesting talks he gave. I want to thank all the staff members, including Jean, Judy, Virggie, Bebe and Gina, who always patiently helped with my questions and warmly welcomed me into the office with big smiles.

To all my friends in Ann Arbor: my graduate life would not have been so wonderful without you. I especially want to thank everyone in my research group and in my cohort. It was a great pleasure to meet and spend time with you. I also want to thank my roommates and all the friends I made on the badminton courts,

whose companion made my graduate life so delightful. I heartily thank family for their unconditional love and support through my ups and downs. Though I have not lived together with them for years, what I know for sure is wherever I go, they would happily see me pursue my dreams and be my strongest support. Knowing this gives me constant courage to conquer any challenges in the future.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

Abstract

Categorical data become increasingly ubiquitous in the modern big data era. In this dissertation, we propose novel statistical learning and inference methods for large-scale categorical data, focusing on latent variable models and their applications to psychometrics. In psychometric assessments, the subjects' underlying aptitude often cannot be fully captured by raw scores due to differing item difficulties. Latent variable models, are popularly used to capture this unobserved proficiency. This dissertation studies two types of latent variable models with categorical responses. The first type assumes multiple discrete latent traits, commonly known as cognitive diagnosis models (CDMs), a special family of discrete latent variable models. The second type assumes a continuous latent score, commonly known as the item response theory (IRT) models. Although both have been widely applied in large-scale assessments, many challenges still exist for efficient learning and statistical inference. This dissertation studies four important problems that arise in these contexts.

The first part develops novel algorithms to estimate large latent Q-matrix in CDMs. Q-matrix plays an important role in CDMs; it specifies the inter-dependence between items and subjects' latent attributes. Accurate knowledge of Q-matrix is critical for cognitive diagnoses, item categorization and assessment design. However, in practice, many assessments either do not have accurate Q-matrix specification or even do not provide Q-matrix. Furthermore, existing methods are not scalable with the size of Q-matrix, despite the prevalence of large Q-matrix. We propose a penalized likelihood approach, with computational complexity growing linearly with Q sizes, to learn large Q-matrix from observational data. The estimation consistency

and the robustness of the proposed method across various CDMs are also established.

The second part develops learning and inference methods for a unidimensional IRT model, the Rasch model, under the missing data setting. Data missingness is prevalent in large-scale assessments; examples include SAT and GRE where subjects' responses are combined from multiple tests administered year-round from a large item pool. Direct inference to compare subjects' latent scores under the missing data setting remains open and challenging in the literature. In this part, we obtain point estimators for the latent scores and derive their asymptotic distribution under a flexible missing-entry design in double asymptotic settings. We show our estimator is statistically efficient and optimal, which is amongst the first results in the binary matrix completion literature.

The third part concerns measurement biases in IRT models. Novel estimation and inference procedures are developed for biases brought by measurement non-invariant items under the differential item functioning (DIF) framework. Existing methods either require knowing anchor items, i.e. DIF-free items or adopt regularization to ensure model identifiability where easy inference is not permitted. We propose a novel minimal L_1 condition for simultaneous DIF detection and model identification. It does not require any knowledge of anchor items and permits easy inference for both binary and multiple groups settings.

The fourth part considers privacy issues for releasing tabular (categorical) data to the public. In the differential privacy (DP) framework, we recommend an optimal mechanism, where data utility is maximized under a privacy constraint. Common users' practices, including merging related cells or integrating multiple data sources,

are considered. Valid inference procedures are developed for the associated privacy-protected data.

CHAPTER I

Introduction

With categorical data becoming increasingly more ubiquitous nowadays, many statistical models have been developed to analyze this type of data. In this dissertation, we study and propose new statistical estimation and inference methods for categorical data, ranging from binary to multi-level count data, with a special focus on latent variable models and their applications to psychometrics. In large-scale psychometric assessments, binary data are often collected where a positive entry denotes a correct response by a subject to a test item while a zero entry denotes otherwise. An overall score is often far from sufficient to reflect the subject's underlying levels of aptitude. This is because items often have differing levels of difficulty and target on different skill sets, and furthermore, different groups of subjects are often assessed based on different sets of test items from a large item pool. Hence, direct comparison amongst subjects based on the overall score is likely to yield biased and unfair ranking results. Subjects' underlying levels of aptitude/proficiency is unobserved and can be modelled naturally by latent variables. This links naturally to the latent

variable models. In this dissertation, we study two types of latent variable models. Both assume categorical manifest variables and are widely applied to the psychometric assessments. The first type assumes that the subject's level of proficiency can be measured by multiple discrete latent attributes across a varying spectrum. While the second type uses a continuous latent score to capture the subject's aptitude. The first type of models is commonly known as the latent class models or the cognitive diagnosis models (CDMs) in the psychometric literature, a special class of discrete latent variable models, and are widely applied to educational assessments (*Junker and Sijtsma*, 2001; *von Davier*, 2008; *García et al.*, 2014), psychiatric diagnosis of mental disorders (*Templin and Henson*, 2006; *de la Torre et al.*, 2018), epidemiological and medical measurement studies (*Wu et al.*, 2016). While the second type of models falls in the category of item response theory (IRT) models. Popular IRT models include the one-parameter Rasch model (*Rasch*, 1960), the two-parameter logistic (2PL) model (*Birnbaum*, 1968), and three-parameter logistic model (3PL) (*Birnbaum*, 1968), amongst many others. In this dissertation, we will study the existing and propose new statistical estimation and inference methodologies for models that arise in these contexts.

Chapter II studies large-scale estimation problems in CDMs. The Q-matrix plays an important role in CDMs. It specifies the inter-dependence structure between the items and the subjects' discrete latent attributes. Knowing the Q-matrix accurately is crucial for valid cognitive diagnoses, item categorization and efficient assessment design. However, in practice, many existing important assessments do not provide accurate Q-matrix, not to mention many others do not even specify the Q-matrix.

Therefore, learning the Q-matrix from observational data is of paramount importance and has drawn great research attention so far. Various approaches have been proposed in the literature to estimate the Q-matrix (*de la Torre*, 2008; *DeCarlo*, 2012; *Liu et al.*, 2012; *Chiu*, 2013; *Chen et al.*, 2015; *de la Torre and Chiu*, 2016; *Xu and Shang*, 2018; *Chung and Johnson*, 2018; *Chen et al.*, 2018; *Culpepper*, 2019). However, existing methods suffer from computational hurdles and are not scalable with the size of the Q-matrix; they either break down or are extremely computationally expensive even when the size of the Q-matrix is moderately large. In practice, assessments with large Q-matrix is not uncommon, for which the existing estimation methods are not feasible. Such examples can be found in many applications, such as educational assessments (*Lee et al.*, 2011; *Choi et al.*, 2015; *González and Wiberg*, 2017) and the medical diagnosis of disease etiology (*Wu et al.*, 2016). Therefore, it remains an open and challenging problem in the literature to learn the large Q-matrix from observational data. In Chapter II, we propose a penalized likelihood approach to learn the large Q-matrix from observational data. More specifically, we identify the similarities between CDMs and the Restricted Boltzmann Machines (RBMs), and proposed a penalized RBM to learn the Q-matrix. As far as we know, our method is among the first ones in the literature that is scalable with the size of the Q-matrix and meanwhile retains high estimation accuracy. Furthermore, we demonstrate that the proposed method is robust to various types of CDMs and its estimation consistency of the Q-matrix is also established. The applicability and effectiveness of the proposed method is illustrated through an application to the TIMSS mathematics data.

Chapter III studies both large-scale learning and statistical inference problems for an IRT model, the Rasch model (*Rasch*, 1960), under the missing data settings. This links naturally to the model-based binary matrix completion problem. Data missingness is prevalent in large-scale assessments; examples include SAT and GRE where responses are collected and combined from multiple tests administered throughout the year from a large item pool. In assessment analysis, one primary goal is to compare/rank subjects based on their underlying levels of aptitude. Further compounded by differing levels of item difficulty, direct comparison based on subjects' overall scores often lead to biased and unfair results. Existing methods, though provide consistency guarantee for point estimates of the latent scores for each candidates, they cannot perform uncertainty quantification under this missing data setting. The latter is important because it helps determine the level of confidence we can have on the ranking/comparison. In Chapter III, we first provide point estimators for the subjects' latent scores and establish theoretical guarantee for their consistency. The remaining parts mainly focus on developing statistical inference theories under this missing data setting in double asymptotic regime. The intrinsic difficulty of the problem lies in the fact that the number of parameters in the model would tend to infinity together with the number of observations, so traditional inference results cannot be directly applied. The problem is further complicated by the nonlinear logistic link in the Rasch model and various missingness patterns present in the response matrix. The asymptotic distribution for any linear form of the proposed estimator for the latent scores is established which permits easy inference on the ranking/comparison. Moreover, a flexible missing-entry design, that

does not require a random sampling scheme, is adopted, which is required by most of the existing literature. The proposed estimator is statistically efficient and optimal, in the sense that the Cramer-Rao lower bound is achieved asymptotically for the model parameters. The results developed can be applied to a much wider horizon. In the real data applications, besides considering the motivating example of linking educational assessment results, we also demonstrate its applicability to linking the US senate voting where the senators' conservativeness scores are ranked even if the senators had not served any overlapping terms.

Measurement biases are not uncommon in psychometric assessments. In large-scale assessments, measurement biases can lead to unfair analysis results against certain gender, races and etc, and is commonly studied under the differential item functioning (DIF) in the literature. Chapter IV considers statistical estimation and inference for measurement bias under an IRT model, the 2PL model (*Birnbaum*, 1968). Measurement bias is brought by measurement non-invariant instruments (e.g. items) across different groups of subjects (*Millsap*, 2012). DIF analysis of item response data aims to detect the measurement non-invariant items (i.e. DIF items) so that biases can be removed by conditioning on the DIF effects. More precisely, a DIF item has a response distribution that depends not only on the latent score measured by the instruments but also the subjects' group membership. Therefore, detection of a DIF item involves comparing the item responses of different groups, conditioning on the latent scores. However, the detection and inference for DIF items remain open and challenging problems in the literature. The complexity of the problem lies in that the individuals' latent scores are not directly observable but are measured

by the instruments which may contain DIF. In addition, different groups may have different latent score distributions. Many approaches have been developed for DIF detection under the IRT framework (*Lord*, 1980; *Thissen*, 1988; *Raju*, 1988, 1990; *Kim et al.*, 1995; *Oort*, 1998; *Steenkamp and Baumgartner*, 1998; *Thissen*, 2001; *Cai et al.*, 2011; *Woods et al.*, 2013; *Tay et al.*, 2015, 2016; *Cao et al.*, 2017). The validity of traditional methods for DIF detection relies heavily on the prior knowledge about the anchor set, i.e. a set of DIF-free items, which is used to identify the latent trait distribution. However, in practice, such anchor set is difficult to locate, and the validity of the analysis results are highly sensitive to the correctness of the anchor set specification (*Kopf et al.*, 2015b). More recently, regularized estimation methods (*Magis et al.*, 2015; *Tutz and Schauberger*, 2015; *Huang*, 2018; *Belzak and Bauer*, 2020; *Bauer et al.*, 2020; *Schauberger and Mair*, 2020) were proposed to tackle this problem that do not require knowledge about anchor items. However, due to additional biases brought by the penalty, no direct/easy inference for DIF effects are permitted. In Chapter IV, we propose a novel minimal L_1 condition for simultaneous DIF detection and model identification. The proposed method can both accurately and computationally efficiently identify DIF items without requiring any prior knowledge on the anchor set. It can also perform statistical inference easily to quantify the uncertainties on the presence of DIF effects for each individual item, yielding valid confidence intervals or p-values. The point estimation and valid inference procedures lead to accurate detection of the DIF items, for which the type-I errors can be controlled by the inference results. The proposed methods can tackle both binary group DIF analysis and multiple group DIF analysis, while the latter is

less frequently considered in the literature but are very common in practice. In the end, we applied our method to PISA 2018 data to study the DIF effects brought by the high-dimensional country factors. Our study demonstrates that massive bias can occur if DIF effects are not taken into account.

Lastly, we consider privacy issues for releasing tabular (categorical) data to the public. Personal privacy protection becomes increasingly more urgent and important as tons of data are collected, transferred and released by different companies, agencies and institutions every day. Removing personal identities from the record alone is far from sufficient to protect individuals' privacy. Data adversaries may still be able to infer about an individual's identity, leading to privacy breach. Famous examples include the recovery of the anonymous location data (*Golle and Partridge*, 2009), privacy loss in genomic data (*Wang et al.*, 2009a) and the Washington State health record identification (*Sweeney*, 2013). Many statistical procedures to process data have been proposed under the data differential privacy (DP) framework (*Dwork et al.*, 2006b) to protect individuals' privacy (*Rubin*, 1993; *Little*, 1993; *Raghunathan et al.*, 2003; *Reiter*, 2005; *Dwork et al.*, 2010; *Drechsler*, 2011; *Mohammed et al.*, 2011; *Chaudhuri et al.*, 2012; *Yu et al.*, 2014; *Wang et al.*, 2015b; *Raab et al.*, 2016; *Friedman et al.*, 2016). However, most of these procedures are developed to target the release of summary statistics. But in practice, releasing the whole data set would be more preferred by the practitioners to perform any analyses desired. This is especially true for tabular (categorical) data where operations such as merging related cells or integrating statistical information obtained across different data sources can be easily performed. On the inference front, dedication to practically conducting

valid inferences on the altered privacy-protected data sets have only drawn attention very recently (*Sheffet*, 2017; *Degue and Le Ny*, 2018; *Ding et al.*, 2018; *Barrientos et al.*, 2019; *Couch et al.*, 2019; *Ferrando et al.*, 2022). Furthermore, no existing work on inference covers the simple but common operations amongst practitioners such as inter- (combine multiple tables) or intra- (combine interior categories within a table) table merging on private tabular data. In Chapter V, we recommend an optimal mechanism under the DP framework, focusing on releasing tabular data, such that an optimal balance between preserving data utility and satisfying privacy requirement can be achieved. Valid inference procedures on the privacy-protected data are developed, including those after performing inter- and intra- table merging operations. In the end, the proposed methods are applied to NCEDL's multi-state study of pre-kindergarten data set (*M. Clifford et al.*, 2017) to demonstrate its practical applicability.

CHAPTER II

Estimation of Large Q-Matrix Using Restricted Boltzmann Machines

2.1 Introduction

Cognitive Diagnosis Models (CDMs) are popular statistical tools widely applied to educational assessments and psychological diagnoses, which have been receiving increasingly more attention in the past two decades. In many modern assessment situations, examiners are concerned with specific attributes that the subjects possess, and thus a simple overall score is no longer sufficient to depict the whole picture of the subjects. As a result, a finer evaluation of the subjects' attributes is desired. CDMs are such tools. They model the relationship between the test items and the subjects' latent skills, which is helpful in assessment design and post-assessment analysis of the subjects' latent attribute patterns. CDMs have seen vast applications in multiple scientific disciplines, including educational assessments (*Junker and Sijtsma*, 2001; *von Davier*, 2008; *García et al.*, 2014), psychiatric diagnosis of mental disorders

(*Templin and Henson*, 2006; *de la Torre et al.*, 2018), epidemiological and medical measurement studies (*Wu et al.*, 2016).

Many CDMs can be viewed as restricted latent class models that directly model the response probabilities as functions of discrete latent attributes. A common goal of cognitive diagnoses is to learn the subjects' latent attributes, such as personalities or skills, based on their responses to a combination of specially designed test items. The Q-matrix plays a critical role in CDMs. It specifies the dependency structure between the test items and the latent attributes. Knowing the Q-matrix accurately is important because it is indispensable to cognitive diagnoses. Besides, the Q-matrix itself can be used to categorize the test items and enable efficient design of future assessments. However, in reality, many existing assessments do not even have the Q-matrix explicitly specified. Even the assessment providers specify the Q-matrix when designing the assessment, the specification may still be inaccurate. In many cases, one test item may potentially be linked to multiple attributes, but usually only the most direct and apparent ones are identified in the pre-designed Q-matrix. Therefore, it is of paramount importance to develop methodologies to efficiently learn the Q-matrix from the observational responses.

Various approaches have been proposed in the literature to learn the Q-matrix. Those methods can be generally classified into two categories, validation of the existing Q-matrix (*de la Torre*, 2008; *DeCarlo*, 2012; *Chiu*, 2013; *de la Torre and Chiu*, 2016) and direct estimation of the Q-matrix from the observational data (*Liu et al.*, 2012; *Chen et al.*, 2015; *Xu and Shang*, 2018; *Chung and Johnson*, 2018; *Chen et al.*, 2018; *Culpepper*, 2019). However, most of the existing estimation methods for the

whole Q-matrix in general suffer from huge computational cost and are not scalable with the size of the Q-matrix; they either break down or are extremely computationally expensive even when the Q-matrix is moderately large. The high computational cost stems from the large number of configurations of the Q-matrix. If we view each binary element of the Q-matrix as a unique parameter, then the number of different configurations would grow exponentially with the size of the Q-matrix. In many applications, the number of latent attributes being tested is large, leading to a high-dimensional space for all possible latent attribute patterns. It is not uncommon that the number of potential attribute patterns is large, sometimes even larger than the sample size, making the estimation even more difficult. Such examples can be found in many applications, such as educational assessments (*Lee et al.*, 2011; *Choi et al.*, 2015) and the medical diagnosis of disease etiology (*Wu et al.*, 2016); for instance, Section 2.5 presents a dataset from the Trends in International Mathematics and Science Study (TIMSS), which has 13 binary latent attributes and $2^{13} = 8192$ attribute patterns while only 757 subjects. On the other hand, the number of items being tested may also be large in many applications. One example is the TIMSS mathematical test which often have more than 100 test items. Another example is the ADM admissions test, which is given twice a year and is used as an entrance test to universities and colleges, contains a total of 200 items (*González and Wiberg*, 2017). Therefore, it remains an open and challenging problem to learn the large Q-matrix from the observational data.

Borrowing the idea from the deep learning literature, we propose to use the restricted Bolzmann machines (RBMs) to learn the large Q-matrix. An RBM is a

generative two-layer neural network that can learn a probability distribution over a collection of inputs (*Smolensky*, 1986). Amongst these inputs, some are observed variables while the others are latent variables that we do not observe, which matches the restricted latent class CDM setting. The weight matrix $\boldsymbol{W}$ in RBMs determines the relationship between the observed variables and the latent variables. By learning this weight matrix $\boldsymbol{W}$ under the framework of RBMs, we show that the structure of the Q-matrix in CDMs can be inferred accordingly. Although this is similar to the maximum likelihood learning approach, by tapping on RBMs, fast learning of the large Q-matrix can be achieved.

Our main contributions are that we identify the relationships between CDMs and RBMs, and proposed a new way of learning the large Q-matrix efficiently. As far as we know, our proposed method is among the first ones in the literature that is scalable with the size of the Q-matrix (with computational cost of $O(J \times K)$) while at the same time retains high estimation accuracy. For example, comparing to *Xu and Shang* (2018) which attains an estimation accuracy of 71.2% in the GDINA setting with five independent latent attributes using 2000 observations, our method achieves more than 86% overall accuracy and much faster computational speed. Another interesting finding is that learning of the Q-matrix by RBMs is robust to different CDMs, including the DINA, ACDM and GDINA models. We provide theoretical guarantees under certain conditions and conduct simulation studies to support our findings. Besides, because of the unsupervised learning nature of RBMs, the traditional cross-validation (CV) procedure are not directly applicable. As a result, we also present a new CV procedure specifically to the Q-matrix learning setting.

The remaining parts of the chapter are organized as follows. Section 2.2 gives a review on CDMs and RBMs, and discussion of their relationships and why the learning of the Q-matrix by RBMs is achieveable across different CDMs. Section 2.3 introduces our proposed estimation method and the new CV procedure. Section 2.4 consists of simulation studies on three typical CDMs. Section 2.5 demonstrates the performance of our proposed method through the data analysis on a TIMSS mathematics data set. Section 2.6 concludes with discussions and potential future directions. All the proofs and additional simulation results can be found in Appendix A. The materials of this chapter are mainly based on *Li et al.* (2022a).

2.2 Estimation of Q-matrix Using RBMs

2.2.1 Review of CDMs

Many CDMs have been developed in recent decades, among which the Deterministic Input Noisy output "And" gate model (DINA, *Haertel*, 1989; *Junker and Sijtsma*, 2001) is one of the most popular and simple models and serves as the foundation for many complex CDMs. Other popularly used CDMs include the Noisy Input Deterministic "And" gate model (NIDA, *Junker and Sijtsma*, 2001), the Reduced Reparametrized Unified Model (R-RUM, *Hartz*, 2002), the General Diagnostic Model (GDM, *von Davier*, 2005), the Deterministic Input Noisy "Or" gate (DINO, *Templin and Henson*, 2006), the Log linear CDM (LCDM, *Henson et al.*, 2008), the Additive CDM (ACDM, *de la Torre*, 2011) and the Generalized DINA model (GDINA, *de la Torre*, 2011).

Consider a CDM with J items and K latent attributes. There are two types of variables for each subject: the observed responses for J items $\mathbf{Y} = (Y_1, ..., Y_J)$ and the latent attribute pattern $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$, which are both assumed to be binary. $Y_j \in \{1, 0\}$ denotes whether the subject answers item j correctly and $\alpha_k \in \{1, 0\}$ denotes possession or non-possession of the attribute k. The Q-matrix, $\boldsymbol{Q} = (q_{j,k}) \in \{0, 1\}^{J \times K}$, specifies the dependence structure between the items and the latent attributes; $q_{j,k} \in \{1, 0\}$ denotes whether a correct response to item j requires the latent attribute k. If we denote the jth row of the Q-matrix to be $\mathbf{q}_j$, then $\mathbf{q}_j$ reflects the full attribute requirements of item j. For a latent attribute pattern $\boldsymbol{\alpha}$, we say $\boldsymbol{\alpha}$ possesses all the required attributes of item j if $\boldsymbol{\alpha} \succeq \mathbf{q}_j$, where $\boldsymbol{\alpha} \succeq \mathbf{q}_j$ means $\alpha_k \geq q_{j,k}$ for all $k = 1, ..., K$. Different CDMs model the item response functions $P(Y_j = 1 \mid \boldsymbol{\alpha})$ differently with the item parameters constrained by the Q-matrix and specific cognitive diagnostic assumptions. Below we introduce three popular CDMs that will be considered in later discussions.

Example II.1 (DINA model)**.** Let $Y_{i,j} \in \{1, 0\}$ denote whether subject i answers item j correctly. Under the DINA model (*Haertel*, 1989; *Junker and Sijtsma*, 2001), for the jth item and the ith subject with the latent attribute pattern $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,K})$, the ideal response variable is defined as $\xi_{i,j} = \prod_{k:q_{j,k}=1} \alpha_{i,k} = \prod_{k=1}^{K} \alpha_{i,k}^{q_{j,k}}$. The ideal response $\xi_{i,j} = 1$ only if $\boldsymbol{\alpha}_i \succeq \boldsymbol{q}_j$, that is, the subject i needs to possess all the latent attributes required by the item j to have a positive ideal response. The uncertainty is further incorporated by two parameters: the slipping parameter s_j and the guessing parameter g_j. Specifically, $s_j = P(Y_{i,j} = 0 \mid \xi_{i,j} = 1)$ and $g_j = P(Y_{i,j} = 1 \mid \xi_{i,j} = 0)$. The slipping parameter and the guessing param-

eter further satisfy $1 - s > g$, which indicates that the capable subjects will have higher positive probability than the incapable ones. The DINA model is one of the most restrictive and interpretable CDMs for dichotomously scored test items. It is a parsimonious model that requires only two parameters for each item regardless of the number of attributes required for the item. It is appropriate when the tasks call for the conjunction of several equally important attributes, and lacking one required attribute for the item is the same as lacking all the required attributes.

Example II.2 (ACDM)**.** In the ACDM, mastering additional required attributes will increase the positive response probability for the items. Specifically, if we take the identity link function in the ACDM, then for the jth item and the ith subject with attribute pattern $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,K})$, we have

$$P(Y_{i,j} = 1 \mid \boldsymbol{\alpha}_i) = \delta_{j,0} + \sum_{k=1}^{K} \delta_{j,k}\alpha_{i,k}q_{j,k}, \tag{2.1}$$

which implies that mastering the kth attribute increases the probability of success on the item j by $\delta_{j,k}$ if the kth latent attribute is required by the item j. Since there is no interaction term in (2.1), the contribution of each latent attribute is independent from one another. If subject i lacks all the required attributes for item j, the term $\sum_{k=1}^{K} \delta_{j,k}\alpha_{i,k}q_{j,k}$ would be 0, and the intercept $\delta_{j,0}$ represents the probability of correctly answering the item j based on pure guessing. Furthermore, even if the ith subject has all the required latent attributes of the item j, $\delta_{j,0} + \sum_{k=1}^{K} \delta_{j,k}\alpha_{i,k}q_{j,k}$ may not sum to 1. In that case, $1 - \left(\delta_{j,0} + \sum_{k=1}^{K} \delta_{j,k}\alpha_{i,k}q_{j,k}\right)$ captures the probability of making a careless mistake. The ACDM is more appropriate to use when the items call

for independent latent attributes but with different contributions to correct response to the items.

Besides the identity link function, other link functions are also proposed. One commonly used link function is the logit link,

$$P(Y_{i,j} = 1 \mid \boldsymbol{\alpha}_i) = \sigma\Big(\delta_{j,0} + \sum_{k=1}^{K} \delta_{j,k}\alpha_{i,k}q_{j,k}\Big). \tag{2.2}$$

where $\sigma(x) = \{1 + \exp(-x)\}^{-1}$. Equation (2.2) is also equivalent to $\text{logit}\big\{P(Y_{i,j} = 1 \mid \boldsymbol{\alpha}_i)\big\} = \delta_{j,0} + \sum_{k=1}^{K} \delta_{j,k}\alpha_{i,k}q_{j,k}$, which is the log-odds of a positive response. The interpretation would then become that each required latent attribute contributes independently to the log-odds of correcting answering item j by $\delta_{j,k}$ in an additive fashion.

Example II.3 (GDINA model)**.** Both the DINA and ACDM models are special cases of the more general GDINA model (*de la Torre*, 2011). In addition to the intercept and the main effects in the ACDMs, the GDINA model also allows interactions amongst the latent attributes. The equation (2.3) gives the item response function for the GDINA model with identity link.

$$\begin{aligned} P(Y_{i,j} = 1 \mid \boldsymbol{\alpha}_i) =& \delta_{j,0} + \sum_{k=1}^{K} \delta_{j,k}\alpha_{i,k}q_{j,k} + \sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} \delta_{jkk'}\alpha_{i,k}\alpha_{i,k'}q_{j,k}q_{j,k'} + ... \\ &+ \delta_{j12...K}\prod_{k=1}^{K} \alpha_{i,k}q_{j,k}. \end{aligned} \tag{2.3}$$

The parameters in (2.3) can be interpreted as follows: δ_0 is the probability of a correct response when none of the required attributes is present; δ_k is the change

in the probability of a correct response when only mastering a single attribute α_k; $\delta_{kk'}$, a first-order interaction effect, is the change in the response probability due to the possession of both α_k and $\alpha_{k'}$ in addition to the main effects of mastering the two individual attributes; and $\delta_{12\ldots K}$ represents the change in the probability of a correct response due to the mastery of all the required attributes in addition to the main effects and all the lower-order interaction effects. Similarly to the ACDM model, $P(Y_{i,j} = 1 \mid \boldsymbol{\alpha}_i)$ is not required to be 1 even when the subject i possesses all the required attribute for the item j. In that case, $1 - P(Y_{i,j} = 1 \mid \boldsymbol{\alpha}_i)$ is the probability of making a careless mistake. Moreover, the intercept $\delta_{j,0}$ and the main effects are typically non-negative, but the interaction effects can take on any values. Therefore, the GDINA model is appropriate if the mixed effects of latent attributes on the probability of a correct response is of interest.

2.2.2 Review of Restricted Boltzmann Machines

RBMs are generative models that can learn probabilistic distributions over a collection of inputs. RBMs were initially introduced under the name Harmonium by *Smolensky* (1986) and gained currency due to their fast learnability in the mid-2000. It has found vast applications in dimension reduction (*Hinton and Salakhutdinov*, 2006), classification (*Larochelle and Bengio*, 2008), collaborative filtering (*Salakhutdinov et al.*, 2007) and many other fields.

RBMs can also be viewed as a probabilistic bipartite graphical models, with observed (visible) units in one part of the graph and latent (hidden) units in the other part. Typically all the hidden units and the visible units are binary. We denote

the visible units by $\boldsymbol{Y} = \{Y_1, ...Y_J\} \in \{0,1\}^J$ and hidden units by $\boldsymbol{\alpha} = \{\alpha_1, ...\alpha_K\} \in \{0,1\}^K$ respectively. One key feature of RBMs is that only interactions between hidden units and visible units are allowed. Neither connections among the visible units, nor any connections among the hidden units are allowed, as shown in Figure 2.1.

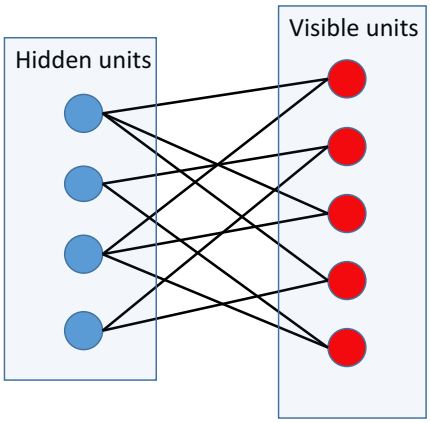

Figure 2.1: A graphical illustration of RBM.

RBMs are characterized by the energy functions with the joint probability distribution specified as

$$P(\boldsymbol{Y}, \boldsymbol{\alpha}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\big\{ - E(\boldsymbol{Y}, \boldsymbol{\alpha}; \boldsymbol{\theta})\big\}, \tag{2.4}$$

where $E(\boldsymbol{Y}, \boldsymbol{\alpha}; \boldsymbol{\theta})$ is known as the energy function and $Z(\boldsymbol{\theta})$ is the partition function,

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{Y} \in \{0,1\}^J} \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} \exp\big\{ - E(\boldsymbol{Y}, \boldsymbol{\alpha}; \boldsymbol{\theta})\big\},$$

which is intractable (*Long and Servedio*, 2010). In specific, the energy function is given by

$$\begin{aligned} E(\boldsymbol{Y}, \boldsymbol{\alpha}; \boldsymbol{\theta}) &= -\boldsymbol{b}^T \boldsymbol{Y} - \boldsymbol{c}^T \boldsymbol{\alpha} - \boldsymbol{Y}^T \boldsymbol{W} \boldsymbol{\alpha} \\ &= -\sum_{j=1}^{J} Y_j b_j - \sum_{k=1}^{K} \alpha_k c_k - \sum_{j=1}^{J} \sum_{k=1}^{K} Y_j w_{j,k} \alpha_k, \end{aligned} \tag{2.5}$$

where $\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{W}\}$ are the model parameters, $\boldsymbol{b} \in \mathbb{R}^J$ are visible biases, $\boldsymbol{c} \in \mathbb{R}^K$ are hidden biases and $\boldsymbol{W} \in \mathbb{R}^{J \times K}$ is the weight matrix encoding the interactions between the visible and the hidden units.

Since no "$\boldsymbol{Y}$-$\boldsymbol{Y}$" or "$\boldsymbol{\alpha}$-$\boldsymbol{\alpha}$" interactions are allowed, the hidden and visible units are conditionally independent given each other, and therefore the joint conditional probability mass functions can be factored in to a product. This can be easily seen from Equations (2.4) and (2.5). Specifically, we have

$$P\big(\boldsymbol{Y} \mid \boldsymbol{\alpha}; \boldsymbol{\theta}\big) = \prod_{j=1}^{J} P\big(Y_j \mid \boldsymbol{\alpha}; \boldsymbol{b}, \boldsymbol{W}\big), \tag{2.6}$$

$$P\big(Y_j = 1 \mid \boldsymbol{\alpha}; \boldsymbol{b}, \boldsymbol{W}\big) = \sigma\Big(b_j + \sum_{k=1}^{K} w_{j,k} \alpha_k\Big), \tag{2.7}$$

and

$$P\big(\boldsymbol{\alpha} \mid \boldsymbol{Y}; \boldsymbol{\theta}\big) = \prod_{k=1}^{K} P\big(\alpha_k \mid \boldsymbol{Y}; \boldsymbol{c}, \boldsymbol{W}\big), \tag{2.8}$$

$$P\big(\alpha_k = 1 \mid \boldsymbol{Y}; \boldsymbol{c}, \boldsymbol{W}\big) = \sigma\Big(c_k + \sum_{j=1}^{J} w_{j,k} Y_j\Big), \tag{2.9}$$

where $\sigma(x) = 1/\{1 + \exp(-x)\}$ is the logistic sigmoid function.

RBMs and CDMs are in fact closely related. The binary observed item responses and the latent attributes in CDMs can be viewed as counterparts to the visible units and the hidden units in RBMs respectively. There is a direct connection between the two. If we fit an ACDM with the logit link, where the conditional probability mass function (2.2) of the observed responses is modeled as a sigmoid function of the latent attributes, then it takes exactly the same form as the conditional probability function (2.7) of a visible unit given the hidden units in RBMs. Moreover, in a CDM, $q_{j,k} = 0$ indicates that there is no interaction between the item j and the latent attribute k, while in the weight matrix of an RBM, $w_{j,k} = 0$ also implies no interaction between the jth visible unit and the kth hidden unit. Therefore we would expect that $w_{j,k} = 0$ in an RBM whenever $q_{j,k} = 0$ in a CDM.

Using the previous example in Figure 2.1 for illustration, on the left of (2.10) is the weight matrix $\boldsymbol{W}$ of an RBM, where $w_{j,k} \neq 0$ indicates the presence of the interaction between the visible unit Y_j and the hidden unit α_k. The corresponding Q-matrix in a CDM can be implied as shown on the right. As we illustrate previously, the non-zero entries in the Q-matrix of an ACDM can be inferred exactly from the non-zero entries in the weight matrix $\boldsymbol{W}$ in an RBM. Interactions among the latent attributes are allowed in the DINA and GDINA models, which violates the assumptions of an RBM. However, the Q-matrix is still estimable in these models. We will give detailed arguments in Section 2.2.3.

$$W = \begin{bmatrix} w_{11} & 0 & w_{13} & 0 \\ 0 & w_{22} & 0 & w_{24} \\ w_{31} & 0 & w_{33} & 0 \\ w_{41} & 0 & 0 & w_{44} \\ 0 & w_{52} & w_{53} & 0 \end{bmatrix} \quad \Longrightarrow \quad Q = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{2.10}$$

2.2.3 Robust Estimation of Q-matrix

In the previous section, we have discussed that RBMs can be used to learn the Q-matrix for the ACDM with logit link. A natural question to ask is whether we can generalize this result to other CDMs such as the DINA and GDINA models. In this section, we will illustrate that under certain conditions, robust estimation of the Q-matrix by RBMs is indeed achievable for common CDMs. In particular, we will demonstrate that the Q-matrix can be estimated correctly under the DINA and GDINA settings.

We focus on the learning of a particular row of the Q-matrix. It is in fact a variable selection problem of the required latent variables for that particular item of interest. Conditional on $\boldsymbol{\alpha}$, we have discussed that RBMs are equivalent to the ACDM with the logit link, while the latter exactly corresponds to the logistic regression with canonical link and additive main effects linear predictor. Therefore in essence, RBMs can also be treated as main effect models. Starting with the simplest case, we shall first study the model selection consistency with linear additive models when the true models are the DINA or the GDINA model. Since it is still an open and challenging problem to establish consistent variable selection under complex latent variable models, here

we start with the ideal case by assuming $\{\alpha_1, ..., \alpha_K\}$ are independent, that is, all the latent variables are independent. Although this is a strong assumption and is rarely fully satisfied in real world scenarios, it can be relaxed in practice which will be discussed in Remark II.7.

Before giving formal statements, we first introduce some notations. Without loss of generality, we focus on the analysis of the response to one single item. For a subject with $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_K\}$, the response to the considered item is denoted by Y, where for clarity, we omit the item index in the notation. Let K^* to be the number of required attributes for the item. Without loss of generality, we let the first K^* attributes be the required attributes for this item, that is, the corresponding row in the Q-matrix is $\mathbf{q} = (1, ..., 1, 0, ..., 0)$ with the first K^* entries being 1 and all the remaining $K - K^*$ entries being 0. For the response Y generated from the DINA or the GDINA model, we denote $\mathbb{E}^*[Y \mid \boldsymbol{\alpha}]$ as the regression mean function for the mis-specified linear regression model of Y on $\alpha_1, ..., \alpha_K$. We show in the following propositions that the mis-specified mean function $\mathbb{E}^*[Y \mid \boldsymbol{\alpha}]$ can identify the required attributes from the non-required ones.

Proposition II.4 (DINA model)**.** *Assume* $\{\alpha_1, \alpha_2, ..., \alpha_K\}$ *are independent with* $\alpha_k \sim$ *Beroulli*(p_k) *where* $p_k \in (0, 1)$*,* $k = 1, 2, ..., K$*. If* Y *is generated from the DINA model, then the mis-specified linear additive model of* Y *regressed on* $(\alpha_1, \alpha_2, ..., \alpha_K)$ *has the mean function in the form of* $\mathbb{E}^*[Y \mid \boldsymbol{\alpha}] = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + ... + \beta_K\alpha_K$ *with* $\beta_l \neq 0$ *for* $l = 1, 2, ..., K^*$ *and* $\beta_k = 0$ *for* $k = K^* + 1, ..., K$*.*

Proposition II.4 states that under the independence condition and if the data is generated from the DINA model, the significant variables included in the true model

can be selected correctly using a mis-specified linear model with additive main effects only.

Proposition II.5 (GDINA model)**.** *Assume* $\{\alpha_1, \alpha_2, ..., \alpha_K\}$ *are independent with* $\alpha_k \sim$ *Bernoulli*(p_k) *where* $p_k \in (0,1)$*,* $k = 1, 2, ...K$*. If* Y *is generated from the GDINA model satisfying the monotonicity assumption (i.e. acquiring an additional required skill* α_k*,* $k = 1, 2, .., K^*$*, will always increase the probability of a correct response), then the mis-specified linear additive model has the corresponding mean function in the form of* $\mathbb{E}^*[Y \mid \boldsymbol{\alpha}] = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + ... + \beta_K\alpha_K$ *with* $\beta_l \neq 0$ *for* $l = 1, 2, ..., K^*$ *and* $\beta_k = 0$ *for* $k = K^* + 1, ..., K$*.*

Similar to Proposition II.4, Proposition II.5 states that under suitable conditions, the significant variables included in the true GDINA model can be selected correctly using a mis-specified linear model with additive main effects only. The detailed proofs for all the propositions can be found in Appendix A.

Propositions II.4 and II.5 demonstrate that the model selection consistency can be achieved using a mis-specified linear main effect model. As we illustrated previously, the conditional probability of a visible unit on the hidden units in RBMs can be regarded as a main effect logistic regression model. Therefore we next give some intuition on why the main effect logistic regression model will give a similar variable selection result to the linear models. Consider a main effect logistic regression model with the canonical link function, that is, $\text{logit}\big(P(Y \mid \boldsymbol{\alpha})\big) = \beta_0 + \beta_1\alpha_1 + ... + \beta_K\alpha_K$. Let $\mathcal{Y} = (Y_i, i = 1, ..., N)$ denote the response vector for all the N subjects, and let $\boldsymbol{\mu} = \big\{\mu_i := P(Y_i \mid \boldsymbol{\alpha}_i), i = 1, ..., N\big\}$ denote the response probabilities for the subjects. We use $\boldsymbol{A} = \big(\boldsymbol{\alpha}_i\big)_{i=1}^N \in \{0,1\}^{N \times K}$ to denote the latent attribute matrix for

the N subjects and $\boldsymbol{A}^*$ to denote the $N \times (K+1)$ matrix $[\mathbf{1}; \boldsymbol{A}]$ with the first column being an all-one vector. In linear models, we usually use the least square estimation to estimate the coefficients, while in logistic regression, the iteratively re-weighted least square (IRLS) method is used. Next we will give some intuition on why these two estimation methods will produce similar variable selection results.

Conditional on $\boldsymbol{\alpha}_i$'s, in the $(t+1)$th step of IRLS, the updating rule for parameter $\boldsymbol{\theta} := (\beta_0, \beta_1, ..., \beta_K)$ is

$$\boldsymbol{\theta}^{(t+1)} = \left(\boldsymbol{A}^{*T}\boldsymbol{W}^{(t)}\boldsymbol{A}^*\right)^{-1}\boldsymbol{A}^{*T}\boldsymbol{W}^{(t)}\boldsymbol{Z}^{(t)},$$

where $\boldsymbol{Z}^{(t)} = \boldsymbol{A}^{*T}\boldsymbol{\theta}^{(t)} + (\boldsymbol{W}^{(t)})^{-1}(\mathcal{Y} - \boldsymbol{\mu}^{(t)})$ is the tth step working response and $\boldsymbol{W}^{(t)} = \text{diag}\big(\mu_1^{(t)}(1-\mu_1^{(t)}), ..., \mu_N^{(t)}(1-\mu_N^{(t)})\big)$ is a diagonal weight matrix with diagonal elements being the variance estimates for each Y_i. Since there is no closed form of IRLS estimator and there is randomness in the convergence process, it is very challenging to study the theoretical properties of the $\boldsymbol{\theta}$ estimated by IRLS. So we only consider a one-step update of IRLS starting from the ideal case of true parameter $\boldsymbol{\theta}_{\text{true}}$ for illustration. It is reasonable to study this ideal case because IRLS will converge close to the $\boldsymbol{\theta}_{\text{true}}$ given the correct model specification and a large sample size. If we start with the true parameters, that is, we let $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}_{\text{true}}$, then,

$$\boldsymbol{\theta}^{(1)} = \left(\boldsymbol{A}^{*T}\boldsymbol{W}_{\text{true}}\boldsymbol{A}^*\right)^{-1}\boldsymbol{A}^{*T}\boldsymbol{W}_{\text{true}}\boldsymbol{Z}_{\text{true}},$$

where the working response, $\boldsymbol{Z}_{\text{true}} = \boldsymbol{A}^{*T}\boldsymbol{\theta}_{\text{true}} + \boldsymbol{W}_{\text{true}}^{-1}(\mathcal{Y} - \boldsymbol{\mu}_{\text{true}})$ is just a linear transformation of observed response $\mathcal{Y}$. Note that this update takes the same form

as the weighted least square estimation of regressing $\boldsymbol{Z}_{true}$ on $\boldsymbol{A}^*$. Hence, the variable selection result in the linear model would be similar to that of the logistic regression. Combining Propositions II.4 and II.5, we have justified that the learning of the Q-matrix by RBMs is achievable across the DINA, ACDM and GDINA models with both identity and logit links.

Remark II.6. In practice, it is not uncommon that some of the 2^K latent attribute patterns do not exist in the collected observations, especially when K is large. How negatively will this impact on the model selection consistency? In the DINA model, we see from the proof of Proposition II.4 (see Section A.2 in Appendix A) that to ensure the variable selection consistency for each required attribute α_k, $k = 1, ..., K^*$, we need to observe data from subjects with $\{\boldsymbol{\alpha} \mid \alpha_k = 0, \alpha_i = 1, i = 1, \ldots, k-1, k+1, \ldots, K^*\}$and $\{\boldsymbol{\alpha} \mid \alpha_i = 1, i = 1, \ldots, K^*\}$. In the GDINA model, from the proof of Proposition II.5, we can see that to ensure the variable selection consistency for each α_k, $k = 1, ..., K^*$, we need to observe data from subjects with $\{\boldsymbol{\alpha} \mid \alpha_k = 0\}$ and $\{\boldsymbol{\alpha} \mid \alpha_k = 1\}$. Therefore, even though some of the latent patterns may not exist in our observed data, the selection consistency is still achievable as long as the required attribute patterns are present.

Remark II.7. The independence assumption on the latent attributes $\{\alpha_1, ..., \alpha_K\}$ can be relaxed to some extent in practice. To see this, consider the setting when $\{\alpha_1, ..., \alpha_K\}$ are possibly dependent but the response Y only directly depends on the first K^* attributes $\{\alpha_1, ..., \alpha_{K^*}\}$. Given $\alpha_1, ..., \alpha_{K^*}$, the response Y is conditionally independent of α_k for all $k = K^* + 1, ..., K$. When only $\alpha_1, ..., \alpha_{K^*}$ are present in the linear regression model of Y regressed on α's, consider adding in one additional

α_k, for any $k = K^* + 1, ..., K$, into the regression model, then its coefficient can be expressed as

$$\beta_k = \frac{Cov\Big(Y - \mathbb{E}^*[Y \mid \alpha_1, ..., \alpha_{K^*}], \quad \alpha_k - \mathbb{E}^*[\alpha_k \mid \alpha_1, ..., \alpha_{K^*}]\Big)}{Var\Big(Y - \mathbb{E}^*[Y \mid \alpha_1, ..., \alpha_{K^*}]\Big)}, \tag{2.11}$$

where we denote $\mathbb{E}^*[A \mid B]$ as the regression mean function of A on B. Since Y and α_k are conditionally independent given $\alpha_1, ..., \alpha_{K^*}$, the numerator of (2.11) is expected to be small. In real implementations, the shrinkage imposed by the L_1 penalty in our proposed method should be able to recover most of these 0's. This is indeed supported by our simulation results in Section 2.4, where we consider moderate to high correlation regimes amongst the latent attributes and our proposed method still achieves satisfactory estimation accuracy of the underlying Q-matrix. Note also that in the special case when $K^* = 1$, the covariance term in (2.11) can be shown exactly equal to zero, in which case β_k can be removed easily in the variable selection process. For a more detailed discussion for the $K^* = 1$ case, please refer to Section A.2 in Appendix A.

Remark II.8. The rigorous consistency theory of using RBMs to learn the Q-matrix under a general CDM setting can be difficult to establish. In the literature, even when the true models are binary RBMs, consistency for training RBMs is an open and challenging problem. Due to the intractable partition function in the binary RBM, an approximate likelihood maximizing approach has to be employed, such as the popularly used Contrastive Divergence (CD) algorithm that will be further intro-

duced in Section 2.3. Even though there are many works in literature studying the asymptotic properties of the CD algorithm (*MacKay*, 2001; *Yuille*, 2004; *Carreira-Perpinan and Hinton*, 2005; *Bengio and Delalleau*, 2009; *Sutskever and Tieleman*, 2010; *Jiang et al.*, 2018), whether and why the CD algorithm provides an asymptotically consistent estimate for binary RBMs are still open questions. Therefore, establishing a consistency theorem using a mis-specified RBM model for the DINA or the GDINA model as in this work is even more challenging, which is left for future exploration. Nevertheless, the CD algorithm in practice has shown empirical success in training RBMs, and our simulation results in Section 2.4 also demonstrate its effectiveness in training RBMs to learn the Q-matrix in CDMs.

2.3 Proposed Estimation Method

In this section, we will introduce our proposed method in detail. As we have illustrated in Section 2.2, non-zero entries in the Q-matrix can be inferred from the corresponding non-zero entries in the weight matrix of RBMs. Therefore, we are interested in a sparse solution of the weight matrix $\boldsymbol{W}$. It is well known that L_1 penalty has the property of producing sparse solutions (*Rosasco*, 2009). Hence, we propose the following L_1 penalized likelihood as our objective function,

$$\min_{\boldsymbol{\theta}} -\log\left\{P(\boldsymbol{Y};\boldsymbol{\theta})\right\} + \lambda \sum_{j=1}^{J}\sum_{k=1}^{K} |w_{j,k}|. \tag{2.12}$$

where $\log\{P(\boldsymbol{Y};\boldsymbol{\theta})\}$ is the marginal log-likelihood of the observed responses $\boldsymbol{Y}$, $\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{W}\}$ are the model parameters, and λ is a non-negative tuning parameter for

the L_1 penalty.

Gradient descent algorithm is a standard numerical method to solve problem (2.12). The likelihood part, following the derivation by *Schlueter* (2014), can be shown that its gradient with respect to the parameters has the following decomposition:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \log\big(P(\boldsymbol{Y};\boldsymbol{\theta})\big) = &- \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} P\big(\boldsymbol{\alpha}|\boldsymbol{Y};\boldsymbol{\theta}\big)\frac{\partial}{\partial \boldsymbol{\theta}} E\big(\boldsymbol{Y},\boldsymbol{\alpha};\boldsymbol{\theta}\big) \\ &+ \sum_{\boldsymbol{y}\in\{0,1\}^J \boldsymbol{\alpha}\in\{0,1\}^K} P\big(\boldsymbol{y},\boldsymbol{\alpha};\boldsymbol{\theta}\big)\frac{\partial}{\partial \boldsymbol{\theta}} E\big(\boldsymbol{y},\boldsymbol{\alpha};\boldsymbol{\theta}\big) \qquad (2.13) \\ =&\mathbb{E}_{P(\boldsymbol{\alpha}|\boldsymbol{Y};\boldsymbol{\theta})}\Big[-\frac{\partial}{\partial \boldsymbol{\theta}} E\big(\boldsymbol{Y},\boldsymbol{\alpha};\boldsymbol{\theta}\big)\Big] - \mathbb{E}_{P(\boldsymbol{y},\boldsymbol{\alpha};\boldsymbol{\theta})}\Big[-\frac{\partial}{\partial \boldsymbol{\theta}} E\big(\boldsymbol{y},\boldsymbol{\alpha};\boldsymbol{\theta}\big)\Big]. \qquad (2.14)\end{aligned}$$

In deep learning literature, this is a well-known decomposition into the positive phase and the negative phase of learning, corresponding to the two expectations in (2.14) respectively. As the two expectations do not have closed forms and are not directly tractable, researchers propose to approximate the gradient by estimating these expectations through Monte Carlo sampling. In particular, the positive phase corresponds to sampling the hidden units given the visible units, while the negative phase corresponds to obtaining the joint hidden and visible samples from the current model.

The bipartite graph structure of RBMs gives the special property of its conditional distributions $P(\boldsymbol{\alpha} \mid \boldsymbol{Y})$ and $P(\boldsymbol{Y} \mid \boldsymbol{\alpha})$ being factorial and simple to compute and sample from, as shown in Section 2.2.2. Therefore, sampling for the positive phase

is straightforward while obtaining samples from the model for negative phase is not since it requires the joint hidden and visible samples. A widely used algorithm to learn RBMs is known as the Contrastive Divergence (CD) algorithm, where the negative phase is approximated by drawing samples from a short alternating Gibbs Markov chain between visible units and hidden units starting from the observed training examples (*Hinton*, 2002). In this work, we use a CD-1 algorithm where Gibbs chains are run for 1 step to approximate the gradient of the log-likelihood part. Specifically, given the original data $\boldsymbol{Y}^{(0)}$, we first sample $\boldsymbol{\alpha}^{(0)}$ according to Equation (2.8) and Equation (2.9) to approximate the positive phase. Then given $\boldsymbol{\alpha}^{(0)}$, we sample $\boldsymbol{Y}^{(1)}$ based on Equation (2.6) and Equation (2.7), and we use $(\boldsymbol{Y}^{(1)}, \boldsymbol{\alpha}^{(0)})$ to approximate the negative phase.

At $(t+1)$th iteration, based on the sampled data, the parameters' updates take the same form as gradient descent if we do not consider L_1 penalty,

$$w'^{(t+1)}_{j,k} \leftarrow w^{(t)}_{j,k} + \gamma^{(t)} \Big\{ \sum_{i=1}^{N} Y^{(0)}_{ij} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}^{(0)}_i; \boldsymbol{\theta}^{(t)}\big) - \sum_{i=1}^{N} Y^{(1)}_{ij} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}^{(1)}_i; \boldsymbol{\theta}^{(t)}\big) \Big\}, \tag{2.15}$$

$$b^{(t+1)}_j \leftarrow b^{(t)}_j + \gamma^{(t)} \Big\{ \sum_{i=1}^{N} Y^{(0)}_{i,j} - \sum_{i=1}^{N} Y^{(1)}_{i,j} \Big\} / N, \tag{2.16}$$

$$c^{(t+1)}_k \leftarrow c^{(t)}_k + \gamma^{(t)} \Big\{ \sum_{i=1}^{N} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}^{(0)}_i; \boldsymbol{\theta}^{(t)}\big) - \sum_{i=1}^{N} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}^{(1)}_i; \boldsymbol{\theta}^{(t)}\big) \Big\} / N, \tag{2.17}$$

where $\boldsymbol{Y}^{(0)}_i = (Y^{(0)}_{i1}, Y^{(0)}_{i2}, \ldots, Y^{(0)}_{iJ})$, $\boldsymbol{Y}^{(1)}_i = (Y^{(1)}_{i1}, Y^{(1)}_{i2}, \ldots, Y^{(1)}_{iJ})$, and $\gamma^{(t)}$ is the learning rate for the tth iteration. Here we denote the updated weight matrix by

$\boldsymbol{W}' = \left(W'_{j,k}\right)_{J\times K}$, since we also need to consider the gradient of the L_1 penalty term later, and thus Equation (2.15) is an intermediate update for the weight matrix. Detailed derivations can be found in the notes written by *Schlueter* (2014). In this work, we use a linearly decreasing learning rate scheme, which is guaranteed to converge as shown in *Collins et al.* (2008).

For the L_1 penalty, we adopt the implementation developed by *Tsuruoka et al.* (2009), which can achieve more stable sparsity structures. As pointed out by *Tsuruoka et al.* (2009), the traditional implementation of L_1 penalty in gradient descent algorithm does not always lead to sparse models because the approximate gradient used at each update is very noisy, which deviates the updates away from zero.

The main idea of the implementation is to keep track of the total penalty and the penalty that has been applied to each parameter, and then the L_1 penalty is applied based on the difference between these cumulative values. By doing so, it is argued that the effect of noisy gradient is smoothed away. To be more specific, at iteration t, let $u^{(t)} := \lambda \sum_{l=1}^{t} \gamma^{(l)}$ be the absolute value of the total L_1 penalty that each parameter could have received up to the point, where $\gamma^{(l)}$ is the learning rate at step l. Let $c_{j,k}^{(t-1)} := \sum_{l=1}^{t-1}(w_{j,k}^{(l+1)} - w'^{(l+1)}_{j,k})$ be the total L_1 penalty that $w_{j,k}$ has actually received up to step t, where $w'^{(l)}_{j,k}$ is the intermediate update at step l calculated by Equation (2.15). Then at iteration $(t+1)$, we update $w_{j,k}^{(t+1)}$ by

$$w_{j,k}^{(t+1)} \leftarrow \max\left\{0, w'^{(t+1)}_{j,k} - (u^{(t)} + c_{j,k}^{(t-1)})\right\} \quad \text{if} \quad w'^{(t+1)}_{j,k} > 0,$$

$$w_{j,k}^{(t+1)} \leftarrow \min\left\{0, w'^{(t+1)}_{j,k} + (u^{(t)} - c_{j,k}^{(t-1)})\right\} \quad \text{if} \quad w'^{(t+1)}_{j,k} \leq 0.$$

Since the updates in Equations (2.15), (2.16) and (2.17) require summations over all the data samples, it would be computationally expensive when the sample size is large. To reduce computational burden, we implement a batch version of the CD-1 algorithm in practice, where we only use a small batch of the whole data set in each iteration. Specifically, we randomly partition the whole data set into B batches, and iterating through all the batches is known as one epoch in machine learning literature. Here we use $\boldsymbol{Y} = \left\{\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}, \ldots, \boldsymbol{Y}_{(B)}\right\}$ to denote the partitions, N_B to denote the batch size, and N_{epoch} to denote the number of epoches. The resulting algorithm is summarized in Algorithm 1.

Algorithm 1: CD-1 algorithm with L_1 penalty

Input: Data $\boldsymbol{Y} = \left\{\boldsymbol{Y}_{(1)}, \boldsymbol{Y}_{(2)}, \ldots, \boldsymbol{Y}_{(B)}\right\}$, λ, γ_0, and N_{epoch}.

Output: Estimates $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{c}}$.

Initialize $w_{j,k}^{(0)}, b_j^{(0)}, c_k^{(0)}$, $u^{(0)} = 0$, $c_{j,k}^{(0)} = c_{j,k}^{(1)} = 0$;

for $e = 0, \ldots, N_{epoch} - 1$ **do**

  for $b = 0, \ldots, B - 1$ **do**

    $t = e \times B + b$ (the number of iterations);

    $\gamma^{(t)} = \frac{\gamma_0}{t+1}$;

    $\boldsymbol{Y}^{(0)} \leftarrow \boldsymbol{Y}_{(b+1)}$;

    Sample $\boldsymbol{\alpha}^{(0)} \sim P\big(\boldsymbol{\alpha} \mid \boldsymbol{Y}^{(0)}; \boldsymbol{c}^{(t)}, \boldsymbol{W}^{(t)}\big)$;

    Sample $\boldsymbol{Y}^{(1)} \sim P\big(\boldsymbol{Y} \mid \boldsymbol{\alpha}^{(0)}; \boldsymbol{b}^{(t)}, \boldsymbol{W}^{(t)}\big)$;

    $u^{(t)} \leftarrow u^{(t-1)} + \lambda\gamma^{(t)}$;

    for $j = 1, \ldots, J, k = 1, \ldots, K$ **do**

      $w'^{(t+1)}_{j,k} \leftarrow w_{j,k}^{(t)} + \gamma^{(t)}\Big\{ \sum_{i=1}^{N_B} Y_{ij}^{(0)} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}_i^{(0)}\big) - \sum_{i=1}^{N_B} Y_{ij}^{(1)} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}_i^{(1)}\big)\Big\}$;

      if $t \geq 2$ **then** $c_{j,k}^{(t-1)} \leftarrow c_{j,k}^{(t-2)} + w_{jk}^{(t)} - w'^{(t)}_{j,k}$;

      if $w'^{(t+1)}_{j,k} > 0$ **then**

        $w_{j,k}^{(t+1)} \leftarrow \max\Big\{0, w'^{(t+1)}_{j,k} - (u^{(t)} + c_{j,k}^{(t-1)})\Big\}$;

      else

        $w_{j,k}^{(t+1)} \leftarrow \min\Big\{0, w'^{(t+1)}_{j,k} + (u^{(t)} - c_{j,k}^{(t-1)})\Big\}$;

    end

  end

  for $j = 1, ..., J$ **do**

    $b_j^{(t+1)} \leftarrow b_j^{(t)} + \gamma^{(t)}\Big\{ \sum_{i=1}^{N_B} Y_{i,j}^{(0)} - \sum_{i=1}^{N_B} Y_{i,j}^{(1)}\Big\}/N_B$;

  end

  for $k = 1, ..., K$ **do**

    $c_k^{(t+1)} \leftarrow c_k^{(t)} + \gamma^{(t)}\Big\{ \sum_{i=1}^{N_B} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}_i^{(0)}\big) - \sum_{i=1}^{N_B} P\big(\alpha_{ik} = 1 \mid \boldsymbol{Y}_i^{(1)}\big)\Big\}/N_B$;

  end

end

In our proposed algorithm, there are two tuning parameters: λ for the L_1 penalty and γ_0 for the learning rate. To get good estimates of our model, we need to select a suitable combination of hyper-parameters λ and γ_0. A popularly used tuning procedure is CV. However, as RBMs are unsupervised learning models, we cannot rely on the so-called "test error" of the labels. Instead, since visible units are re-sampled at each iteration in the CD algorithm, we may use the reconstruction error of the visible units to assess the goodness of fit. Nevertheless, the visible reconstruction error will always increase as the penalty coefficient λ increases, because larger penalty would introduce more bias. Therefore, the traditional CV procedures would not work here. To solve this problem, given values of λ and γ_0, instead of directly using the $\hat{\boldsymbol{W}}_{\lambda,\gamma_0}$ obtained from a penalized RBM to compute the reconstruction error, we propose to debias the non-zero entries in $\hat{\boldsymbol{W}}_{\lambda,\gamma_0}$ by training an RBM with no penalty but fixing the zero positions the same as those in $\hat{\boldsymbol{W}}_{\lambda,\gamma_0}$. The proposed CV producedure is summarized below.

1. Split the data into M partitions. Each time we use one partition as the validation set and the remaining as the training set.

2. Apply the penalized CD Algorithm 1 to train the RBM on the training set with pre-specified λ and γ_0, and obtain the estimates $\hat{\boldsymbol{W}}_{\lambda,\gamma_0}$ and $\hat{\boldsymbol{Q}}_{\lambda,\gamma_0}$.

3. Use the training set again to debias the non-zero entries of $\hat{\boldsymbol{W}}_{\lambda,\gamma_0}$. Specifically, we use $\hat{\boldsymbol{W}}_{\lambda,\gamma_0}$ as the initial value and set $\lambda = 0$ in Algorithm 1 to train an unpenalized RBM, and only update the non-zero entries of $\hat{\boldsymbol{W}}_{\lambda,\gamma_0}$ while keeping the zero entries unchanged. Hidden bias $\boldsymbol{c}$ and visible bias $\boldsymbol{b}$ are updated at

each step as usual. This step give us the de-biased weight matrix $\check{\boldsymbol{W}}_{\lambda,\gamma_0}$.

4. Compute the reconstruction error on the validation set. In specific, at each iteration of the CD algorithm, we fix $\boldsymbol{W} = \check{\boldsymbol{W}}_{\lambda,\gamma_0}$, and only update the hidden and visible biases. The reconstruction error is computed as the mean batch squared error between the latest sampled visible batches $\{\boldsymbol{Y}_1^{(1)}, ..., \boldsymbol{Y}_m^{(1)}\}$ and the observational batches $\{\boldsymbol{Y}_1^{(0)}, ..., \boldsymbol{Y}_m^{(0)}\}$ in the validation set.

5. For each combination of λ and γ_0 in the subject set, we repeat Steps 2-4 across all M validation sets. The $\hat{\boldsymbol{Q}}_{\lambda^*,\gamma_0^*}$ corresponding to the smallest mean batch squared error (see Section 2.4 for definition) is taken as the final estimate of the Q-matrix.

Another main difference from the traditional CV procedure is we select the Q-matrix corresponding to the smallest validation error instead of taking average of the validation errors and then training a new RBM with the best tuning parameters according to the smallest mean error. There are two advantages. On one hand, the traditional way of averaging errors, though more stable, is very time-consuming in this problem. On the other hand, the gradient descent steps in the CD algorithm may only produce locally optimal results. To avoid being stuck in sub-optima, we run the CD algorithm M times with different initializations and different training and validation sets for each combination of λ and γ_0, and select the estimated Q-matrix corresponding to the smallest validation error. By doing so, the Q-matrix is expected to be more accurately estimated.

Remark II.9. The computational cost of our proposed method only grows linearly

in K and this enables estimation of very large Q-matrices. As far as we know, the current methods in the literature have computational cost greater than $O(K)$, with the majority growing exponentially with K. For example, in *Xu and Shang* (2018), they proposed to learn the Q-matrix by estimating the coefficients in the LCDM plus a penalty term with the EM algorithm. In the E-step of the EM algorithm, 2^K posterior probabilities for each of the attribute patterns need to be updated.

2.4 Simulation Studies

We conduct simulation studies on three popular CDMs, the DINA, ACDM and GDINA models, to study the performance of our proposed method in learning the Q-matrix under different CDM settings. In particular, we examine the scalability to the size of the Q-matrix and the estimation accuracy of the proposed algorithm.

We first introduce the metrics used to evaluate the performance of the proposed estimation method. To measure the convergence of the algorithm, we investigate the change in the mean batch errors against time. The mean batch error is the reconstruction error between the latest sampled visible batches $\left\{\boldsymbol{Y}_{(1)}^{(1)}, ..., \boldsymbol{Y}_{(B)}^{(1)}\right\}$ and the original observed batches $\left\{\boldsymbol{Y}_{(1)}^{(0)}, ..., \boldsymbol{Y}_{(B)}^{(0)}\right\}$, where $\left\{\boldsymbol{Y}_{(1)}^{(0)}, ..., \boldsymbol{Y}_{(B)}^{(0)}\right\}$ partitions the whole observed data set into B batches. Given the batch-size N_B, the mean batch error is defined as

$$\frac{1}{BN_B}\sum_{b=1}^{B}\sum_{i=1}^{N_B}\sum_{j=1}^{J}\left(Y_{(b),i,j}^{(1)} - Y_{(b),i,j}^{(0)}\right)^2.$$

To evaluate the estimation accuracy, we report entry-wise overall percentage error (OE), out of true positives percentage error (OTP) and out of true negatives per-

centage error (OTN). Specifically,

$$\text{OE} := \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{1}\{\hat{q}_{j,k} \neq q_{j,k}\},$$

which is the percentage of wrongly estimated entries out of the total number of entries in the Q-matrix.

$$\text{OTP} := \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{1}\{\hat{q}_{j,k} = 0, q_{j,k} = 1\}}{\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{1}\{q_{j,k} = 1\}},$$

which is defined as the percentage of wrongly estimated entries out of all true positive entries (i.e. entries 1) in the Q-matrix.

$$\text{OTN} := \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{1}\{\hat{q}_{j,k} = 1, q_{j,k} = 0\}}{\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbb{1}\{q_{j,k} = 0\}},$$

which is defined as the percentage of wrongly estimated entries out of all true negatives (i.e. entries 0) in the Q-matrix. A challenge in computing these errors arises because the estimated Q-matrix can only be identified up to column permutations. To solve this problem, we apply the Hungarian algorithm to match the columns of the estimated $\hat{Q}$ to the true Q-matrix by jointly minimizing the total column-wise matching errors. Details of the Hungarian algorithm can be found in *Kuhn* (1955).

We consider different number of latent attributes $K = 5, 10, 15, 20, 25$. To ensure the Q-matrix is identifiable so that it can be learned from the observational data, we

specify it as follows:

$$Q = \begin{bmatrix} I_K \\ Q_1 \\ Q_2 \end{bmatrix}, \tag{2.18}$$

where I_K is a K dimensional identity matrix; $Q_1 \in \{0,1\}^{K\times K}$ with value 1 in the (i,i)th entries for $i = 1, ..., K$ and the $(i, i+1)$th entries for $i = 1, ..., K-1$, and values 0 for all the other entries; $Q_2 \in \{0,1\}^{K\times K}$ with value 1 in entries (i,i) for $i = 1, ..., K$, $(i, i-1)$ for $i = 2, ..., K$ and $(i, i+1)$ for $i = 1, ..., K-1$, and value 0 for all the remaining entries. The above construction sets the number of items to be $J = 3K$. This Q-matrix satisfies the identifiability conditions in *Gu and Xu* (2019) and therefore is identifiable under the DINA setting in Simulation Study 2.4.1. Moreover, this construction also ensures the (generic) identifiability of the ACDM and GDINA models considered in Simulation Studies 2.4.2 and 2.4.3 (see *Xu*, 2017; *Gu and Xu*, 2020b,a). A random design of the Q-matrix, in which its identifiability is not be guaranteed, is also considered in Section A.1.1 of Appendix A.

In each simulation study, we consider two different sample sizes $N = 2000$ or 10000. Both independent and dependent settings of latent attributes are explored. Denote the latent attribute matrix by $\boldsymbol{A} = \left(\boldsymbol{\alpha}_i\right)_{i=1}^{N} \in \{0,1\}^{N\times K}$, which depicts the latent attribute patterns of the N subjects. We use two steps to simulate the latent patterns (*Chen et al.*, 2015). First, a Gaussian latent vector is generated for each subject $\boldsymbol{z}_i = (z_{i1}, ..., z_{iK}) \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$ for $i = 1, ..., N$, where $\Sigma = (1-\rho)\mathbf{1}_K + \rho\mathbf{1}_K\mathbf{1}_K^\top$, $\mathbf{1}_K = (1, \ldots, 1)_K^\top$, and ρ is the correlation between any two different latent attributes.

In practice, since some attributes may be harder to master than others, different thresholds are applied in the sampling of attribute profiles. In particular, for a given K, we specify the thresholds ranging from -0.5 to 0.5, with a step size of $1/(K-1)$, for each of attribute $1, 2, ...K$ respectively. Then $\alpha_{ik} = 1$ if z_{ik} is greater than its respective threshold and $\alpha_{ik} = 0$ otherwise. For the independent setting, we set $\rho = 0$, while for the dependent settings, we consider both a low correlation with $\rho = 0.25$ and a high correlation with $\rho = 0.75$. For the tuning of hyper-parameters, we take the subject sets as $\lambda \in \{0.003, 0.004, ..., 0.015\}$ and $\gamma_0 \in \{0.5, 1, ..., 5.5\}$, and perform 5-fold CV to select the best estimated Q-matrix. For each setting, 100 repetitions are simulated. The batch size and the number of epochs are fixed at 50 and 300 respectively.

2.4.1 Simulation Study 1. DINA Model

For the DINA test items, we consider two uncertainty levels, $g_j = s_j = 0.1$ or $g_j = s_j = 0.2$ for all $j = 1, ..., J$. Figure 2.2 plots the mean batch errors against time for the independent case with $K = 5$ (the first row) and $K = 25$ (the second row) across different sample sizes and different noise levels. When $K = 5$, we can see that the CD-1 algorithm converge well after 6 seconds for all different sample sizes under different noise levels. This suggests that with a small number of latent attributes, the sample sizes and the uncertainty levels do not affect the convergence speed a lot. Focusing on the second row of Figure 2.2, we note that although the size of the Q-matrix increases from 75 ($K = 5$) to 1875 ($K = 25$), the convergence time only increases by around 10 seconds, and the CD-1 algorithm converges well

after just 15 seconds even when $K = 25$. This indicates that the proposed method is scalable with the size of the Q-matrix. Dependent settings have similar convergence rates and hence the results are omitted.

Figure 2.3 and 2.4 plot different estimation errors against the sizes of the Q-matrix for independent and dependent settings respectively. For the independent case, in Figure 2.3, we can see that the OE stays below 16% across all the settings. There is a decreasing trend in the OE as the Q-matrix size increases due to the increasing sparsity of the true underlying Q-matrix. Our proposed method performs significantly better than the baseline method predicting all the entries of the Q-matrix to be 0 (which would produce OE of 36% for $K = 5$). Furthermore, we note that increasing uncertainty level will deteriorate the OTP, making the estimation of positive entries harder. Increasing the sample size N would in general help improve the estimation accuracy. For the dependent case, in Figure 2.4, we observe that the results in the low correlation setting are very similar to that of the independent setting. This suggests that our proposed method is robust when moderate correlations amongst latent attributes exist. On the other hand, when the correlations amongst the attributes are high, we see increments in all the three error metrics, OE, OTP and OTN. The correlations amongst the attributes would compound the difficulty in estimation of the Q-matrix. However, all the OE's still stay well below 20%. Hence, our proposed method can still achieve effective learning of the Q-matrix when the correlations amongst the attributes are high.

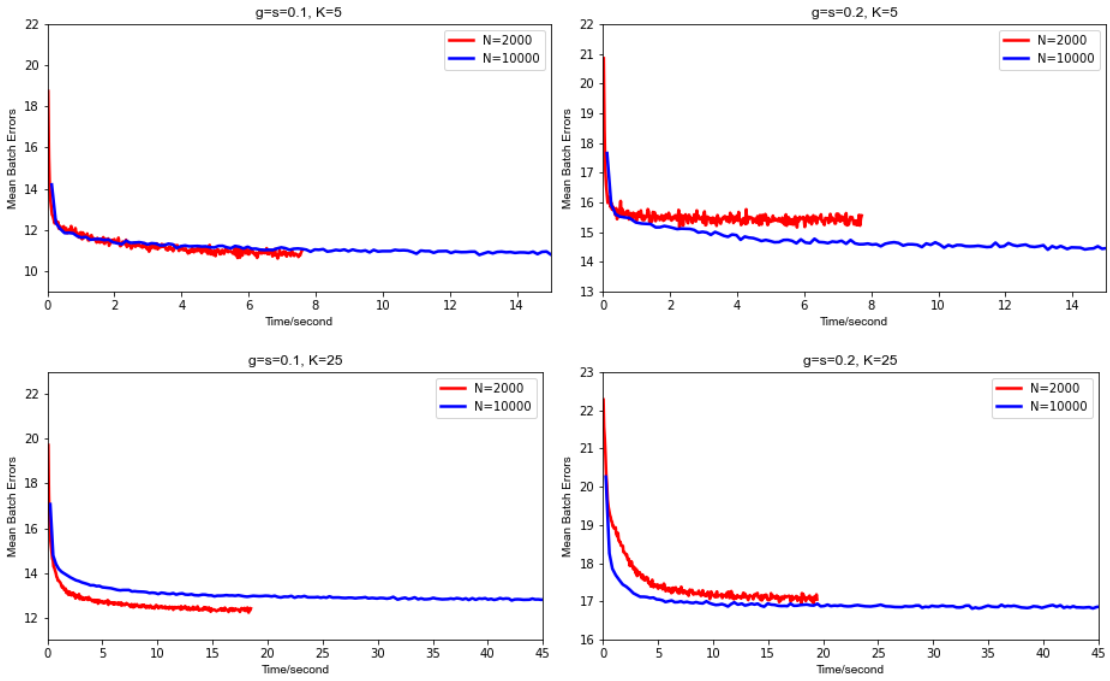

Figure 2.2: Plots of mean batch errors against time for the DINA data.

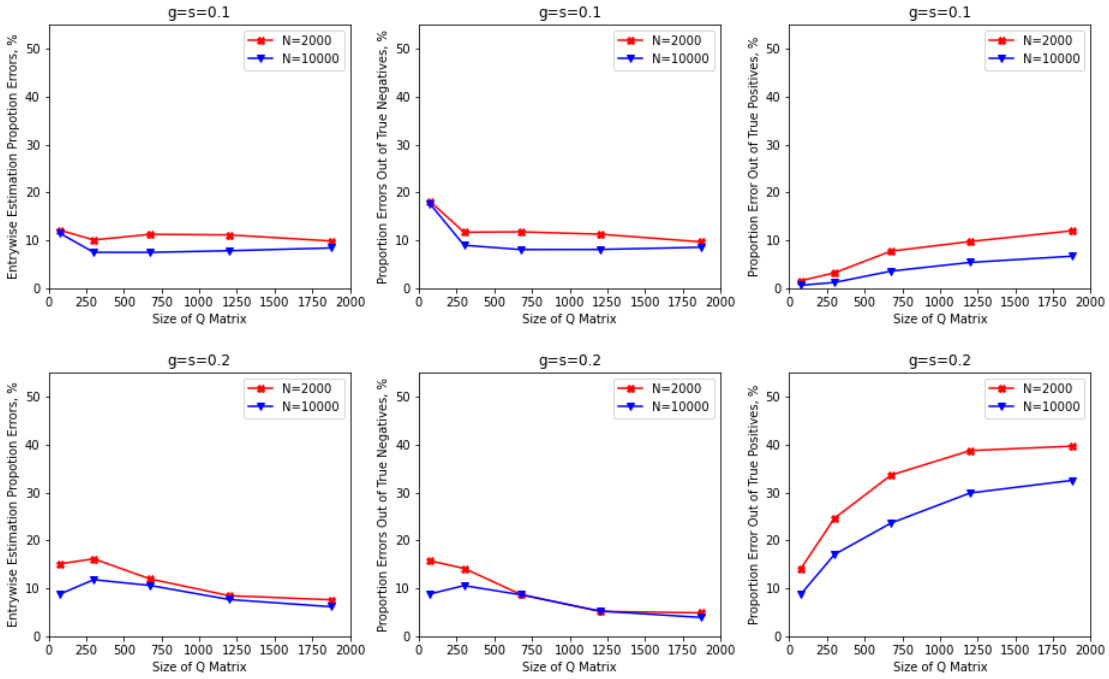

Figure 2.3: Plots of different performance metrics against the size of the Q-matrix for the DINA data (independent case).

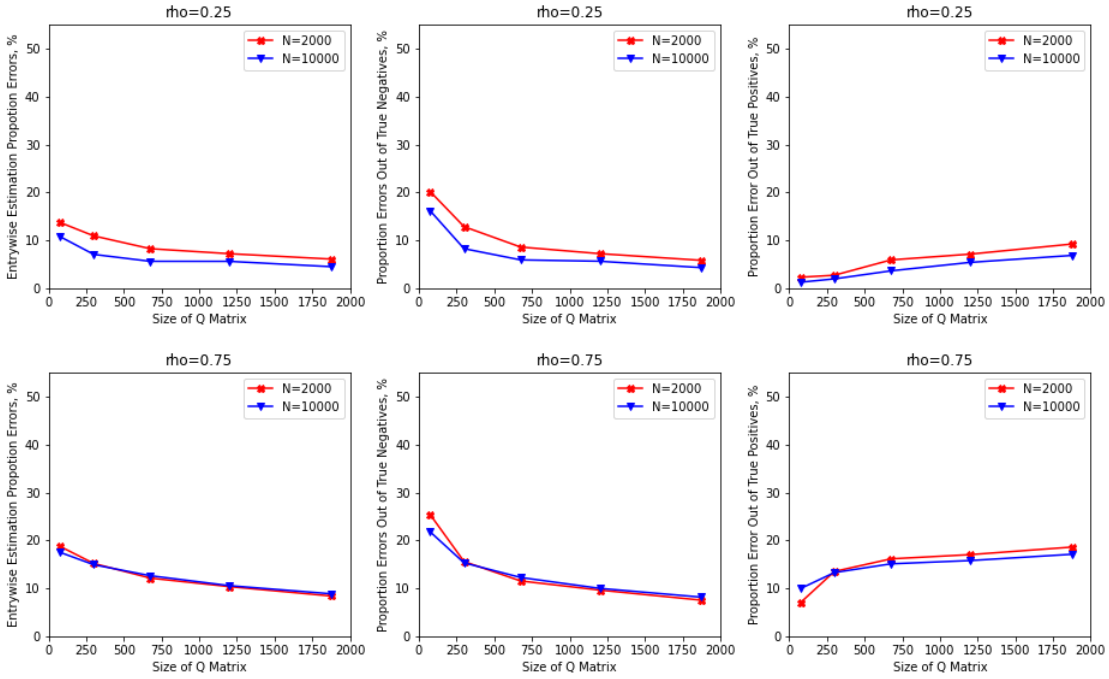

Figure 2.4: Plots of different performance metrics against the size of the Q-matrix for the DINA data (dependent case with $g = s = 0.1$). Rows 1 and 2 correspond to correlation settings 0.25 and 0.75 respectively.

2.4.2 Simulation Study 2. ACDM Model

We conduct similar analysis using data generated from the ACDM to examine the convergence speed and estimation accuracy of our proposed method. Define K_j^* to be the number of required attributes for the item j. Without loss of generality, we let the first K_j^* attributes be the required attributes for item j, i.e., the corresponding row in the Q-matrix is $\boldsymbol{q}_j = (1, ..., 1, 0, ..., 0)$ with the first K_j^* entries being 1 and all the remaining $K - K_j^*$ entries being 0. For an ACDM with the identity link function 2.1, we have $P(Y_j = 1 \mid \mathbf{1}_K) = \delta_{j,0} + \sum_{k=1}^{K_j^*} \delta_{j,k} := p_j$, the highest success probability achievable for the most capable subjects. Similar to the DINA setting, two different

uncertainty levels are considered: case 1. $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$ and case 2. $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$. For $k = 1, ..., K_j^*$, $\delta_{j,k}$ is set to be $(p_j - \delta_{j,0})/K_j^*$, that is, the contribution of each required attribute to the success probability is equal.

Figure 2.5 shows the convergence speed of our proposed method under the independent setting. We observe similar patterns as in the DINA case: uncertainty levels and samples sizes do not have significant impacts on the convergence speed. Our proposed algorithm is scalable with the size of the Q-matrix in the ACDM setting. Figure 2.6 and 2.7 plot different estimation metrics against the size of the Q-matrix for independent and dependent settings respectively. From Figure 2.6, we can see that the results are very similar to those observed in the DINA model setting, which demonstrates that our proposed methods is effective in the ACDM data. Furthermore, for the dependent setting in Figure 2.7, we observe that when the correlation is of 0.25, the estimation accuracy remains similar to that in the independent settings. When the correlation is of 0.75, unlike in the DINA setting, the OE, OTP and OTN only increase very slightly. In particular, the OE stays well below 16.5% when $K = 5, 10, ..., 25$. This suggests that when the true data generating model is the ACDM, our proposed method is robust when the correlations amongst the attributes are high.

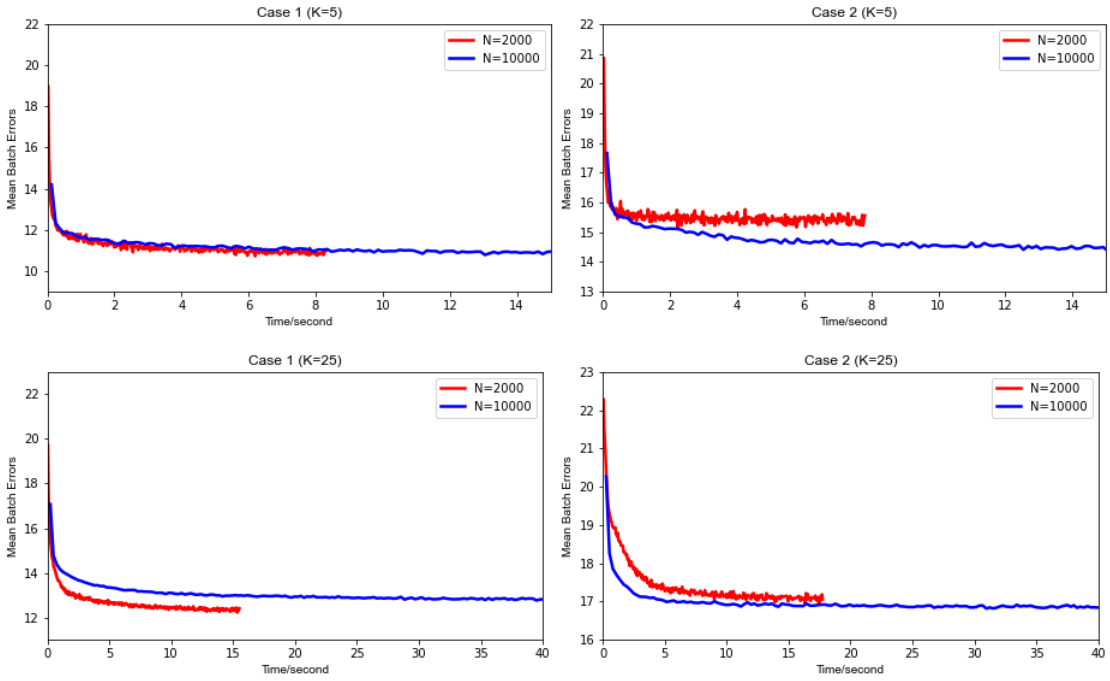

Figure 2.5: Plots of mean batch errors against the time for the ACDM data. Case 1 represents the setting when $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$. Case 2 represents the setting when $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$.

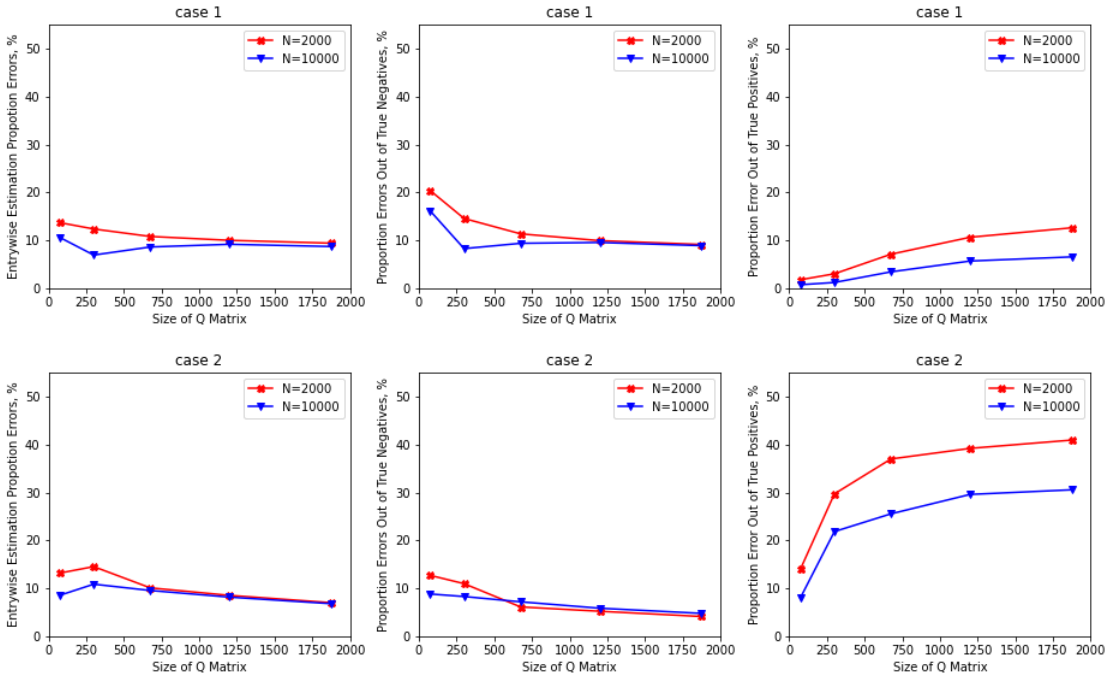

Figure 2.6: Plots of different performance metrics against the size of the Q-matrix for the ACDM data (independent case). Case 1 represents the setting when $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$. Case 2 represents the setting when $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$.

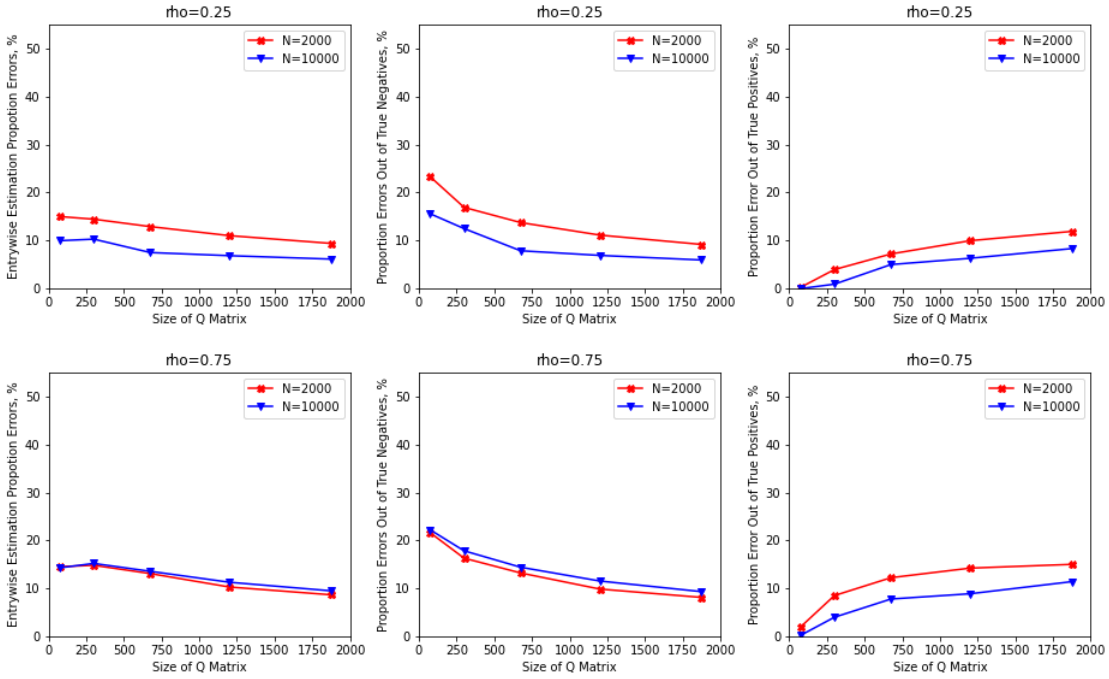

Figure 2.7: Plots of different performance metrics against the size of the Q-matrix for the ACDM data (dependent case with $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$). Rows 1 and 2 correspond to correlation settings 0.25 and 0.75 respectively.

2.4.3 Simulation Study 3. GDINA Model

Let the highest success probability achievable for the most capable subjects be $P(Y_j = 1 \mid \mathbf{1}_K) := p_j$ from Equation (2.3). Similar to the ACDM setting, we consider two uncertainty levels: case 1. $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$ and case 2. $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$. Using the Q-matrix specified at the beginning of this section, for each item j, we may have $K_j^* = 1, 2$ or 3. When $K_j^* = 1$, we set $\delta_{j,k} = p_j - \delta_{j,0}$. When $K_j^* = 2$, we let $\delta_{j,k} = \delta_{jkk'} = (p_j - \delta_{j,0})/3$ and when $K_j^* = 3$ we set $\delta_{j,k} = \delta_{jkk'} = \delta_{jkk'k''} = (p_j - \delta_{j,0})/7$. As such, the main effects and the interaction terms are all assumed to have the same contributions to the probability of a positive

response. Both independent and dependent settings are considered.

Convergence rates under independent setting are summarized in Figure 2.8. Similar patterns to the DINA and the ACDM settings can be observed, indicating that our algorithm is scalable to the size of the Q-matrix in the GDINA model. As before, dependent settings have similar convergence patterns, and hence the results are not presented here. Behaviors of different estimation metrics over the size of the Q-matrix for both the independent and dependent settings are summarized in Figure 2.9 and 2.10 respectively.

For the independent setting in Figure 2.9, slightly better estimation accuracy can be observed than in the DINA and the ACDM settings. This suggests our proposed methods is effective in the learning the Q-matrix from data generated using the GDINA model. One thing to emphasize is that our method is competitive amongst the existing algorithms in the literature. For example, comparing to a similar simulation study in *Xu and Shang* (2018) for $K = 5$ independent attributes and $N = 2000$, our overall estimation accuracy of around 87% is significantly better than theirs, whose overall accuracy is 71.2%. Moreover, our method also has much smaller computational cost than their method. For the dependent setting in Figure 2.10, we observe that the estimation accuracy remains similar to the independent setting when the correlation is of 0.25. When the correlations increase to 0.75, all the three error metrics only increase very slightly. This observation is similar to the ACDM setting. The OE's remain well below 16.5% for all $K = 5, 10, ..., 25$. This suggests that when the true data generating model is the GDINA model, the proposed method is fairly robust to high attribute correlations.

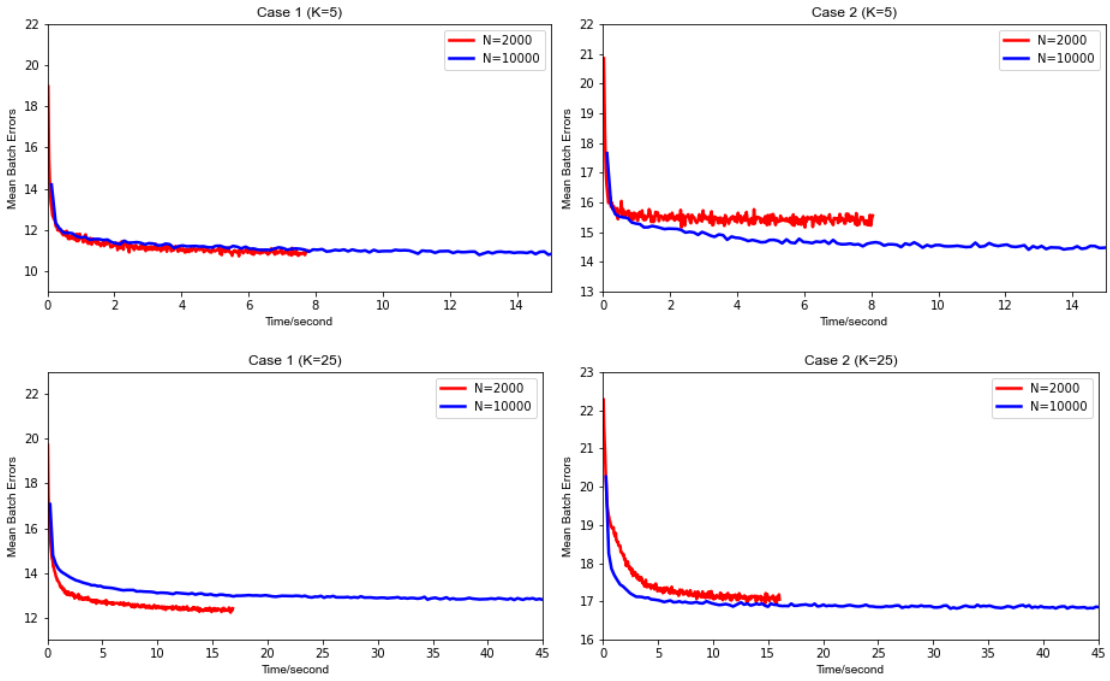

Figure 2.8: Plots of mean batch errors against the size of the Q-matrix for the GDINA data. Case 1 represents the setting when $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$. Case 2 represents the setting with higher uncertainty levels when $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$.

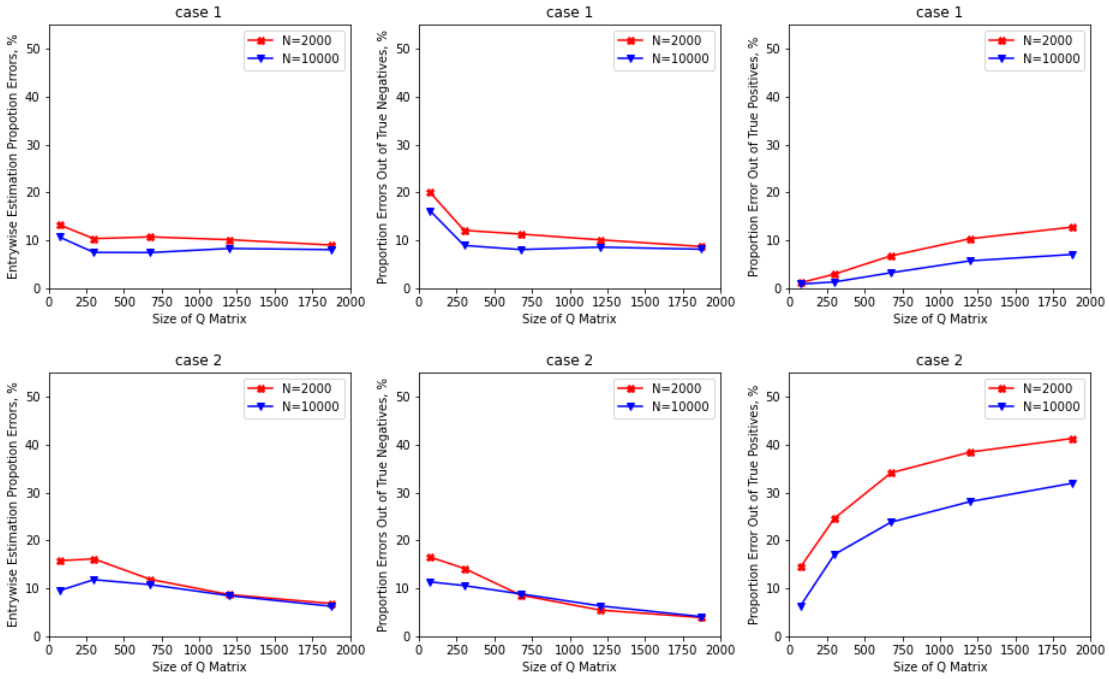

Figure 2.9: Plots of different performance metrics against the size of the Q-matrix for the GDINA data (independent case). Case 1 represents the setting when $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$. Case 2 represents the setting with higher uncertainty levels when $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$.

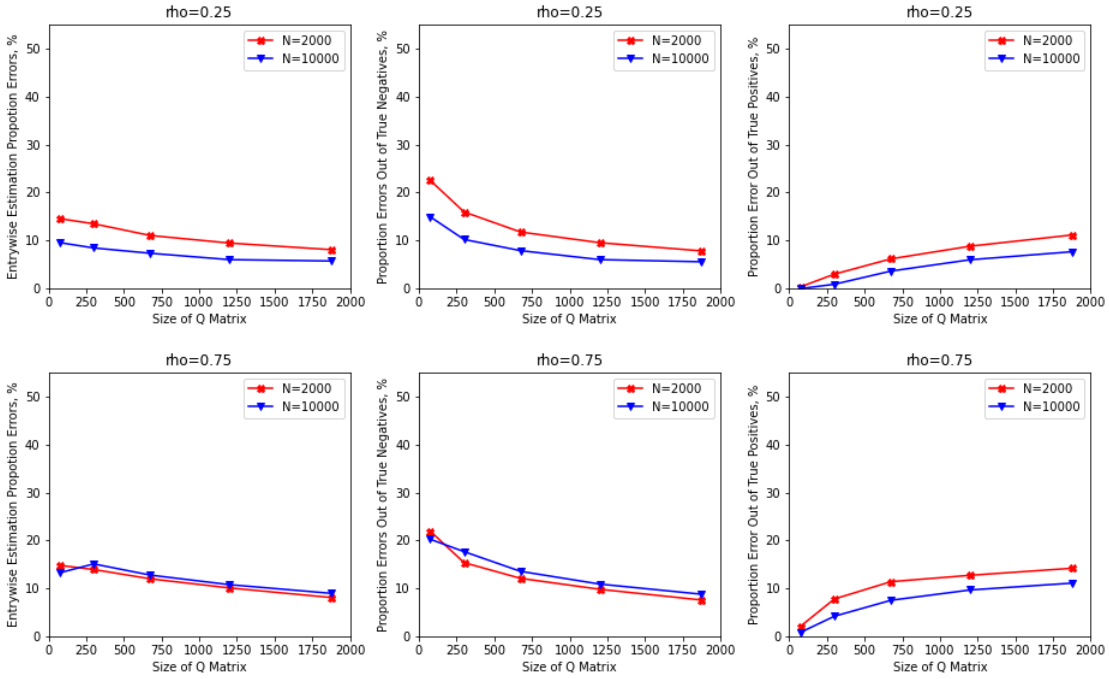

Figure 2.10: Plots of different performance metrics against the size of the Q-matrix for the GDINA data (dependent case with $\delta_{j,0} = 0.1$, $p_j = 0.9$ for all $j = 1, ..., J$). Rows 1 and 2 correspond to correlation settings 0.25 and 0.75 respectively.

2.4.4 Attribute Classifications

As discussed in Section 2.2.3, the marginal distributions of the attributes are mis-specified in RBMs, in which a conditional independent structure is assumed. However, in practice, the latent attributes are often highly correlated and the conditional independence assumption may not hold. This mis-specification in latent attribute distributions is expected to bring in additional errors in the estimated Q-matrix. In order to understand the practical implications of the mis-specification in the estimated Q-matrix, we compare the commonly used attribute classification accuracy (ACC) rate obtained using the estimated Q-matrix ($\hat{Q}$) and the true Q-

matrix (Q). In particular, when there are N subjects, the ACC of the k'th attribute is defined as

$$ACC(k) := \frac{1}{N}\sum_{i=1}^{N}|\hat{\alpha}_{ik} - \alpha_{ik}|,$$

where $\hat{\alpha}_{ik}$ and α_{ik} represent the estimated and the true attribute values, respectively.

The simulation set-ups remain the same as the dependent settings in Section 2.4. All the DINA data, the ACDM data and the GDINA data are considered. Attribute classifications are performed using the estimated $\hat{Q}$ and the true Q under the corresponding true underlying CDMs. The results are summarized in Table 2.1.

Not surprisingly, we observe that the ACC rates obtained using $\hat{Q}$ are worse than that using Q in all settings across all models. The errors in $\hat{Q}$ stem from two sources, the mis-specification error in the latent attributes' marginal distribution and the sample estimation error. On the other hand, we also note that the ACC rates obtained using $\hat{Q}$ do not deteriorate too much from using the true Q when sample size is large, especially under the ACDM and GDINA models. This suggests the Q-matrix estimation accuracy in the ACDM and GDINA models may be less prone to the mis-specification in the latent attributes' marginal distribution. Furthermore, the ACC rates drop as the number of attributes increases in the model. This reflects the increasing difficulty in attribute classifications as the number of attributes increments. Surprisingly, the ACC rates are generally higher when the correlation amongst attributes is higher. This may be because the higher dependency among the attributes results in fewer numbers of possible attribute patterns, making the estimation relatively easier. Not so surprisingly, we also observe that increasing sample size can in general help improve ACC rates.

We also conduct simulation studies to explore the potential of using the proposed method to perform latent attribute classifications directly. The performance of the proposed method in attribute classifications is satisfactory. For more details on the additional simulation results, please refer to Appendix A.

		$N = 2000$				$N = 10000$			
		$\rho = 0.25$		$\rho = 0.75$		$\rho = 0.25$		$\rho = 0.75$	
	Model	$\hat{Q}$	Q	$\hat{Q}$	Q	$\hat{Q}$	Q	$\hat{Q}$	Q
$K = 5$	DINA	0.806	0.944	0.888	0.956	0.830	0.945	0.890	0.957
	ACDM	0.812	0.926	0.911	0.946	0.921	0.928	0.915	0.948
	GDINA	0.918	0.928	0.935	0.949	0.928	0.929	0.947	0.950
$K = 10$	DINA	0.801	0.939	0.898	0.954	0.811	0.940	0.894	0.956
	ACDM	0.815	0.922	0.913	0.946	0.910	0.925	0.906	0.950
	GDINA	0.885	0.924	0.939	0.949	0.926	0.926	0.899	0.951

Table 2.1: Average ACC rates out of 100 repetitions for $K = 5, 10$ attributes respectively obtained using the true CDMs. $\hat{Q}$ and Q denote the estimated Q-matrix from the proposed method and the true Q-matrix respectively.

2.5 Real Data Analysis

We apply our proposed method to a TIMSS data set. TIMSS provides data on the mathematics and science curricular achievement of the fourth and the eighth grade students across countries such as the U.S. The data set contains 23 mathematical items from TIMSS 2003 items and is packed in the CDM package in R (*Robitzsch et al.*, 2020). Both a binary scored subjects' response matrix and an associated expert constructed Q-matrix are included in the data set. In particular, the binary response matrix consists of 757 observations, and it is therefore of dimension 757 by 23. The Q-matrix on the other hand specifies how the 23 items are related to 13

binary mathematical skill attributes, as summarized in Table 2.2.

Skill attributes	Items
1. Understand concepts of a ratio and a unit rate and use language appropriately	1, 7, 20
2. Use ratio and rate reasoning to solve real world and mathematical problems	3, 11, 15, 19, 22
3. Compute fluently with multi-digit numbers and find common factors and multiples	12, 18
4. Apply and extend previous understandings of numbers to the system of rational numbers	4, 17, 23
5. Apply and extend previous understandings of arithmetic to algebraic expressions	8, 13, 16, 21
6. Reason about and solve one-variable equations and inequalities	2, 5, 6,10,14
7. Recognize and represent proportional relationships between quantities	3, 6
8. Use proportional relationships to solve multi-step ratio and percent problems	11
9. Apply and extend previous understandings of operations with fractions to add, subtract, multiply, and divide rational numbers	4, 8, 18, 23
10. Solve real-life and mathematical problems using numerical and algebraic expressions and equations	5
11. Compare two fractions with different numerators and different denominators; Understand a fraction a/b with as $a > 1$ a sum of fractions $1/b$	1, 9, 18
12. Solve multi-step word problems posed with whole numbers and having whole number answers using the four operations, including problems in which remainders must be interpreted. Represent these problems using equations with a letter standing for the unknown quantity; Generate a number or shape pattern that follows a given rule. Identify apparent features of the pattern that were not explicit in the rule itself	5 , 15
13. Use equivalent fraction as a strategy to add and subtract fractions	1, 12, 18

Table 2.2: Clusters of the 23 TIMSS 2003 mathematics items according to the underlying skill attributes.

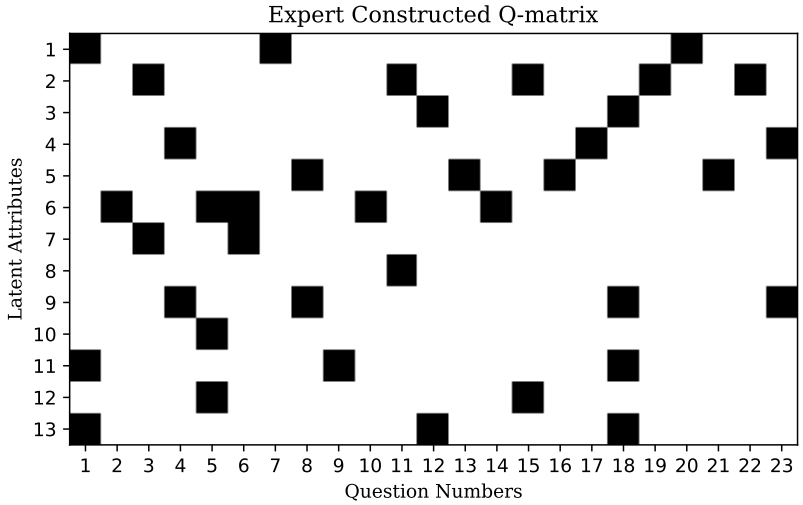

Figure 2.11: Heat-plot of the expert constructed Q^0. The white/black blocks correspond to $q^0_{ij} = 0/1$ respectively.

Note that the provided Q-matrix may not fully represent the ground truth because the construction of the Q-matrix by experts is almost always subjective. In this case, the provided Q-matrix was constructed from the consensus of two experts. When they are not able to reach an agreement for any item through discussion, a third expert would step in to resolve the conflict. The percentage of two experts' overall agreement for the constructed Q-matrix is only 88.89%, according to *Su et al.* (2013). We denote this expert constructed Q-matrix as Q^0 and its (i, j)th entry as q^0_{ij}. A heat-plot of Q^0 is summarized in Figure 2.11. To demonstrate the practical implications of our proposed method, we start with this expert constructed Q^0 and explore further whether our proposed method can improve on the quality of the Q-matrix to better represent the ground truth.

We initialize the weight matrix with Q^0 in our proposed method. The estimated

Q-matrix is denoted as $\hat{Q}$ and its (i,j)th entry as $\hat{q}_{ij}$. If we treat the expert constructed Q^0 as the truth for evaluation purpose, then the entry-wise proportional "error" rate, the out of true positives "error" rate and out of true negatives "error" rate of $\hat{Q}$ are 0.126, 0.053 and 0.139 respectively. The low "error" rates suggest Q^0 and $\hat{Q}$ are similar and our proposed method can indeed recover the main latent structure, especially the positive entries, in the expert constructed Q^0.

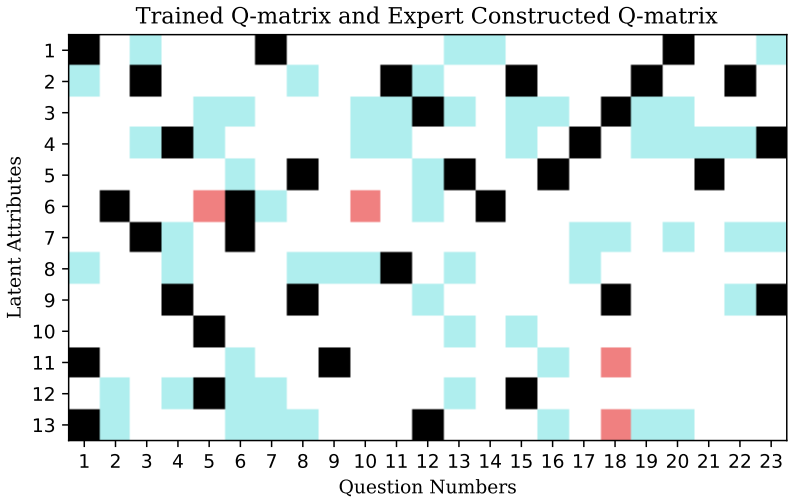

Figure 2.12: Heat-plot to compare between the estimated $\hat{Q}$ and the expert constructed Q^0. The white blocks represent entries (i,j) when both $\hat{q}_{ij} = q^0_{ij} = 0$. The black blocks represent entries (i,j) when both $\hat{q}_{ij} = q^0_{ij} = 1$. The red blocks represent entries (i,j) when $\hat{q}_{ij} = 0$ and $q^0_{ij} = 1$. The blue blocks represent entries (i,j) when $\hat{q}_{ij} = 1$ and $q^0_{ij} = 0$.

Figure 2.12 presents the heat-plot of the comparison between the estimated $\hat{Q}$ and the expert constructed Q^0. In particular, white and black entries represent the cases when $\hat{q}_{ij} = q^0_{ij} = 0$ and when $\hat{q}_{ij} = q^0_{ij} = 1$ respectively. While blue and red entries represent the cases when $\hat{q}_{ij} = 1, q^0_{ij} = 0$ and when $\hat{q}_{ij} = 0, q^0_{ij} = 1$ respectively. We

see that the majority of the positive entries in Q^0 are picked up by $\hat{Q}$, and only 4 of them are predicted to be 0 in $\hat{Q}$, as represented by the red blocks in Figure 2.12. This suggests the proposed method can estimate the Q-matrix with high sensitivity. Some of these false negatives do make sense. For example, item 5 describes three figures arranged in matchsticks with some patterns and asks for the total number of matchsticks that would be used to construct figure 10 if the pattern continues. It is a pattern recognition problem and does not seem to be closely related to attribute 6, "reason about and solve one-variable equations and inequalities". However, we acknowledge that this data driven approach can sometimes make mistakes. For example, the other three false negatives predicted may not make much sense. Take item 10 for example, which reads "inequality equivalent to $x/3 > 8$". It clearly requires the knowledge of attribute 6, which is not successfully identified by the proposed method. On the other hand, the white regions, representing the agreed entry 0's, occupy the majority of the plot. This suggests the specificity is controlled. Moreover, we see some blue blocks scattering in Figure 2.12, which represent the entries that are 0 in Q^0 but are predicted to be 1 in $\hat{Q}$. Some of these blocks capture information that is neglected by the expert when constructing the Q-matrix. Take item 22 for example, whose description is "At a play, 3/25 of the people in the audience were children. What percent of audience is this?" In the expert constructed Q-matrix, this item only requires mastering attribute 2. However, in our estimated $\hat{Q}$, this item is further related to attribute 4, "understanding of rational numbers", 7, "recognizing proportional relationships" and 9, "applying operations with fractions". Nevertheless, we also want to point out that the proposed method may over-select,

resulting in redundant attributes being selected. For example, item 8 reads "If $x = 3$, what is the value of $-3x$". The proposed method predicts that it is related to attribute 2, "Use ratio and rate reasoning to solve real world and mathematical problems", but in fact item 8 does not seem to be related to attribute 2. Therefore, careful examination of the predicted entries is still needed, but it can potentially help to improve the quality of the Q-matrix.

We further compare the goodness-of-fit of Q^0 and $\hat{Q}$ across different CDMs, including the DINA, ACDM and GDINA models using both AIC and BIC as criteria. We note that out of the three models tested, the ACDM gives the smallest values of both AIC and BIC. Moreover, using $\hat{Q}$ gives much smaller AIC (19348.71) than using Q^0 (19568.17) under the ACDM, which suggests the estimated $\hat{Q}$ fits better under the ACDM than the expert constructed Q^0 in terms of AIC. On the other hand, using $\hat{Q}$ achieves a BIC value of 20313.98, slightly worse than a BIC value of 20286.44 obtained by using Q^0. However, the two values are comparable in size and the improvement is not significant. Nonetheless, since we do not know what the true underlying model and the Q-matrix are, consultation to the domain experts is still needed to make assertive conclusions about which of $\hat{Q}$ and Q^0 is better.

2.6 Discussions

In conclusion, our proposed method using RBMs with L_1 penalty can achieve both fast and accurate learning of the large Q-matrices in different types of CDMs. This is shown by both the theoretical proofs developed in Section 2.2.3 and the simulation studies carried out in Section 2.4. The real data analysis on TIMSS data

set further suggests that our method can also work well in real world scenarios, and thus it would provide a powerful tool in large-scale exploratory cognitive diagnosis assessments.

We discuss some potential use cases of our proposed method. One potential use case is to provide a reasonably accurate Q-matrix for cognitive diagnoses such as latent attribute classifications, when no Q-matrix or only an inaccurately specified Q-matrix is available. Depending on the accuracy requirements, the estimated Q-matrix can either be used directly in CDMs to perform latent attribute classifications or can serve as a starting point for domain experts for further refinement before use. Another potential use case is to provide a Q-matrix estimate for test item categorizations and enabling efficient design for future assessments. Similarly, whether the estimated Q-matrix can be used directly depends on the accuracy requirements in different real settings. To add reliability and confidence for direct usage, goodness-of-fit measures such as AIC or BIC can always be evaluated and compared between the estimated Q-matrix and the potentially inaccurate specified Q-matrix if it is available, as a first step. If the goodness-of-fit of the estimated Q-matrix is bad, then either the model used is not appropriate or the estimated Q-matrix is inaccurate. In these cases, consultation to domain experts is still necessary. Nevertheless, our proposed method may help reduce the burden placed on the experts. Based on the estimated Q-matrix, if one finds out that additional items with specific q-vectors need to be included in the test, then it is likely such an item is indeed missing from the original test design. In this scenario, we recommend to include the additional item into the test design to keep safe. Furthermore, in the case when the accuracy

requirement is exceptionally high, our proposed method can still help. In this scenario, we recommend to set the penalty term to be 0 and apply CD Algorithm 1 to train the original RBM on the whole data set to obtain $\hat{\boldsymbol{W}}$. Then for each item j, experts can rank $\{|\hat{w}_{jk}| : k = 1, ..., K\}$ in a descending order first and pay more attention to those $\hat{w}_{jk}$ with large absolute values as those correspond to the q_{jk} that are most likely to be 1's.

Note that by initializing the RBM parameters $\boldsymbol{W}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ randomly, the proposed estimation method assumes no prior knowledge of the Q-matrix. In practice, we may have partial knowledge of the Q-matrix, using which we could potentially obtain a better initialization of the parameters. For example, we may have a pre-specified Q-matrix design with possible mis-specifications in some entries; in such cases, we can initialize the weight matrix $\boldsymbol{W}$ and the visible bias vector $\boldsymbol{b}$ based on our prior knowledge of the Q-matrix. Note that $w_{j,k}$ in $\boldsymbol{W}$ correspond to $\delta_{j,k}q_{j,k}$ in the ACDM. From the initialization perspective, we find what affects the learning accuracy most significantly are the signs of the initial values. So, to keep things simple, we can initialize $\boldsymbol{W}$ with the partially available Q-matrix directly. For the visible biases, if the underlying model is believed to be the DINA model, by considering $\boldsymbol{\alpha} = \mathbf{0}$, we can derive $b_j = \log(g_j/(1 - g_j))$. Under the ACDM or the GDINA model, we can obtain $b_j = \log(\delta_{j,0}/(1 - \delta_{j,0}))$ using a similar argument. Though we do not know g_j or $\delta_{j,0}$ in reality, very likely these values are between 0 and 0.5, in which case $b_j < 0$. It is therefore reasonable to initialize each b_j from a Uniform$(-5, 0)$ distribution. This would help improve the estimation accuracy.

Some limitations of our method include it does not take into account the inter-

actions between the latent attributes due to the assumptions imposed on RBMs. In many real world scenarios, it is not uncommon that the latent attributes interact with one another and have joint effects on the distribution of the observed responses. One potential way to solve this problem is to apply deep Boltzmann machines (DBMs) to model the distribution of the responses. Since DBMs allow interactions between the latent attributes, it will capture the interactions between the latent attributes and take that into account. Moreover, this work focuses more on the estimation part while inference on the estimated Q-matrix is not discussed. It would be interesting to pin down the asymptotic distributional form of this Q-matrix estimator to facilitate inferences such as hypothesis testing and constructing confidence intervals.

CHAPTER III

Statistical Inference for Noisy Binary Matrix Completion

3.1 Introduction

Noisy low-rank matrix completion is concerned with the recovery of a low-rank matrix when only a fraction of noisy entries are observed. This topic has received much attention in the past decade, as a result of its vast applications in practical contexts such as collaborative filtering (*Goldberg et al.*, 1992), system identification (*Liu and Vandenberghe*, 2010) and sensor localisation (*Biswas et al.*, 2006). While the majority of the literature considers the completion of real-valued observations (*Candès and Recht*, 2009; *Candès and Tao*, 2010; *Keshavan et al.*, 2010; *Koltchinskii et al.*, 2011; *Negahban and Wainwright*, 2012; *Chen et al.*, 2020a), many practical problems involve categorical-valued matrices, such as the famous Netflix challenge. Several works have been done on the completion of categorical matrix, including *Davenport et al.* (2014) and *Bhaskar and Javanmard* (2015) for binary matrix, and

Klopp et al. (2015) and *Bhaskar* (2016) for categorical matrix, whose entries can take multiple discrete values. In these works, low-dimensional nonlinear probabilistic models are assumed to handle the categorical data.

Despite the importance of uncertainty quantification to matrix completion, most of the matrix completion literature focus on point estimation and prediction, while statistical inference has received attention only recently. Specifically, *Chen et al.* (2019a) and *Xia and Yuan* (2021) considered statistical inference under the linear models and derived asymptotic normality results. The statistical inference for categorical matrices is more challenging due to the involvement of nonlinear models. To our best knowledge, no work has been done to provide statistical inference for the completion of categorical matrices. In addition to nonlinearity, another challenge in modern theoretical analysis of matrix completion concerns the double asymptotic regime where both the numbers of rows and columns are allowed to grow to infinity. Under this asymptotic regime, both the dimension of the parameters and the number of observable entries grow with the numbers of rows and columns. However, existing theory on the statistical inference for diverging number of parameters (*Portnoy*, 1988; *He and Shao*, 2000; *Wang*, 2011) is not directly applicable, as the dimension of the parameter space in the current problem grows faster than that is typically needed for asymptotic normality; see Section 3.3 for further discussions.

In this chapter, we move one step further towards the statistical inference for the completion of categorical matrix. Specifically, we consider the inference for binary matrix completion under a unidimensional nonlinear factor analysis model with the logit link. Such a nonlinear factor model is one of the most popular models for

multivariate binary data, having received much attention from the theoretical perspective (*Andersen*, 1970; *Haberman*, 1977; *Lindsay et al.*, 1991; *Rice*, 2004), as well as wide applications in various areas, including educational testing (*van der Linden and Hambleton*, 2013), word acquisition analysis (*Kidwell et al.*, 2011), syntactic comprehension (*Gutman et al.*, 2011), and analysis of health outcomes (*Hagquist and Andrich*, 2017). It is also referred to as the Rasch model (*Rasch*, 1960) in psychometrics literature. Despite the popularity and extensive research of the model, its use to binary matrix completion and related statistical inferences for the latent factors and model parameters have not been explored. The considered nonlinear factor model is also closely related to the Bradley-Terry model (*Bradley and Terry*, 1952; *Simons and Yao*, 1999; *Han et al.*, 2020) for directed random graphs and the β-model (*Chatterjee et al.*, 2011; *Rinaldo et al.*, 2013) for undirected random graphs. In fact, the considered model can be viewed as a Bradley-Terry model or β-model for bipartite graphs (*Rinaldo et al.*, 2013). However, the analysis of bipartite graphs is more involved, for which the results and proof strategies in the existing works no longer apply and new technical tools are needed.

Specifically, we introduce a likelihood-based estimator under the nonlinear factor analysis model for binary matrix completion. Under a very flexible missing-entry setting that does not require a random sampling scheme, asymptotic normality results are established that allow us to draw statistical inference. These results suggest that our estimator is asymptotically efficient and optimal, in the sense that the Cramer-Rao lower bound is achieved for model parameters. The proposed method and theory are applied to two real-world problems including (1) linking two forms of a college

admission test that have common items and (2) linking the voting records from multiple years in the United States senate. In the first application, the proposed method allows us to answer the question "for subjects A and B who took different test forms, would subject A perform significantly better than subject B, if they had taken the same test form?". In the second application, it can answer the questions such as "Is Republican senator Marco Rubio significantly more conservative than Republican senator Judd Gregg?". Note that Marco Rubio and Judd Gregg had not served in the United States senate at the same time. We point out that the entry missingness in these applications does not satisfy the commonly assumed random sampling schemes for matrix completion.

The rest of the chapter is organized as follows. In Section 3.2, we introduce the considered Rasch model and discuss its application to binary matrix completion. In Section 3.3, we establish the asymptotic normality for the maximum likelihood estimator. A simulation study is given in Section 3.4 and two real-data applications are presented in Section 3.5. We conclude with discussions on the limitations of the current work and future directions in Section 3.6. The materials of this chapter are mainly based on *Chen et al.* (2021b).

3.2 Model and Estimation

Let Y be a binary matrix with N rows and J columns and $Y_{ij} \in \{0, 1\}$ be the entries of Y, $i = 1, ..., N$, and $j = 1, ..., J$. Some entries of Y are not observable. We use z_{ij} to indicate the missing status of entry Y_{ij}, where $z_{ij} = 1$ indicates that Y_{ij} is observed and $z_{ij} = 0$ otherwise. We let $Z = (z_{ij})_{N \times J}$ be the indicator matrix

for data missingness. The main goal of binary matrix completion is to estimate $E(Y_{ij}|z_{ij} = 0)$.

This problem is typically tackled under a probabilistic model (see e.g., *Cai and Zhou*, 2013; *Davenport et al.*, 2014; *Bhaskar and Javanmard*, 2015), which assumes that Y_{ij}, $i = 1, ..., N$, $j = 1, ..., J$, are independent Bernoulli random variables, with success probability $\exp(m_{ij})/\{1 + \exp(m_{ij})\}$ or $\Phi(m_{ij})$, where m_{ij} is a real-valued parameter and Φ is the cumulative distribution function of the standard normal distribution. It is further assumed that the matrix $M = (m_{ij})_{N\times J}$ is either exactly low-rank or approximately low-rank, where the approximate low-rankness is measured by the nuclear norm of M. Finally, a random sampling scheme is typically assumed for z_{ij}. For example, *Davenport et al.* (2014) considered a uniform sampling scheme where z_{ij} are independent and identically distributed Bernoulli random variables, and *Cai and Zhou* (2013) considered a non-uniform sampling scheme. Under such a random sampling scheme, Z and Y are assumed to be independent and thus data missingness is ignorable.

It is of interest to draw statistical inference on linear forms of M, including the inference of individual entries of M. This is a challenging problem under the above general setting for binary matrix completion, largely due to the presence of a non-linear link function. In particular, the existing results on the inference for matrix completion as established in *Xia and Yuan* (2021) and *Chen et al.* (2019a) are under a linear model that observes $m_{ij} + \epsilon_{ij}$ for the non-missing entries, where ϵ_{ij} are mean-zero independent errors. Their analyses cannot be directly applied to non-linear models.

As the first inference work of binary matrix completion with non-linear models, we start with a basic setting in which we assume the success probability takes a logistic form of M and each m_{ij} depends on a row effect and a column effect only. Asymptotic normality results are then established for the inference of M. Specifically, this model assumes that

(1) given M, Y_{ij}, $i = 1, ..., N$, $j = 1, ..., J$, are independent Bernoulli random variables whose distributions do not depend on the missing indicators Z,

(2) the success probability for Y_{ij} is assumed to be $\exp(m_{ij})/\{1 + \exp(m_{ij})\}$ that follows a logistic link,

(3) M has the model parameterization that $m_{ij} = \theta_i - \beta_j$.

In the rest, θ_i and β_j will be referred to as the row and column parameters, respectively. This parameterization allows the success probability of each entry to depend on both a row effect and a column effect. We now introduce two real-world application examples and discuss the interpretations of the row and column parameters in these applications.

Example III.1. In educational testing, each row of the data matrix represents an subject and each column represents an item (i.e., exam question). Each binary entry Y_{ij} records whether subject i correctly answers item j. The row parameter θ_i is interpreted as the ability of subject i, which is an individual-specific latent score, and the column parameter β_j is interpreted as the difficulty of item j, as the probability of correctly answering an item increases with one's ability θ_i and decreases with the difficulty level β_j of the item.

In Section 3.5, we apply the considered model to link two forms of an educational test, an important practical issue in educational assessment (*Kolen and Brennan*, 2014). That is, consider two groups of subjects taking two different forms of an educational test, where the two forms share some common items but not all, resulting in missingness of the data matrix. As the two test forms may have different difficulty levels, it is usually not fair to directly compare the total scores of two students who take different forms. The proposed method allows us to compare subjects' performance as if they had taken the same test form and to also quantify the estimation uncertainty.

Example III.2. Consider senators' roll call voting records in the United States senate, and in this application, each row of the data matrix corresponds to a senator and each column corresponds to a bill voted in the senate. Each binary response Y_{ij} records whether the senator voted for or against the bill. It has been well recognized in the political science literature (*Poole et al.*, 1991; *Poole and Rosenthal*, 1991) that senate voting behavior is essentially unidimensional, though slightly different latent variable models are used in that literature. That is, it is believed that senators' voting behavior is driven by a unidimensional latent factor, often interpreted as the conservative-liberal political ideology. Moreover, it is a consensus that the Republican senators tend to lie on the conservative side of the factor and the Democratic senators tend to lie on the liberal side, though there are sometimes a very small number of exceptions. To apply the our method to senators' roll call voting records, we pre-process the data as follows. If bill j is more supported by the Republican party than the Democratic party and senator i voted for the bill, then we let $Y_{ij} = 1$.

If bill j is more supported by the Democratic party and senator i voted against the bill, we let $Y_{ij} = 1$. Otherwise, $Y_{ij} = 0$. More details about this data pre-processing can be found in Section 3.5. Under the considered model, the row parameter may be interpreted as the conservativeness score of senator i. That is, the higher the conservativeness score of a senator, the higher chance for him/her to support a bill favored by the Republican party and to vote against a bill favored by the Democratic party. The column parameter characterizes the bill effect.

In Section 3.5, we apply the model to link the roll call voting records from multiple years, where different senators serve different terms in the senate, resulting in missingness of the data matrix. The model allows us to compare senators in terms of their conservative-liberal political ideology, even if they have not served in the senate at the same time.

As mentioned previously, the considered nonlinear factor model can also be viewed as a Bradley-Terry model (*Bradley and Terry*, 1952) for directed graphs that is commonly used for modeling pairwise comparisons. In Remark III.3 below, we discuss this connection and explain the reason why the existing result such as *Han et al.* (2020) does not apply to the current setting.

Remark III.3. Data Y under our model setting can be viewed as a bipartite graph with $N + J$ nodes. Its adjacency matrix takes the form

$$\begin{pmatrix} \mathrm{NA}_{N,N} & Y \\ (1_{N,J} - Y)^T & \mathrm{NA}_{J,J} \end{pmatrix}, \tag{3.1}$$

where $\mathrm{NA}_{N,N}$ and $\mathrm{NA}_{J,J}$ are two matrices whose entries are missing and $1_{N,J}$ is a

matrix with all entries being 1. We let the value of $1 - Y_{ij}$ be missing if Y_{ij} is missing (i.e., $z_{ij} = 0$). Such a directed graph can be modeled by the Bradley-Terry model; see *Bradley and Terry* (1952). In *Han et al.* (2020), asymptotic normality results are established for n-by-n adjacency matrices that follow the Bradley-Terry model when the graph size n grows to infinity. However, *Han et al.* (2020) only consider a uniformly missing setting. That is, the probability that the edges between two nodes are missing is assumed to be the same for all pairs of nodes. This assumption is not satisfied for the adjacency matrix (3.1), due to the two missing matrices on the diagonal. In fact, the asymptotic analysis under the current setting is more involved, due to the need of simultaneously considering two indices N and J and the increased complexity in approximating the asymptotic variance of model parameters.

Given data $\{Y_{ij} : z_{ij} = 1, i = 1, ..., N, j = 1, ..., J\}$, the log-likelihood function for parameters $\theta = (\theta_1, ..., \theta_N)^T$ and $\beta = (\beta_1, ..., \beta_J)^T$ takes the form

$$l(\theta, \beta) = \sum_{i,j:z_{ij}=1} \left[Y_{ij}(\theta_i - \beta_j) - \log\{1 + \exp(\theta_i - \beta_j)\} \right]. \tag{3.2}$$

The identifiability of parameters θ and β is subject to a location shift. That is, the distribution of data remains unchanged, if we add a common constant to all the θ_i and β_j, as the likelihood function in (3.2) only depends on the all differences $\theta_i - \beta_j$. To avoid ambiguity, we require $\sum_{i=1}^{N} \theta_i = 0$ in the rest. We point out that this requirement does not play a role when we draw inference about any linear form of M, as the location shift of θ and β does not affect the value of M. We estimate θ

and β by the maximum likelihood estimator

$$(\hat{\theta}, \hat{\beta}) = \arg\max_{\theta,\beta} \ l(\theta, \beta), s.t., \sum_{i=1}^{N} \theta_i = 0. \tag{3.3}$$

The maximum likelihood estimator of θ and β further leads to the maximum likelihood estimator of M, $\hat{m}_{ij} = \hat{\theta}_i - \hat{\beta}_j$. It is easy to see that (3.3) is a convex optimization problem. Thanks to the low-rank structure of M, this problem can be efficiently solved by performing alternating maximization, as often used for estimating low-rank matrices (*Chen et al.*, 2019b, 2020b; *Udell et al.*, 2016). Such an algorithm is implemented for the numerical studies, whose details are provided in Algorithm 2 below.

3.3 Statistical Inference

In this section, we consider the statistical inference of any linear form of M. Specifically, we use $g : \mathbb{R}^{N\times J} \mapsto \mathbb{R}$ to denote a linear function of M that takes the form

$$g(M) = \sum_{i=1}^{N} \sum_{j=1}^{J} w_{ij} m_{ij}, \tag{3.4}$$

where the weights w_{ij} are pre-specified. It is straightforward that a point estimate of $g(M)$ is given by $g(\hat{M}) = \sum_{i=1}^{N} \sum_{j=1}^{J} w_{ij} \hat{m}_{ij}$. Our goal is to establish the asymptotic normality for $g(\hat{M})$, based on which we can test hypothesis about $g(M)$ or construct confidence intervals. We provide two examples of $g(M)$ that may be of interest in practice.

Algorithm 2: Alternating Gradient Ascent Algorithm

Input: Partially observed data matrix Y, learning rate γ, tolerance threshold tol.

Output: Estimates $\hat{\theta}$, $\hat{\beta}$.

Initialize $\theta^{(0)} = \{\theta_i^{(0)} : i = 1, ..., N\}, \beta^{(0)} = \{\beta_j^{(0)} : j = 1, ..., J\}$ with $\theta_i^{(0)}, \beta_j^{(0)} \sim \text{Uniform}(-c, c)$, $i = 1, ..., N, j = 1, ..., J$,

$\text{JML}^{(0)} = 0$,

$\text{JML}^{(1)} = \sum_{i,j:z_{ij}=1} \Big(y_{ij} \{\theta_i^{(0)} - \beta_j^{(0)}\} - \log\left[1 + \exp\left\{\theta_i^{(0)} - \beta_j^{(0)}\right\}\right] \Big)$;

while $JML^{(1)} - JML^{(0)} > tol$ **do**

  $\text{JML}^{(0)} = \text{JML}^{(1)}$;

  for $i = 1, \ldots, N$ **do**

    $\theta_i^{(1)} \leftarrow \theta_i^{(0)} + \gamma \Big(\sum_{j:z_{ij}=1}^{J} \Big[y_{ij} - \Big\{ e^{\theta_i^{(0)} - \beta_j^{(0)}} \Big\} / \Big\{ 1 + e^{\theta_i^{(0)} - \beta_j^{(0)}} \Big\} \Big] \Big)$;

  end

  for $j = 1, \ldots, J$ **do**

    $\beta_j^{(1)} \leftarrow \beta_j^{(0)} + \gamma \Big(\sum_{i:z_{ij}=1}^{N} \Big[y_{ij} + \Big\{ e^{\theta_i^{(1)} - \beta_j^{(0)}} \Big\} / \Big\{ 1 + e^{\theta_i^{(1)} - \beta_j^{(0)}} \Big\} \Big] \Big)$;

  end

  $\theta^{(1)} = \theta^{(1)} - N^{-1} \sum_{i=1}^{N} \theta_i^{(1)}$;

  $\beta^{(1)} = \beta^{(1)} - N^{-1} \sum_{i=1}^{N} \theta_i^{(1)}$;

  $\text{JML}^{(1)} = \sum_{i,j:z_{ij}=1} \Big(y_{ij} \{\theta_i^{(1)} - \beta_j^{(1)}\} - \log\left[1 + \exp\left\{\theta_i^{(1)} - \beta_j^{(1)}\right\}\right] \Big)$;

  $\theta^{(0)} = \theta^{(1)}$;

  $\beta^{(0)} = \beta^{(1)}$;

end

Example III.4. Consider $g(M) = m_{ij}$ for entry (i, j) that is not observed, i.e., $z_{ij} = 0$. The asymptotic normality of $\hat{m}_{ij}$ allows us to quantify the uncertainty in our prediction $\exp(\hat{m}_{ij})/\{1 + \exp(\hat{m}_{ij})\}$ of the unobserved entry.

Example III.5. Consider $g(M) = \sum_{j=1}^{J} (m_{ij} - m_{i'j})/J = \theta_i - \theta_{i'}$, that is of interest in both educational testing and ranking. If we interpret the model as the Rasch model in educational testing, then θ_i can be regarded as subject i's ability level. Subject i is more likely to answer any question correctly than subject i' if $\theta_i > \theta_{i'}$,

and vise versa. Therefore, even when two subjects do not answer the same test form, the statistical inference of this quantity will allow us to compare their performance and further quantify the uncertainty in this comparison. On the other hand, if we draw connections to Bradley-Terry model in ranking, then θ_i can be interpreted as subject i's ranking criteria. The statistical inference on $(\theta_i - \theta_{i'})$ for any combination of i, i' would allow us to quantify the uncertainty in the rankings of all N subjects.

We first establish the existence and consistency for M, θ, and β. We denote

$$J_* = \min\Big\{\sum_{j=1}^{J} z_{ij} : i = 1, ..., N\Big\} \text{ and } J^* = \max\Big\{\sum_{j=1}^{J} z_{ij} : i = 1, ..., N\Big\}$$

as the minimum and maximum numbers of observed entries per row, respectively. Similarly, we denote

$$N_* = \min\Big\{\sum_{i=1}^{N} z_{ij} : j = 1, ..., J\Big\} \text{ and } N^* = \max\Big\{\sum_{i=1}^{N} z_{ij} : j = 1, ..., J\Big\}$$

as the minimum and maximum numbers of observed entries per column, respectively. Let $\|x\|_\infty = \max\{|x_i| : i = 1, ..., n\}$ be the infinity norm of a vector $x = (x_1, ..., x_n)^T$. Let θ^*, β^* and M^* be the true values of θ, β and M, respectively. Without loss of generality, we assume $N \geq J$. For simplicity, we also assume $N_* > J_*$ and $N^* > J^*$. The following conditions are required.

Condition III.6. As N and J grow to infinity, the following are satisfied:

(a) There exists a constant $k > 0$, such that $N_* \geq kN^{2/3}$ and $J_* \geq kJ^{2/3}$;

(b) $(J_*)^{-1}(\log N)$ converge to 0;

(*c*) There exist positive constants k_1 and k_2 such that $k_1 J_* \leq J^* \leq k_2 J_*$.

Condition III.7. There exists a constant $c < \infty$ such that $\|\theta^*\|_\infty < c$ and $\|\beta^*\|_\infty < c$.

Condition III.8. For any (i, j), there exists $1 \leq i_1, i_2, ..., i_k \leq N$ and $1 \leq j_1, j_2, ..., j_k \leq J$ such that $z_{ij_1} = z_{i_1 j_1} = z_{i_1 j_2} = z_{i_2 j_2} = ... = z_{i_k j_k} = z_{i_k j} = 1$.

We provide some discussions on Conditions III.6 and III.8. Condition III.6(a) requires the number of observations for each parameter to grow to infinity at a suitable rate. Under this requirement, the proportion of observable entries is allowed to decay to zero at the rate $(NJ)^{-\frac{1}{3}}$. Condition III.6(b) is a very mild technical condition that requires J_* to grow faster than $\log(N)$. Condition III.6(c) requires that J_* and J^* are of the same order. This assumption essentially requires a balanced missing data pattern that has a similar spirit as the random sampling regimes for missingness adopted in *Cai and Zhou* (2013) and *Davenport et al.* (2014). Condition III.8 is necessary and sufficient for the identifiability of θ and β; see Proposition III.9 for a formal statement.

Proposition III.9. *If Condition III.8 holds, then θ and β are uniquely determined by equations $\sum_{i=1}^N \theta_i = 0$ and $\theta_i - \beta_j = m_{ij}$, $i = 1, ..., N, j = 1, ..., J$, for which $z_{ij} = 1$.*

If Condition III.8 does not hold, then there exists $(\tilde{\theta}, \tilde{\beta}) \neq (\theta, \beta)$, such that $\sum_{i=1}^N \tilde{\theta}_i = 0$, $\sum_{i=1}^N \theta_i = 0$, and $\theta_i - \beta_j = \tilde{\theta}_i - \tilde{\beta}_j$, $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$.

Theorem III.10 below guarantees the existence and consistency of the maximum likelihood estimator, when both N and J grow to infinity.

Theorem III.10. *Assume Conditions III.6, III.7 and III.8 hold. Then, as N, J grow to infinity, maximum likelihood estimator $(\hat{\theta}, \hat{\beta})$ exists, with probability tending to 1. Furthermore, as N and J grow to infinity, we have*

$$\|\hat{\theta} - \theta^*\|_\infty = O_p\big\{(\log N)^{\frac{1}{2}} J_*^{-\frac{1}{2}}\big\}, \quad \|\hat{\beta} - \beta^*\|_\infty = O_p\big\{(\log J)^{\frac{1}{2}} N_*^{-\frac{1}{2}}\big\},$$

and

$$\max_{i,j} |\hat{m}_{ij} - m_{ij}^*| = O_p\big\{(\log J)^{\frac{1}{2}} N_*^{-\frac{1}{2}} + (\log N)^{\frac{1}{2}} J_*^{-\frac{1}{2}}\big\}.$$

To state the asymptotic normality result for $g(\hat{M})$, we reexpress

$$g(M) = w_g^T \theta + \tilde{w}_g^T \beta,$$

where $w_g = (w_{g1}, \cdots, w_{gN})^T$ and $\tilde{w}_g = (\tilde{w}_{g1}, \cdots, \tilde{w}_{gJ})^T$. Note that this expression always exists, by letting $w_{gi} = \sum_{j=1}^J w_{ij}$ and $\tilde{w}_{gj} = -\sum_{i=1}^N w_{ij}$. We introduce some notations. Let $\sigma_{ij}^2 = \text{var}(Y_{ij}) = \exp(\theta_i^* - \beta_j^*)/\{1 + \exp(\theta_i^* - \beta_j^*)\}^2$, $\sigma_{i+}^2 = \sum_{j=1}^J z_{ij}\sigma_{ij}^2$, and $\sigma_{+j}^2 = \sum_{i=1}^N z_{ij}\sigma_{ij}^2$. Further denote $\hat{\sigma}_{ij}^2 = \exp(\hat{\theta}_i - \hat{\beta}_j)/\{1 + \exp(\hat{\theta}_i - \hat{\beta}_j)\}^2$, $\hat{\sigma}_{i+}^2 = \sum_{j=1}^J z_{ij}\hat{\sigma}_{ij}^2$, and $\hat{\sigma}_{+j}^2 = \sum_{i=1}^N z_{ij}\hat{\sigma}_{ij}^2$ to be the corresponding plug-in estimates. We use $\|\cdot\|_1$ to denote the L_1 norm of a vector. The result is summarized in Theorem III.11 below.

Theorem III.11. *Assume Conditions III.6, III.7 and III.8 hold and $J_*^{-2} N_* (\log N)^2 \to 0$ as $N \to \infty$. Consider a linear function $g(M) = w_g^T \theta + \tilde{w}_g^T \beta$ with $g(M) \neq 0$. Further suppose that there exists a constant $C > 0$ such that $\|w_g\|_1 < C$ and $\|\tilde{w}_g\|_1 < C$.*

Then

$$\tilde{\sigma}(g)^{-1}\big\{g(\hat{M}) - g(M^*)\big\} \to N(0,1) \text{ in distribution},$$

where $\tilde{\sigma}^2(g) = \sum_{i=1}^{N} w_{gi}^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^{J} \tilde{w}_{gj}^2(\sigma_{+j}^2)^{-1}$.

Moreover, $\tilde{\sigma}(g)$ *can be replaced by its plug-in estimator, i.e.,*

$$\hat{\sigma}(g)^{-1}\big\{g(\hat{M}) - g(M^*)\big\} \to N(0,1) \text{ in distribution}, \tag{3.5}$$

where $\hat{\sigma}^2(g) = \sum_{i=1}^{N} w_{gi}^2(\hat{\sigma}_{i+}^2)^{-1} + \sum_{j=1}^{J} \tilde{w}_{gj}^2(\hat{\sigma}_{+j}^2)^{-1}$.

We now discuss the implications of Theorem III.11. For each θ_i, $\mathrm{var}(\hat{\theta}_i) = (\sigma_{i+}^2)^{-1}\big\{1 + o(1)\big\}$. It is worth noting that by the classical theory for maximum likelihood estimation, $(\sigma_{i+}^2)^{-1}$ is the Cramer-Rao lower bound for the estimation of θ_i, when the column parameters β are known. Thus, the result of Theorem III.11 implies that $\hat{\theta}_i$ is an asymptotically optimal estimator for θ_i. Similarly, for each β_j, $\mathrm{var}(\hat{\beta}_j) = (\sigma_{+j}^2)^{-1}\big\{1 + o(1)\big\}$, which also achieves the Cramer-Rao lower bound asymptotically, when the row parameters θ are known. Moreover, $\mathrm{var}(\hat{m}_{ij}) = \mathrm{var}(\hat{\theta}_i - \hat{\beta}_j) = \big\{(\sigma_{i+}^2)^{-1} + (\sigma_{+j}^2)^{-1}\big\}\big\{1 + o(1)\big\}$. We end this section with a remark.

Remark III.12. The derived asymptotic theory is different from that for non-linear regression models of increasing dimensions that has been studied in *Portnoy* (1988), *He and Shao* (2000) and *Wang* (2011). To achieve asymptotic normality under the setting of these works, one at least requires the number of observations to grow faster than the square of the number of parameters. Under the setting of the current work, the model has $N + J - 1$ free parameters, while the number of observed entries is allowed to grow as slow as $O((NJ)^{\frac{2}{3}})$, which is much slower than $(N + J - 1)^2$. Even

when there is no missing entries, the number of observed entries is NJ which does not grow as fast as $(N + J - 1)^2$.

3.4 Simulation Study

We study the finite-sample performance of the likelihood-based estimator. We consider two settings: (1) $N = 5000$ and $J = 200$, and (2) $N = 10000$ and $J = 400$. Missing data are generated under a block-wise design. That is, we split the rows into five equal-sized clusters and the columns into four equal-sized clusters. We let each row cluster correspond to the columns from a distinct combination of two column clusters. Rows from the same cluster have the same missing pattern. Specifically, their entries are observable and only observable, on the columns that this row cluster correspond to. This missing data pattern can be illustrated by a five-by-four block-wise matrix $\{(1,0,0,1,0)^T, (1,1,0,0,1)^T, (0,1,1,1,0)^T, (0,0,1,0,1)^T\}$, where 1 and 0 represent a submatrix with $z_{ij} = 1$ and 0, respectively. An illustration of the missing pattern Z is illustrated in Figure 3.1. Under the first setting, $N_* = 2000, N^* = 3000$, and $J_* = J^* = 100$. Under the second setting, $N_* = 4000, N^* = 6000$, and $J_* = J^* = 200$. For each setting, θ is simulated from a uniform distribution over the space $\{x = (x_1, ..., x_N)^T : \sum_{i=1}^N x_i = 0, -2 \leq x_i \leq 2\}$, and β is obtained by simulating β_j independently from the uniform distribution over the interval $[-2, 2]$. For each setting, 2000 independent datasets are generated from the considered model.

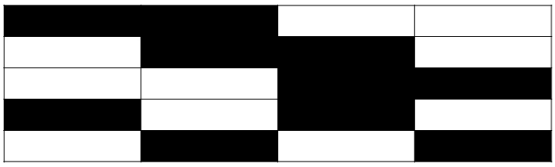

Figure 3.1: Heat map of Z. The black and white regions correspond to $z_{ij} = 1$ and 0, respectively.

Under setting (1), the mean squared estimation errors for M, θ and β are 0.067, 0.064 and 0.0028, respectively, across all relevant entries and all 2000 independent samples. Under setting (2), these values read 0.033, 0.031 and 0.0013, respectively. Unsurprisingly, increasing sample sizes can improve estimation accuracy.

We then examine the variance approximation in Theorem III.11. We compare $\hat{\sigma}^2(g)$, $\tilde{\sigma}^2(g)$ and $s^2(g)$, where $s^2(g)$ denotes the sample variance of $g(\hat{M})$ that is calculated based on the 2000 simulations. As $\hat{\sigma}^2(g)$ varies across the datasets, we calculate $\bar{\sigma}^2(g)$ as the average of $\hat{\sigma}^2(g)$ over 2000 simulated datasets. We consider functions $g(M) = m_{ij}, \theta_i, \beta_j,\ i = 1, ..., N, j = 1, ..., J$. The results are given in Figure 3.2, where panels (a)-(c) show the scatter plots of $s^2(g)$ against $\bar{\sigma}^2(g)$ and panels (d)-(f) show those of $s^2(g)$ against $\tilde{\sigma}^2(g)$. These plots suggest that $\bar{\sigma}^2(g)$, $\tilde{\sigma}^2(g)$, and $s^2(g)$ are close to each other, for the specific forms of g that are examined.

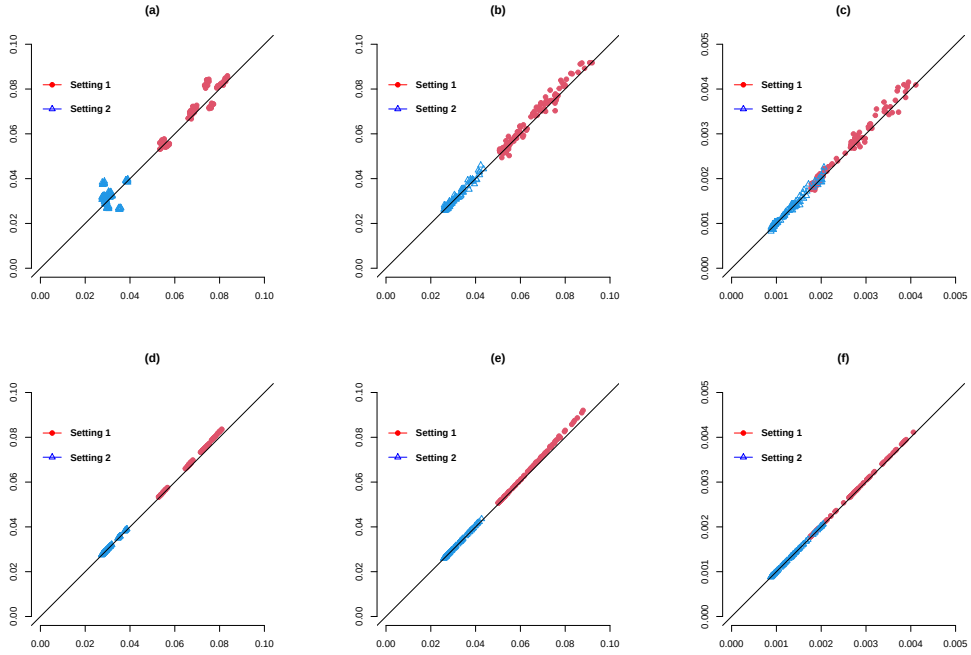

Figure 3.2: Panels (a)-(c) plot $s^2(g)$ against $\bar{\sigma}^2(g)$ for $g(M) = m_{ij}$, θ_i, and β_j, respectively. Panels (d)-(f) plot $s^2(g)$ against $\tilde{\sigma}^2(g)$ for $g(M) = m_{ij}$, θ_i and β_j, respectively. Each panel shows 100 randomly sampled m_{ij}, θ_i, or β_j under each setting. The line $y = x$ is given as a reference.

To validate asymptotic normality, we compare the empirical densities of the 2000 sample estimates of m_{11}, θ_1 and β_1 against their respective theoretical normal density curves in Figure 3.3 for illustration. We can observe from Figure 3.3 that the empirical distributions of the estimates agree well with their corresponding theoretical distributions.

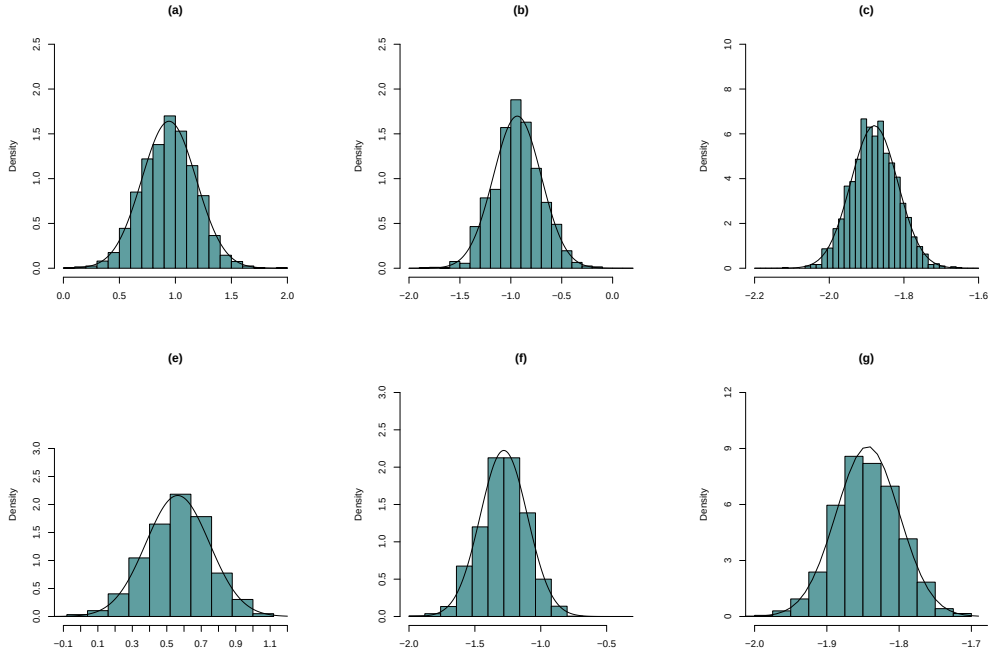

Figure 3.3: Panels (a)-(c) presents the empirical densities (histograms) of $\hat{m}_{11}$, $\hat{\theta}_1$ and $\hat{\beta}_1$ under setting (1), respectively, out of 2000 simulations. Panels (e)-(g) presents the empirical densities of $\hat{m}_{11}$, $\hat{\theta}_1$ and $\hat{\beta}_1$ under setting (2), respectively, out of 2000 simulations. The curves are theoretical density curves of $\mathrm{N}(m_{11}, \tilde{\sigma}^2(m_{11}))$, $\mathrm{N}(\theta_1, \tilde{\sigma}^2(\theta_1))$ and $\mathrm{N}(\beta_1, \tilde{\sigma}^2(\beta_1))$, respectively, included as references.

Furthermore, for each m_{ij}, θ_i, and β_j, we construct its 95% Wald interval based on (3.5), for which the empirical coverage based on 2000 independent replications is computed. This result is shown in Figure 3.4, where the two panels correspond to the two simulation settings, respectively. In each panel, the three box-plots show the empirical coverage probabilities for entries of M, θ, and β, respectively. As we can see, all these empirical coverage probabilities are close to the nominal level 95%.

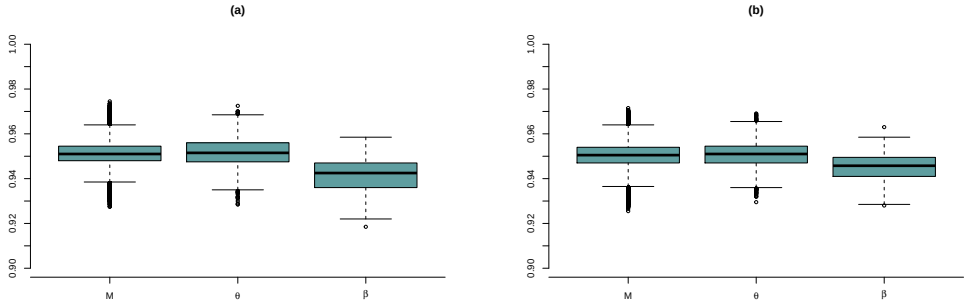

Figure 3.4: Panels (a) and (b) show the empirical coverage rates for the 95% Wald intervals under settings (1) and (2), respectively.

3.5 Real-data Applications

In what follows, we consider two real-data applications.

3.5.1 Application to Educational Testing

We first apply the proposed method to link two forms of an educational test that share common items. The dataset is a benchmark dataset for studying linking methods for educational testing (*González and Wiberg*, 2017). It contains binary responses from two forms of a college admission test. Each form has 120 items and is answered by 2000 subjects. There are 40 common items shared by the two test forms. There is no missing data within each test. Thus, $N = 4000$, $J = 200$, and 40% of the data entries are missing. We apply the proposed method to this dataset. Making use of Theorem III.11, 95% confidence intervals are obtained for both the row (i.e., person) parameters and the column (i.e., item) parameters. The results allow us to compare students who took different test forms, as well as non-common items

from the two forms. For illustration, we randomly choose 100 row parameters and 100 column parameters and show their 95% confidence intervals in Figure 3.5. Such uncertainty quantification can be vital for colleges when making admission decisions.

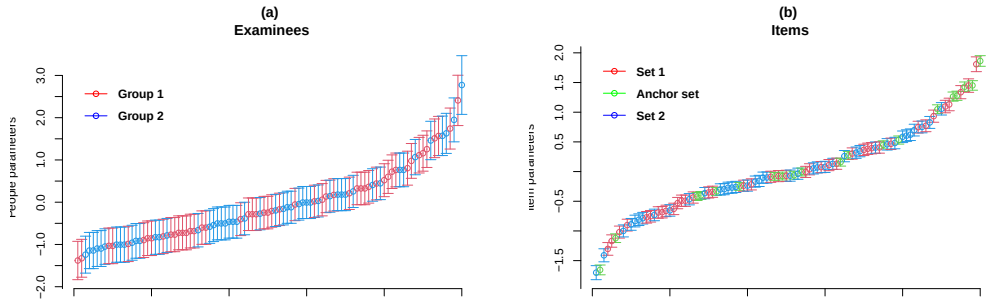

Figure 3.5: (a) 95% confidence intervals of 100 row parameters, with 50 randomly selected from each group. (b) 95% confidence intervals of the 100 column parameters, with 40 each randomly chosen from group 1 and group 2 and 20 randomly selected from anchor items (i.e., common items).

3.5.2 Application to Senate Voting

We now apply the proposed method to the United States senate roll call voting data. Data are from the 111th through the 113th congress that include the voting records from January 11, 2009 to December 16, 2014. Quite a few senators did not serve for the entire period.

To apply our method to senators' roll call voting records with θ_i being interpreted as the conservativeness score of senator i, we pre-process the data as follows. First, five senators who did not serve for more than half a year during the period are removed from the dataset, including Edward M. Kennedy, Joe Biden, Hilary Clinton, Julia Salazar and Carte Goodwin. Second, 191 bills are removed, as all the observed

votes to each of these bills are the same and consequently their maximum likelihood estimates do not exist. After these two steps, the resulting dataset contains $N = 139$ senators and $J = 1648$ bills. Finally, for bill j that has a higher percentage support within the Republican party than that within the Democratic party, we let $Y_{ij} = 1$ if senator i voted for the bill and $Y_{ij} = 0$ if senator i voted against it. For bill j that has a higher percentage support within the Democratic party than that within the Republican party, we let $Y_{ij} = 1$ if senator i voted against the bill and $Y_{ij} = 0$ if he/she voted for it. The value of Y_{ij} is missing, if the senator chose not to vote or he/she was not in the senate when this bill was voted. For the final data being analyzed, the proportion of missing entries is 26.1% and the connectedness Condition III.8 is satisfied. The missingness pattern of the dataset is given in Figure 3.6. Note that in this example, $N < J$. However, our asymptotic results are still applicable if we simply switch the roles of N and J in the required conditions.

Our asymptotic results allow us to compare senators' ideological position, even if they did not serve in the senate at the same time. For example, Judd Gregg served in the senate between January 3, 1993 and January 3, 2011, while Marco Rubio started his first term as a senator since January 3, 2011. In our model, Judd Gregg (θ_i) and Marco Rubio (θ_k) have estimated conservativeness scores of 2.59 and 4.25, respectively. Applying our asymptotic results, we have $\hat{\theta}_i - \hat{\theta}_k = -1.66$ and its standard error is 0.169. If we test $H_0 : \theta_i = \theta_k$ against $H_1 : \theta_i \neq \theta_k$, we obtain an extremely small p-value of 9.0×10^{-23}. Therefore, we conclude that senator Marco Rubio is significantly more conservative than senator Judd Gregg.

In addition, we present in Tables 3.1 and 3.2 the ten senators with the largest

row parameter estimates, and the ten senators with the smallest row parameter estimates. These results align well with the public perceptions about these senators. For example, Jim Demint, who is ranked the most conservative senator in this dataset by our method, was also identified by Salon as one of the most conservative members of the senate (*Kornacki*, 2011). Our method ranks Mike Lee the second, though his conservativeness score is not significantly different from that of Demint. In fact, in 2017, the New York Times used the NOMINATE system (*Poole and Rosenthal*, 2001) to arrange Republican senators by ideology and ranked Lee as the most conservative member of the Senate (*Parlapiano et al.*, 2017). For another example, Brian Schatz who is ranked to be the most liberal senator by our method, is well-known as a liberal Democrat. During his time in the senate, he voted with the Democratic party on most issues.

Finally, the 95% confidence intervals for all the row parameters are shown in Figure 3.7 and a full list of rankings for all the 139 senators is given in the Appendix section, where the corresponding row parameter estimates and their standard errors are also presented.

Figure 3.6: Heat map of Z. The black and white regions correspond to $z_{ij} = 1$ and 0, respectively.

Table 3.1: Ranking of the top 10 most conservative senators predicted by the model. Rep and Dem represent the Republican party and the Democratic party, respectively.

Rank	Senator (party)	State	Conservativeness Score (s.e.($\hat{\theta}$))
1	Jim DeMint (Rep)	South Carolina	5.87 (0.157)
2	Mike Lee (Rep)	Utah	5.73 (0.138)
3	Ted Cruz (Rep)	Texas	5.65 (0.195)
4	Tom Coburn (Rep)	Oklahoma	5.25 (0.114)
5	Rand Paul (Rep)	Kentucky	5.24 (0.129)
6	Tim Scott (Rep)	South Carolina	5.17 (0.176)
7	Jim Bunning (Rep)	Kentucky	4.92 (0.204)
8	Ron Johnson (Rep)	Wisconsin	4.84 (0.119)
9	James Risch (Rep)	Idaho	4.81 (0.102)
10	Jim Inhofe (Rep)	Oklahoma	4.69 (0.103)

Table 3.2: Ranking of the top 10 most liberal senators predicted by the model. Rep and Dem represent the Republican party and the Democratic party, respectively.

Rank	Senator (party)	State	Conservativeness Score (s.e.($\hat{\theta}$))
1	Brian Schatz (Dem)	Hawaii	-4.74 (0.468)
2	Roland Burris (Dem)	Illinois	-4.43 (0.297)
3	Mazie Hirono (Dem)	Hawaii	-4.17 (0.383)
4	Cory Booker (Dem)	New Jersey	-4.14 (0.572)
5	Tammy Baldwin (Dem)	Wisconsin	-3.90 (0.352)
6	Sherrod Brown (Dem)	Ohio	-3.89 (0.168)
7	Tom Udall (Dem)	New Mexico	-3.85 (0.165)
8	Dick Durbin (Dem)	Illinois	-3.83 (0.164)
9	Ben Cardin (Dem)	Maryland	-3.82 (0.163)
10	Sheldon Whitehouse (Dem)	Rhode Island	-3.74 (0.163)

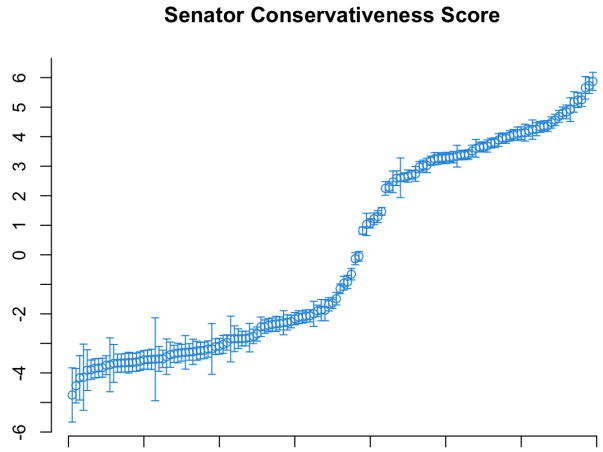

Figure 3.7: 95% confidence intervals of 139 row (i.e. senator) parameters in the senate voting application.

3.6 Discussions

This chapter considers the statistical inference for binary matrix completion under a unidimensional nonlinear factor model, the Rasch model. Asymptotic normality results are established. Our results suggest that the maximum likelihood estimator is statistically efficient, even though the number of parameters diverges. Our simulation study shows that the developed asymptotic result provides a good approximation to finite sample data, and two real-data examples demonstrate its usefulness in the areas of educational testing and political science.

The current results can be easily extended to matrix completion problems with quantized measurement that has a similar natural exponential family form. Admittedly, the model considered may be oversimple for complex application problems, for

example, certain collaborative filtering problems for which the rank of the underlying matrix M may be higher than considered here and the underlying latent factors may be multidimensional. The extension of the current results to more flexible models is left for future investigation. As the first inference result for binary matrix completion, we believe the current results will shed light on the statistical inference for more general matrix completion problems.

CHAPTER IV

DIF Statistical Inference and Detection without Knowing Anchoring Items

4.1 Introduction

Measurement invariance refers to psychometric equivalence of an instrument (e.g., a questionnaire or test) across several specified groups, such as gender and ethnicity. The lack of measurement invariance suggests that the instrument has different structures or meanings to different groups, leading to biases in measurements (*Millsap*, 2012).

Measurement invariance is typically assessed by differential item functioning (DIF) analysis of item response data that aims to detect the measurement non-invariant items (i.e. DIF items). More precisely, a DIF item has a response distribution that depends on not only the latent trait measured by the instrument but also respondents' group membership. Therefore, the detection of a DIF item involves comparing the item responses of different groups, conditioning on the latent traits.

The complexity of the problem lies in that individuals' latent trait levels cannot be directly observed but are measured by the instrument that may contain DIF items. In addition, different groups may have different latent trait distributions. This problem thus involves identifying the latent trait, and then conducting the group comparison given individuals' latent trait levels.

Many statistical methods have been developed for DIF analysis. Traditional methods for DIF analysis require prior knowledge about a set of DIF-free items, which is known as the anchor set. This anchor set is used to identify the latent trait distribution. These methods can be classified into two categories. Methods in the first category (*Mantel and Haenszel*, 1959; *Dorans and Kulick*, 1986; *Swaminathan and Rogers*, 1990; *Shealy and Stout*, 1993; *Zwick et al.*, 2000; *Zwick and Thayer*, 2002; *May*, 2006; *Soares et al.*, 2009; *Frick et al.*, 2015) do not explicitly assume an item response theory (IRT) model, and methods in the second category (*Thissen*, 1988; *Lord*, 1980; *Kim et al.*, 1995; *Raju*, 1988, 1990; *Cai et al.*, 2011; *Woods et al.*, 2013; *Thissen*, 2001; *Oort*, 1998; *Steenkamp and Baumgartner*, 1998; *Cao et al.*, 2017; *Woods et al.*, 2013; *Tay et al.*, 2015, 2016) are developed based on IRT models. Compared with non-IRT-based methods, an IRT-based method defines the DIF problem more clearly, at the price of potential model mis-specification. Specifically, an IRT model represents the latent trait as a latent variable and further characterizes the item-specific DIF effects by modeling each item response distribution as a function of the latent variable and group membership.

Anchor-set-based methods rely heavily on a correctly specified set of DIF-free items. The mis-specification of some anchor items can lead to invalid statistical

inference results (*Kopf et al.*, 2015b). To address this issue, item purification methods (*Candell and Drasgow*, 1988; *Clauser et al.*, 1993; *Fidalgo et al.*, 2000; *Wang and Yeh*, 2003; *Wang and Su*, 2004; *Wang et al.*, 2009b; *Kopf et al.*, 2015a,b) have been proposed, which iteratively select an anchor set by stepwise model selection methods. More recently, regularized estimation methods (*Magis et al.*, 2015; *Tutz and Schauberger*, 2015; *Huang*, 2018; *Belzak and Bauer*, 2020; *Bauer et al.*, 2020; *Schauberger and Mair*, 2020) have been proposed that use LASSO-type regularized estimation procedures for simultaneous model selection and parameter estimation. Moreover, *Yuan et al.* (2021) proposed a visualization method for the detection of DIF under the Rasch model (*Rasch*, 1960). Their methods are based on testing differential item pair functioning, which does not require prior information of an anchor set. Unfortunately, unlike many anchor-set-based methods with a correctly specified anchor set, all these methods do not provide valid statistical inference for testing the null hypothesis of "item j is DIF-free", for each item j. Consequently, the type-I error for testing the hypothesis cannot be guaranteed to be controlled at a pre-specified significance level. Furthermore, although the regularized estimation methods have been shown to accurately detect DIF items, they are typically computationally intensive, since they involve solving multiple regularized maximum likelihood estimation problems with different tuning parameters.

This chapter proposes a new method that can statistically accurately and computationally efficiently estimate the DIF effects without requiring prior knowledge. It draws statistical inference on the DIF effects of individual items, yielding valid confidence intervals and p-values. The point estimation and statistical inference lead

to accurate detection of the DIF items, for which the type-I error of DIF detection can be controlled by the inference results. The method is proposed under a 2PL IRT measurement model. The key to this method is a minimal L_1 norm assumption for identifying the true model.

Two-Parameter Logistic Model

The 2PL model (*Birnbaum*, 1968) is widely used to model binary item responses (e.g., wrong/right or absent/present). In the absence of DIF, the 2PL model assumes a logistic relationship between Y_{ij} and θ_i, which is independent of the value of x_i. That is,

$$P(Y_{ij} = 1|\theta_i = \theta) = \frac{\exp(a_j\theta + d_j)}{1 + \exp(a_j\theta + d_j)}, \tag{4.1}$$

where the slope parameter a_j and intercept parameter d_j are typically known as the discrimination and easiness parameters, respectively. The right hand side of (4.1) as a function of θ is known as the 2PL item response function.

The rest of the chapter is organized as follows. The first part, as summarized in Section 4.2, focuses on developing binary group DIF analysis methods. In specific, Section 4.2.1 introduces the model setup with DIF effects. Section 4.2.2 introduces the related literature to our proposed method, including the likelihood ratio test method and the $L1$ regularization type of methods. In Section 4.2.3, a new method is proposed for the statistical estimation and inference on binary group DIF effects. Simulation studies and a real data application to EPQ-R data are given in Sections 4.2.4 and 4.2.5, respectively. The second part, as summarized in Section 4.3, extends the results in the first part of the chapter to deal with multiple group DIF

analysis, with an emphasis on the application to PISA 2018 data. In particular, Section 4.3.1 specifies the model set-up of the intended setting to handle multiple group DIF problem. Section 4.3.2 gives the details of our proposed estimation, inference and a group clustering method based on detected DIF structures. A succinct simulation study is included in Section 4.3.3 and an in-depth analysis on the PISA 2018 data can be found in Section 4.3.4. Lastly, we conclude this chapter with some discussions on the potential limitations and some future directions of this work. The materials of this chapter are mainly based on *Chen et al.* (2021a).

4.2 Binary Group DIF Detection and Inference

Consider N individuals answering J items. Let $Y_{ij} \in \{0, 1\}$ be a binary random variable, denoting individual i's response to item j. Let y_{ij} be the observed value, i.e., the realization, of Y_{ij}. For the ease of exposition, we use $\mathbf{Y}_i = (Y_{i1}, ..., Y_{iJ})$ to denote the response vector of individual i. We assume the individuals are from two groups, indicated by $x_i = 0, 1$, where 0 and 1 are referred to as the reference and focal groups, respectively. We further introduce a latent variable θ_i, which represents the latent trait that the items are designed to measure. DIF occurs when the distribution of $\mathbf{Y}_i$ depends on not only θ_i but also x_i. More precisely, DIF occurs if $\mathbf{Y}_i$ is not conditionally independent of x_i, given θ_i. Seemingly a simple group comparison problem, DIF analysis is non-trivial due to the latency of θ_i. In particular, the distribution of θ_i may depend on the value of x_i, which confounds the DIF analysis.

4.2.1 Model Set-up

When the items potentially suffer from DIF, then the item response functions may depend on the group membership x_i. In that case, the item response function can be modeled as follows,

$$P(Y_{ij} = 1|\theta_i = \theta, x_i) = \frac{\exp(a_j\theta + d_j + \gamma_j x_i)}{1 + \exp(a_j\theta + d_j + \gamma_j x_i)}, \tag{4.2}$$

where γ_j is an item-specific parameter that characterizes its DIF effect. More precisely,

$$\frac{P(Y_{ij} = 1|\theta_i = \theta, x_i = 1)/(1 - P(Y_{ij} = 1|\theta_i = \theta, x_i = 1))}{P(Y_{ij} = 1|\theta_i = \theta, x_i = 0)/(1 - P(Y_{ij} = 1|\theta_i = \theta, x_i = 0))} = \exp(\gamma_j).$$

That is, $\exp(\gamma_j)$ is the odds ratio for comparing two individuals from two groups who have the same latent trait level. Item j is DIF-free under this model when $\gamma_j = 0$. We further make the local independence assumption as in most IRT models. That is, Y_{i1}, ..., Y_{iJ} are assumed to be conditionally independent, given θ_i and x_i.

We assume the conditional distribution of θ_i given x_i to follow a normal distribution,

$$\theta_i|x_i \sim N(\mu x_i, 1).$$

Note that the latent trait distribution for the reference group is set to a standard normal distribution to identify the location and scale of the latent trait. A similar assumption is typically adopted in IRT models for a single group of individuals.

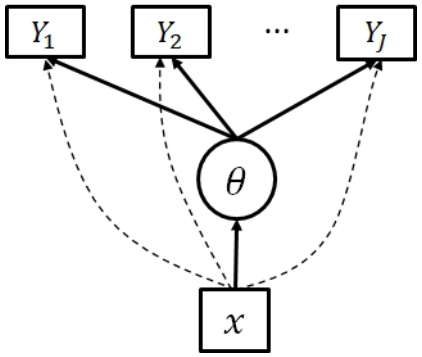

Figure 4.1: The path diagram of the proposed model for DIF analysis. The subscript i is omitted for simplicity. The dashed lines from x to Y_j indicate the DIF effects.

The marginal likelihood function for the proposed model takes the form

$$L(\Xi) = \prod_{i=1}^{N} \int \left(\prod_{j=1}^{J} \frac{\exp(y_{ij}(a_j\theta + d_j + \gamma_j x_i))}{1 + \exp(a_j\theta + d_j + \gamma_j x_i)} \right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(\theta - \mu x_i)^2}{2} \right) d\theta, \quad (4.3)$$

where $\Xi = \{\mu, a_j, d_j, \gamma_j, j = 1, ..., J\}$ denotes all the fixed model parameters.

The goal of DIF analysis is to detect the DIF items, i.e., the items for which $\gamma_j \neq 0$. Unfortunately, without further assumptions, this problem is ill-posed due to the non-identifiability of the model. We discuss this identifiabibility issue below.

Note that for any constant c, the model remains equivalent, if we simultaneously replace μ and γ_j by $\mu + c$ and $\gamma_j - a_j c$, respectively, and keep a_j unchanged. This identifiability issue is due to that all the items are allowed to suffer from DIF, resulting in an unidentified latent trait. In other words, without further assumptions, it is impossible to disentangle the DIF effects and the difference between the two groups' latent trait distributions.

4.2.2 Related Works

Many of the IRT-based DIF analyses (*Thissen et al.*, 1986; *Thissen*, 1988; *Thissen et al.*, 1993) require prior knowledge about a subset of DIF-free items, which are known as the anchor items. More precisely, we denote this known subset by A. Under the model described above, it implies that the constraints $\gamma_j = 0$ are imposed for all $j \in A$ in the estimation. With these zero constraints, the γ_j parameters cannot be freely transformed, and thus, the model becomes identifiable. Therefore, for each non-anchor item $j \notin A$, the hypothesis of $\gamma_j = 0$ can be tested, for example, by a likelihood ratio test. The DIF items can then be detected based on the statistical inference of these hypothesis tests.

The validity of the anchor-item-based analyses relies on the assumption that the anchor items are truly DIF-free. If the anchor set includes one or more DIF items, then the results can be misleading (*Kopf et al.*, 2015b). To address the issue brought by the mis-specification of the anchor set, item purification methods (*Candell and Drasgow*, 1988; *Clauser et al.*, 1993; *Fidalgo et al.*, 2000; *Wang and Yeh*, 2003; *Wang and Su*, 2004; *Wang et al.*, 2009b; *Kopf et al.*, 2015b,a) have been proposed that iteratively purify the anchor set. These methods conduct model selection using a stepwise procedure to select the anchor set, implicitly assuming that there exists a reasonably large set of DIF items. Then DIF is assessed by hypothesis testing given the selected anchor set. This type of methods also has several limitations. First, the model selection results may be sensitive to the choice of the initial set of anchor items and the specific stepwise procedure (e.g., forward or backward selection), which is a common issue with stepwise model selection procedures (e.g., stepwise variable

selection for linear regression). Second, the model selection step has uncertainty. As a result, there is no guarantee that the selected anchor set is completely DIF-free, and furthermore, the post-selection statistical inference of items may not be valid in the sense that the type-I error may not be controlled at the targeted significance level.

Regularized estimation methods (*Magis et al.*, 2015; *Tutz and Schauberger*, 2015; *Huang*, 2018; *Belzak and Bauer*, 2020; *Bauer et al.*, 2020; *Schauberger and Mair*, 2020) have also been proposed for identifying the anchor items, which also implicitly assumes that many items are DIF-free. These methods do not require prior knowledge about anchor items, and simultaneously select the DIF-free items and estimate the model parameters using a LASSO-type penalty (*Tibshirani*, 1996). A regularized estimation procedure solves the following optimization problem,

$$\hat{\Xi}^{\lambda} = \arg\min_{\Xi} -\log L(\Xi) + \lambda \sum_{j=1}^{J} |\gamma_j|, \tag{4.4}$$

where $\lambda > 0$ is a tuning parameter that determines the sparsity level of the estimated γ_j parameters. Generally speaking, a larger value of λ leads to a more sparse vector $\hat{\boldsymbol{\gamma}}^{\lambda} = (\hat{\gamma}_1^{\lambda}, ..., \hat{\gamma}_J^{\lambda})$. A regularization method (*Belzak and Bauer*, 2020) solves the optimization problem (4.4) for a sequence of λ values, and then selects the tuning parameter λ based on the Bayesian Information Criterion (*Schwarz*, 1978). Let $\hat{\lambda}$ be the selected tuning parameter. Items for which $\hat{\gamma}_j^{\hat{\lambda}} \neq 0$ are classified as DIF items and the rest are classified as DIF-free items. While the regularization methods are computationally more stable than stepwise model selection in the item purification

methods, they also suffer from some limitations. First, they involve solving non-smooth optimization problems like (4.4) for different tuning parameter values, which is not only computationally intensive but also requires tailored computation code that is not available in most statistical packages/software for DIF analysis. Second, these methods may be sensitive to the choice of the tuning parameter. Although methods and theories have been developed in the statistics literature to guide the selection of the tuning parameter, there is no consensus on how the tuning parameter should be chosen, leaving ambiguity in the application of these methods. Finally, as a common issue of regularized estimation methods, obtaining valid statistical inference from these methods is not straightforward. That is, regularized estimation like (4.4) does not provide a valid p-value for testing the null hypothesis $\gamma_j = 0$. In fact, post-selection inference after regularized estimation was conducted in *Bauer et al.* (2020), where the type I error cannot be controlled at the targeted level under some simulation scenarios.

4.2.3 Proposed Method

In what follows, we propose a new method for DIF analysis that does not require prior knowledge about anchor items. As will be shown in the rest, the proposed method can not only accurately detect the DIF items, but also provide valid statistical inference for testing the hypotheses of $\gamma_j = 0$.

We impose a condition for identifying the true model in the same spirit as the sparsity assumption that is implicitly imposed in the regularized estimation and item purification methods. Recall that the model remains equivalent, if we simultaneously

replace μ and γ_j by $\mu + c$ and $\gamma_j - a_j c$, respectively, and keep a_j unchanged. Let a_j^*, d_j^*, γ_j^*, and μ^* denote the true model parameters. To identify the true model, we require the following minimal L_1 (ML1) condition to hold

$$\sum_{j=1}^{J} |\gamma_j^*| < \sum_{j=1}^{J} |\gamma_j^* - a_j^* c|, \tag{4.5}$$

for all $c \neq 0$. This assumption implies that, among all models that are equivalent to the true model, the true parameter vector $\boldsymbol{\gamma}^*$ has the smallest L_1 norm. Since the L_1 norm measures the sparsity level of a vector (*Tibshirani*, 1996), the ML1 condition tends to hold when $\boldsymbol{\gamma}^*$ is sparse. Equivalently, we can rewrite (4.5) as

$$\arg\min_c h(c) = 0, \tag{4.6}$$

where $h(c) = \sum_{j=1}^{J} |\gamma_j^* - a_j^* c|$. We show the function $h(c)$ in Figure 4.2, where $J = 10$, $a_j^* = 1$ for all j, $\gamma_j^* = 0$ and 1 when $j = 1, ..., 8$ and $j = 9, 10$, respectively. The following proposition provides a sufficient and necessary condition for the ML1 condition to hold. The proof is given in Appendix C.

Proposition IV.1. *Condition* (4.5) *holds if and only if*

$$\sum_{j=1}^{J} |a_j^*| \left(-I\big(\frac{\gamma_j^*}{a_j^*} \geq 0\big) + I\big(\frac{\gamma_j^*}{a_j^*} < 0\big) \right) < 0$$

and

$$\sum_{j=1}^{J} |a_j^*| \left(-I\big(\frac{\gamma_j^*}{a_j^*} > 0\big) + I\big(\frac{\gamma_j^*}{a_j^*} \leq 0\big) \right) > 0,$$

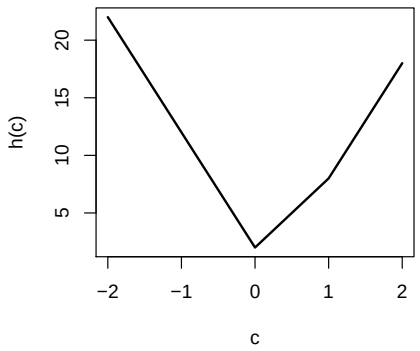

Figure 4.2: Function $h(c) = \sum_{j=1}^{J} |\gamma_j^* - a_j^* c|$, where $J = 10$, $a_j^* = 1$ for all j, $\gamma_j^* = 0$ and 1 for $j = 1, ..., 8$ and $j = 9, 10$, respectively. The minimal value of $h(c)$ is achieved when $c = 0$.

where $I(\cdot)$ is the indicator function.

Consider a special case when $a_j^* = 1$ for all j, i.e., the measurement model is a one-parameter logistic model when there is no DIF. Then according to Proposition IV.1, the ML1 condition holds, if and only if

$$\sum_{j=1}^{J} I(\gamma_j^* \geq 0) > \sum_{j=1}^{J} I(\gamma_j^* < 0)$$

and

$$\sum_{j=1}^{J} I(\gamma_j^* \leq 0) > \sum_{j=1}^{J} I(\gamma_j^* > 0),$$

i.e., the median of γ_1^*, ..., γ_J^* is 0. These inequalities hold for the example in Figure 4.2, where $\sum_{j=1}^{J} I(\gamma_j^* \geq 0) = 10$, $\sum_{j=1}^{J} I(\gamma_j^* < 0) = 0$, $\sum_{j=1}^{J} I(\gamma_j^* \leq 0) = 8$ and $\sum_{j=1}^{J} I(\gamma_j^* > 0) = 2$.

Corollary IV.2. *Condition* (4.5) *holds if*

$$\sum_{j=1}^{J} |a_j^*| I(\gamma_j^* = 0) > \max \left\{ \sum_{j=1}^{J} |a_j^*| I(\gamma_j^*/a_j^* < 0), \sum_{j=1}^{J} |a_j^*| I(\gamma_j^*/a_j^* > 0) \right\}.$$

Corollary IV.2 can be derived from Proposition IV.1 that provides a sufficient condition for the ML1 condition, which suggests that the ML1 condition holds when $\gamma_j^* = 0$ for a sufficient number of items.

Parameter Estimation

We now propose a procedure for estimating the model under the ML1 condition. This procedure is described in Algorithm 3 below.

Algorithm 3: ML1 Algorithm for Binary Group DIF Detection

Step 1: Solve the following MML estimation problem

$$\tilde{\Xi} = \arg\max_{\Xi} \log L(\Xi), \quad s.t. \quad \gamma_1 = 0. \tag{4.7}$$

Step 2: Solve the optimization problem

$$\hat{c} = \arg\min_{c} \sum_{j=1}^{J} |\tilde{\gamma}_j - \tilde{a}_j c| \tag{4.8}$$

Output: The ML1 estimate $\hat{\gamma}_j = \tilde{\gamma}_j - \tilde{a}_j \hat{c}$, $\hat{\mu} = \tilde{\mu} + \hat{c}$, $\hat{\alpha}_j = \tilde{\alpha}_j$, $\hat{d}_j = \tilde{d}_j$.

We provide some remarks about these steps. The estimator (4.7) in Step 1 can be viewed as the MML estimator of the model, treating item 1 as an anchor item. We emphasize that the constraint $\gamma_1 = 0$ in Step 1 is an arbitrary but mathematically

convenient constraint for ensuring the estimability of the model when solving (4.8). It does not require item 1 to be truely a DIF-free item. This constraint can be replaced by any equivalent constraint, for example, $\gamma_2 = 0$, while not affecting the final estimation result. Step 2 finds the transformation that leads to the ML1 solution among all the models that are equivalent to the estimated model from Step 1. The optimization problem (4.8) is convex that takes the same form as the Least Absolute Deviations (LAD) objective function in median regression (*Koenker*, 2005). Consequently, it can be solved using standard statistical packages/software for quantile regression. In particular, the R package "*quantreg*" (*Koenker et al.*, 2018) is used in our simulation study and real data analysis.

The ML1 condition (4.5), together with some additional regularity conditions, guarantees the consistency of the above ML1 estimator. That is, $\hat{\Xi}$ will converge to Ξ^* as the sample size N grows to infinity. This result is formalized in Theorem IV.3, with its proof given in the Appendix C.

Theorem IV.3. *Let* $\Xi^* = \{\mu^*, \gamma_j^*, d_j^*, a_j^*, j = 1, ..., J\}$ *be the true model parameters, and* $\Xi^\dagger = \{\mu^\dagger, \gamma_j^\dagger, d_j^\dagger, a_j^\dagger, j = 1, ..., J\}$ *be the true parameter values of the equivalent model with constraint* $\gamma_1^\dagger = 0$. *Assume this equivalent model satisfies the standard regularity conditions in Theorem 5.14 of van der Vaart (2000) that concerns the consistency of maximum likelihood estimation. Further, assume that the ML1 condition (4.5) holds. Then* $|\hat{\mu} - \mu^*| = o_P(1), |\hat{\gamma}_j - \gamma_j^*| = o_P(1)$, $|\hat{a}_j - a_j^*| = o_P(1)$, *and* $|\hat{d}_j - d_j^*| = o_P(1)$ *as* $N \to \infty$.

We further discuss the connection between the proposed estimator and the regularized estimator (4.4). Note that $\hat{\Xi}$ is the one with the smallest $\sum_{j=1}^{J} |\gamma_j|$ among all

equivalent estimators that maximize the likelihood function (4.3). When the solution path of (4.4) is smooth and the solution to the ML1 problem (4.8) is unique, it is easy to see that $\hat{\Xi}$ is the limit of $\hat{\Xi}^{\lambda}$ when the positive tuning parameter λ converges to zero. In other words, the proposed estimator can be viewed as a limiting version of the LASSO estimator (4.4). According to Theorem IV.3, this limiting version of the LASSO estimator is sufficient for the consistent estimation of the true model.

Note that the consistency in Theorem IV.3 further implies that the DIF items can be consistently selected by a simple hard-thresholding approach. Similar hard-thresholding methods perform well for variable selection in regression models (*Meinshausen and Yu*, 2009). Given our ML1 estimate $\hat{\Xi}$ and a pre-specified threshold δ, this hard-thresholding method classifies the items for which $|\hat{\gamma}_j| \leq \delta$ as the non-DIF items and the rest as DIF items. Theorem IV.4 below describes the result on selection consistency.

Theorem IV.4. *Let* $\{\hat{\gamma}_1, ..., \hat{\gamma}_J\}$ *be estimators of the DIF parameters returned by Algorithm 3. For any fixed* δ *satisfying* $0 < \delta < \min\{|\gamma_1^*|, ..., |\gamma_J^*|\}$*, the probability*

$$P(1_{\{|\hat{\gamma}_j| \leq \delta\}} = 1_{\{\gamma_j^* = 0\}})$$

converges to 1, as the sample size N *goes to infinity, for all* $j = 1, ..., J$.

In practice, the value of δ can be chosen by BIC, similar to the choice of λ for the LASSO procedure in *Belzak and Bauer* (2020). That is, we consider a sequence of δ values. For each δ, the hard-thresholding method is applied to obtain an estimated set of non-DIF items. We then refit the model with the γ_j parameters fixed to zero

for j in this estimated non-DIF item set, and compute the corresponding BIC value. We choose the δ that yields the smallest BIC values. The final classification of the items is given by the results based on the chosen δ.

Statistical Inference

The statistical inference of individual γ_j parameters is of particular interest in the DIF analysis. In fact, without the bias brought by the regularization tuning parameter, we can draw valid statistical inference on the DIF parameters γ_j.

Note that the uncertainty of $\hat{\gamma}_j$ is inherited from $\tilde{\Xi}$, where $\sqrt{N}(\tilde{\Xi} - \Xi^{\dagger})$ asymptotically follows a mean-zero multivariate normal distribution[1] by the large-sample theory for maximum likelihood estimation; see Appendix C for more details. We denote this multivariate normal distribution by $N(\mathbf{0}, \Sigma^{*})$, where a consistent estimator of Σ^{*}, denoted by $\hat{\Sigma}_N$, can be obtained based on the marginal likelihood. We define a function

$$G_j(\Xi) = \gamma_j - a_j \times \arg\min_{c} \sum_{l=1}^{J} |\gamma_l - a_l c|,$$

where $\Xi = \{\mu, a_l, d_l, \gamma_l, l = 1, ..., J\}$. Note that the function G_j maps an arbitrary parameter vector of the model to the γ_j parameter of the equivalent ML1 parameter vector. To draw statistical inference, we need the distribution of

$$\hat{\gamma}_j - \gamma_j^{*} = G_j(\tilde{\Xi}) - G_j(\Xi^{\dagger}).$$

By the asymptotic distribution of $\sqrt{N}(\tilde{\Xi} - \Xi^{\dagger})$, we know that the distribution of

[1]Note that this is a degenerated multivariate normal distribution since $\tilde{\gamma}_1 = \gamma_1^{\dagger} = 0$.

$G_j(\tilde{\Xi}) - G_j(\Xi^\dagger)$ can be approximated by that of $G_j(\Xi^\dagger + \mathbf{Z}/\sqrt{N}) - G_j(\Xi^\dagger)$, and the latter can be further approximated by $G_j(\tilde{\Xi} + \mathbf{Z}/\sqrt{N}) - G_j(\tilde{\Xi})$, where $\mathbf{Z}$ follows a normal distribution $N(\mathbf{0}, \hat{\Sigma}_N)$. Therefore, although function G_j does not have an analytic form, we can approximate the distribution of $\hat{\gamma}_j - \gamma_j^*$ by Monte Carlo simulation. We summarize this procedure in Algorithm 4 below.

Algorithm 4: Inference

Input: The number of Monte Carlo samples M and significance level α.

Step 1: Generate M i.i.d. samples from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\hat{\Sigma}_N$. We denote these samples as $\mathbf{Z}_1$, ..., $\mathbf{Z}_M$.

Step 2: Obtain $e_m = G_j(\tilde{\Xi} + \mathbf{Z}_m/\sqrt{N}) - G_j(\tilde{\Xi})$, for $m = 1, ..., M$.

Step 3: Obtain the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical distribution of $(e_1, ..., e_M)$, denoted by $q_{\alpha/2}$ and $q_{1-\alpha/2}$, respectively.

Output: Level $1-\alpha$ confidence interval for γ_j^* is given by $(\hat{\gamma}_j - q_{1-\alpha/2}, \hat{\gamma}_j - q_{\alpha/2})$. In addition, the p-value for a two-sided test of $\gamma_j^* = 0$ is given by

$$\frac{\sum_{i=1}^{M} 1_{\{|e_i| > |\hat{\gamma}_j|\}}}{M}.$$

Algorithm 4 only involves sampling from a multivariate normal distribution and solving a convex optimization problem based on the LAD objective function, both of which are computationally efficient. The value of M is set to 10,000 in our simulation study and 50,000 in the real data example below.

The p-values can be used to control the type-I error, i.e., the probability of mistakenly detecting a non-DIF item as a DIF one. To control the item-specific type-I errors to be below a pre-specified threshold α (e.g., $\alpha = 0.05$), we detect the

items for which the corresponding p-values are below α. Besides the type-I error, we may also consider the False Discovery Rate (FDR) for DIF detection (*Bauer et al.*, 2020) to account for multiple comparisons, where the FDR is defined as the expected ratio of the number of non-DIF items to the total number of detections. To control the FDR, the Benjamini-Hochberg (B-H) procedure (*Benjamini and Hochberg*, 1995) can be employed to the p-values.

4.2.4 Simulation Study

Simulation studies were conducted to evaluate the performance of the proposed method and compare it with the likelihood ratio test (LRT) method (*Thissen*, 1988) and the LASSO method (*Bauer et al.*, 2020). Note that the LRT method requires a known anchor item set. Correctly specified anchor item sets with different sizes will be considered when applying the LRT method.

In the simulation, we set the number of items $J = 25$, and consider two settings for the sample sizes, $N = 500$, and 1000. The parameters of the true model are set as follows. First, the discrimination parameters are set between 1 and 2, and the easiness parameters are set between -1 and 1. Their true values are given in Table 4.1. The observations are split into groups of equal sizes, indicated by $x_i = 0$, and 1. The parameter μ in the structural model is set to 0.5, so that the latent trait distribution is standard normal and $N(0.5, 1)$ for the reference and focal groups, respectively. We consider two settings for the DIF parameters, one with smaller absolute DIF parameter values, and the other with larger absolute DIF parameter values. Their true values are given in Table 4.1. For both sets of the DIF parameters,

the ML1 condition is satisfied. The combinations of settings for the sample sizes and DIF parameters lead to four settings in total. For each setting, 100 independent datasets are generated.

Table 4.1: Discrimination, easiness and DIF parameter values used in the simulation studies.

Item number	a_j	d_j	γ_j (small DIF)	γ_j (large DIF)
1	1.30	0.80	0.00	0.00
2	1.40	0.20	0.00	0.00
3	1.50	-0.40	0.00	0.00
4	1.70	-1.00	0.00	0.00
5	1.60	1.00	0.00	0.00
6	1.30	0.80	0.00	0.00
7	1.40	0.20	0.00	0.00
8	1.50	-0.40	0.00	0.00
9	1.70	-1.00	0.00	0.00
10	1.60	1.00	0.00	0.00
11	1.30	0.80	0.00	0.00
12	1.40	0.20	0.00	0.00
13	1.50	-0.40	0.00	0.00
14	1.70	-1.00	0.00	0.00
15	1.60	1.00	0.00	0.00
16	1.30	0.80	-0.60	-1.00
17	1.40	0.20	0.60	1.30
18	1.50	-0.40	-0.65	-0.90
19	1.70	-1.00	0.70	1.20
20	1.60	1.00	0.65	1.00
21	1.30	0.80	-0.60	-1.00
22	1.40	0.20	0.60	1.30
23	1.50	-0.40	-0.65	-0.90
24	1.70	-1.00	0.70	1.20
25	1.60	1.00	0.65	1.00

We first evaluate the accuracy of the proposed estimator $\hat{\Xi}$ given by Algorithm 3. Table 4.2 shows the mean squared errors (MSE) for μ and the average MSEs for

a_js, d_js, and γ_js that are obtained by averaging the corresponding MSEs over the J items. As we can see, these MSEs and average MSEs are small in magnitude and decrease as the sample size increases under each setting. This observation aligns with our consistency result in Theorem IV.3.

Table 4.2: Average mean squared errors of the estimated parameters in the simulation studies. Mean squared errors are first evaluated by averaging out of 100 replications and then averaged across 25 items to obtain the average mean squared errors for $\boldsymbol{a}$, $\boldsymbol{d}$ and $\boldsymbol{\gamma}$. The mean squared errors for μ is presented.

	Small DIF		Large DIF	
	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
$\boldsymbol{a}$	0.034	0.017	0.032	0.017
$\boldsymbol{d}$	0.035	0.018	0.037	0.018
$\boldsymbol{\gamma}$	0.058	0.027	0.061	0.028
μ	0.023	0.011	0.023	0.011

We then compare the proposed method and the LRT method in terms of their performances on statistical inference. Specifically, we focus on whether FDR can be controlled when applying the B-H procedure to the p-values obtained from the two methods. The comparison results are given in Table 4.3. As we can see, FDR is controlled to be below the targeted level for the proposed method and the LRT method with 1, 5, and 10 anchor items, under all settings.

When anchor items are known, the standard error can be computed for each estimated γ_j and thus the corresponding Wald interval can be constructed. We compare the coverage rates of the confidence intervals given by Algorithm 4 and the Wald intervals that are based on five anchor items. The results are shown in Figure 4.3. We see that the coverage rates from both methods are comparable across

Table 4.3: Comparison of the FDR of the proposed p-value based method and the LRT method with 1, 5 and 10 anchor items respectively at the FDR control of 5%. The values are averaged out of 100 replications.

	$N = 500$		$N = 1000$	
	Small DIF	Large DIF	Small DIF	Large DIF
p-value based method	0.034	0.029	0.027	0.025
LRT (1 anchor item)	0.029	0.039	0.031	0.034
LRT (5 anchor items)	0.021	0.026	0.027	0.027
LRT (10 anchor items)	0.011	0.014	0.018	0.017

all settings and are close to the 95% targeted level. Note that these coverage rates are calculated based on only 100 replicated datasets, which may be slightly affected by the Monte Carlo errors.

Furthermore, we compare the proposed hard-thresholding procedure with the LASSO procedure, in terms of the accuracy in the detection of DIF items. For the hard-thresholding procedure, 20 values of threshold δ are considered that are equally spaced in $[0, 0.9]$ and $[0, 1.8]$ under the small and large DIF settings, respectively. For the LASSO procedure, 20 values are considered for the tuning parameter λ that are equally spaced in $[0.008, 0.12]$ and $[0.01, 0.15]$ under the small and large DIF settings, respectively. The optimal values of λ and δ are chosen by BIC, which yield the selection of DIF items for the two procedures, respectively. The selection accuracy is evaluated by two metrics, the true positive rate (TPR) and false positive rate (FPR), where the TPR is the expected ratio of the number of correctly detected DIF items to the number of truely DIF items, and the FPR is the expected ratio of the number of falsely detected DIF items to the number of non-DIF items. Table 4.4 shows the TPR and FPR that are estimated based on the 100 simulation replications. As we can

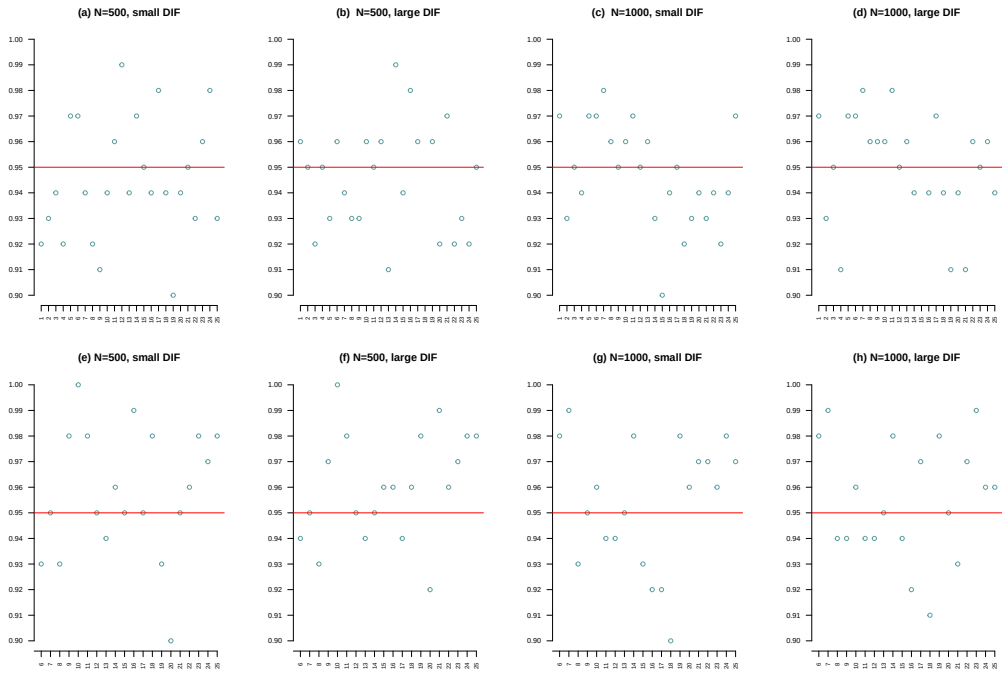

Figure 4.3: Scatter plots of the coverage rates of the 95% confidence intervals for γ_j^*'s. x-axes and y-axes are labeled with item numbers and coverage rates respectively. Panels (a) - (d) correspond to our proposed method, and panels (e) - (h) correspond to the Wald intervals constructed with five anchor items.

see, the two methods have comparable TPR and FPR under all settings. It is worth noting that the hard-thresholding method is computationally much faster than the LASSO method, since it does not involve maximizing a regularized likelihood under different tuning parameters.

Finally, we compare the detection power of different methods based on the receiver operating characteristic (ROC) curves. For a given method, an ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings. More specifically, ROC curves are constructed for the hard-thresholding and the LASSO methods by varying the corresponding

Table 4.4: Comparison of the TPR and FPR of the proposed hard-thresholding method and the LASSO method. The optimal thresholds for the hard-thresholding method and the optimal penalties for the LASSO method are selected using the BIC criteria. The results are averaged out of 100 replications.

		$N = 500$		$N = 1000$	
		Small DIF	Large DIF	Small DIF	Large DIF
TPR	Hard-thresholing	0.598	0.967	0.855	1.000
	LASSO	0.607	0.962	0.863	1.000
FPR	Hard-thresholding	0.019	0.015	0.015	0.009
	LASSO	0.015	0.011	0.025	0.006

tuning parameters δ and λ. ROC curves are also constructed by thresholding the p-values from the proposed method and the LRT method with 1, 5, and 10 anchor items, respectively. Note that for the LRT method, the TPR and FPR are calculated based on the non-anchor items. For each method, an average ROC curve is obtained based on the 100 replications, for which the area under the ROC curve (AUC) is calculated. A larger AUC value indicates the better detection power. The AUC values for different methods across our simulation settings are given in Table 4.5. According to the AUC values, the two proposed procedures (i.e., thresholding the p-values from Algorithm 4 and the hard-thresholding procedure) and the LRT method with 10 anchor items have similar detection power and perform better than the rest. That is, without knowing any anchor items, the proposed procedures perform better than the LRT method that knows 1 or 5 anchor items, and perform similarly as the LRT method that knows 10 anchor items. The superior performance of the proposed procedures is brought by the use of the ML1 condition, which identifies the model parameters using information from all the items. Based on the AUC values, we also

see that the LASSO procedure performs similarly as the proposed procedures under the large DIF settings, but is less accurate under the small DIF settings.

Table 4.5: Comparison of AUC of the proposed p-value based method, the hard-thresholding method, the LASSO method and the LRT method with 1, 5 and 10 anchor items respectively.

	$N = 500$		$N = 1000$	
	Small DIF	Large DIF	Small DIF	Large DIF
p-value based	0.937	0.990	0.986	0.992
Hard-thresholding	0.935	0.995	0.983	0.999
LASSO	0.918	0.992	0.956	0.999
LRT (1 anchor item)	0.840	0.957	0.907	0.986
LRT (5 anchor items)	0.931	0.988	0.979	0.990
LRT (10 anchor items)	0.950	0.992	0.979	0.989

4.2.5 Application to EPQ-R Data

DIF methods have been commonly used for assessing the measurement invariance of personality tests (e.g., *Escorial and Navas* (2007), *Millsap* (2012), *Thissen et al.* (1986)). In this section, we apply the proposed method to the Eysenck Personality Questionnaire-Revised (EPQ-R, *Eysenck et al.* (1985)), a personality test that has been intensively studied and received applications worldwide (*Fetvadjiev and van de Vijver*, 2015). The EPQ-R has three scales that measure the Psychoticism (P), Neuroticism (N) and Extraversion (E) personality traits, respectively. We analyze the long forms of the three personality scales that consist of 32, 24, and 23 items, respectively. Each item has binary responses of "yes" and "no" that are indicated by 1 and 0, respectively. This analysis is based on data of an EPQ-R study collected from 1432 participants in the United Kingdom. Among these participants, 823 are

females and 609 are males. Females and males are indicated by $x_i = 0$ and 1, respectively. We study the DIF caused by gender. The three scales are analyzed separately using the proposed methods.

The results are shown through Tables 4.6 - 4.8, and Figure 4.4. Specifically, Tables 4.6 - 4.8 present the p-values for testing $\gamma_j = 0$ and the detection results for the P, E, N scales, respectively. For each table, the items are ordered by the p-values in an increasing order. The items indicated by "F" are the ones detected by the B-H procedure with FDR level 0.05, and those indicated by "H" are the ones detected by the hard-thresholding procedure whose threshold δ is chosen by BIC. The item IDs are consistent with those in Appendix 1 of *Eysenck et al.* (1985), where the item descriptions are given. The three panels of Figure 4.4 further give the point estimate and confidence interval for each γ_j parameter, for the three scales, respectively. Under the current model parameterization, a positive DIF parameter means that a male participant is more likely to answer "yes" to the item than a female participant, given that they have the same personality trait level. We note that the absolute values of $\hat{\gamma}_j$ are all below 1, suggesting that there are no items with very large gender-related DIF effects.

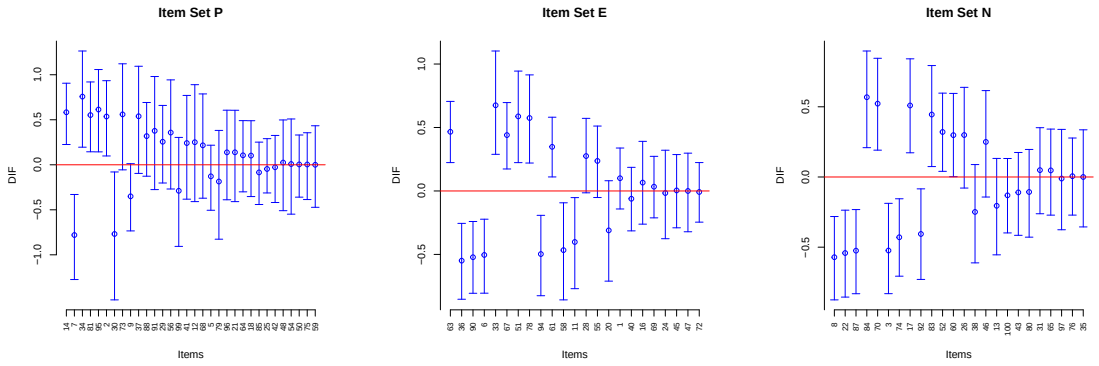

Figure 4.4: Plots of 95% confidence intervals for the DIF parameters $\gamma_j^{*'}s$ on scale P, N, and E data sets. The red horizontal lines denote $\gamma = 0$. Items are arranged according to the increasing p-values.

Table 4.6: P-values for testing $\gamma_j^* = 0$ for items in P scale. Note that the items are ordered in increasing p-values. Items selected by the B-H procedure with FDR control at 5% and the proposed hard-thresholding method are identified using "F" and "H", respectively, besides the item numbers.

Item	14 FH	7 FH	34 H	81 H	95 H	2 H	30 H	73 H
p-value	0.0014	0.0015	0.0057	0.0061	0.0104	0.0140	0.0364	0.0619
Item	9	37 H	88	91	29	56	99	41
p-value	0.0625	0.0681	0.1235	0.2217	0.2304	0.2442	0.3389	0.3780
Item	12	68	5	79	96	21	64	18
p-value	0.4389	0.4557	0.4567	0.5187	0.5515	0.5529	0.5819	0.5888
Item	85	25	42	48	54	50	75	59
p-value	0.6080	0.7527	0.8441	0.8787	0.9447	0.9528	0.9559	0.9616

Table 4.7: P-values for testing $\gamma_j^* = 0$ for items in E scale. Note that the items are ordered in increasing p-values. Items selected by the B-H procedure with FDR control at 5% and the proposed hard-thresholding procedure are identified using "F" and "H", respectively, besides the item numbers.

Item	63 FH	36 FH	90 FH	6 FH	33 FH	67 FH	51 FH	78 FH
p-value	0.0000	0.0004	0.0006	0.0011	0.0013	0.0013	0.0016	0.0019
Item	94 FH	61 FH	58 FH	11 H	28	55	20	1
p-value	0.0031	0.0051	0.0199	0.0310	0.0644	0.0958	0.1278	0.4073
Item	40	16	69	24	45	47	72	
p-value	0.6185	0.6439	0.7819	0.8371	0.9291	0.9364	0.9391	

Table 4.8: P-values for testing $\gamma_j^* = 0$ for items in N scale. Note that the items are ordered in increasing p-values. Items selected by the B-H procedure with FDR control at 5% and the proposed hard-thresholding procedure are identified using "F" and "H", respectively, besides the item numbers.

Item	8 FH	22 FH	87 FH	84 FH	70 FH	3 FH	74 FH	17 FH
p-value	0.0004	0.0006	0.0007	0.0014	0.0016	0.0026	0.0026	0.0037
Item	92 FH	83 FH	52	60	26	38	46	13
p-value	0.0130	0.0152	0.0264	0.0487	0.0994	0.1553	0.1856	0.2337
Item	100	43	80	31	65	97	76	35
p-value	0.3365	0.4417	0.4694	0.7116	0.7376	0.9220	0.9531	0.9550

From Tables 4.6 - 4.8, we see that all the three scales have some items whose p-values are close to zero, suggesting gender DIF may exists across the three scales. We list some example items with the smallest p-values for each of the three scales. For the P scale, item 14 "Do you dislike people who don't know how to behave themselves?", 7 "Would being in debt worry you?" and 34 "Do you have enemies who want to harm you?" have the smallest p-values. As shown in Figure 4.4, the estimated DIF parameters are positive for items 14 and 34 and negative for item 7. For the E scale, items 63 "Do you nearly always have a 'ready answer' when people talk to you?", 36 "Do you have many friends?" and 90 "Do you like plenty of bustle and excitement around you?" have the smallest p-values, where item 63 has positive $\hat{\gamma}_j$s, and items 36 and 90 have negative $\hat{\gamma}_j$. For the N scale, the top three items are 8 "Do you ever feel 'just miserable' for no reason?", 22 "Are your feelings easily hurt?" and 87 "Are you easily hurt when people find fault with you or the work you do?", where $\hat{\gamma}_j$ are negative for all the items.

From Tables 4.6 through 4.8, we see that the selection based on the B-H procedure with FDR level 0.05 is more conservative than the hard-thresholding approach

since the number of items selected by the B-H procedure tend to be smaller than that selected by the hard-thresholding approach. More precisely, the B-H procedure detects 2, 11, and 10 items as DIF items from the P, E, and N scales, respectively, and the hard-threhsolding procedure selects 9, 12, and 10 items, respectively. We remark that the hard-thresholding procedure is based on the point estimate of the items and thus does not always select the items with the smallest p-values. For example, it detects item 37 as a DIF item but does not detects item 9, though item 37 has a larger p-value. This is because item 37 has a larger absolute value of $\hat{\gamma}_j$ than item 9.

4.3 Multiple Group DIF Detection and Inference

We extend the work from Section 4.2 to handle multiple group DIF analysis.

4.3.1 Model Setup

Here we introduce an additional parameter vector $\boldsymbol{\gamma}_j$ into (4.1) to incorporate the potential DIF effect in item j brought by a p-dimensional categorical background feature $\boldsymbol{x}_i$. $\boldsymbol{x}_i$ segregates the individuals into different groups so that subjects within each group have the similar proficiency distribution and suffer the same level of DIF from each item. In specific, the p-dimensional categorical feature vector for individual i is denoted as $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$, with one dimension taking value of one and all other dimensions taking the value of zero. We make the assumption that the log odds ratio of a correct response of individual i to item j depends linearly on the DIF effect $\boldsymbol{\gamma}_j^T \boldsymbol{x}_i$.

In this case, the item response function is modeled as

$$P(Y_{ij} = 1|\theta_i = \theta, \boldsymbol{x}_i) = \frac{\exp\left\{a_j\theta + d_j + \boldsymbol{\gamma}_j^T \boldsymbol{x}_i\right\}}{1 + \exp\left\{a_j\theta + d_j + \boldsymbol{\gamma}_j^T \boldsymbol{x}_i\right\}}, \tag{4.9}$$

where we assume $\theta_i \mid \boldsymbol{x}_i \sim N(\mu_1 x_{i1} + ... + \mu_p x_{ip}, 1_{(\boldsymbol{x}_i=\boldsymbol{0})} + \sigma_1^2 x_{i1} + ... + \sigma_p^2 x_{ip})$, i.e. group k is assumed to have a mean proficiency distribution of μ_k and a group specific variance of σ_k^2. The indicator $1_{(\boldsymbol{x}_i=\boldsymbol{0})}$ is included for model identifiability purpose. We point out that this formulation corresponds to using a multiple group categorical feature vector with $(p + 1)$ categories, with one category captured by the baseline when $\boldsymbol{x}_i = \boldsymbol{0}$, in which case the latent proficiency distribution is fixed to be a standard normal. As such, each group k would be characterized by a unique mean μ_k and variance σ_k^2, except for the baseline group whose mean and variance are fixed to be zero and one respectively.

However, similar to the binary group case, model (4.9) is not identifiable if no additional constraints are imposed. Though $\boldsymbol{a}, \boldsymbol{d}$ and $\boldsymbol{\sigma}$ can be uniquely determined, $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are indeterminate. To see this, consider $\eta_{ij} = a_j\theta_i + d_j + \boldsymbol{\gamma}_j^T \boldsymbol{x}_i$. The expectation of η_{ij} would remain invariant, if we replace μ_k with $\mu_k + c_k$ and γ_{jk} with $\gamma_{jk} - a_j c_k$ for $c_k \in \mathbb{R}$ and $j = 1, ..., J$. Consequently, model (4.9) would remain invariant in these parameter settings, therefore not identifiable from response data.

4.3.2 Proposed Method

Let $\mu_k^*, \sigma_k^*, a_j^*, d_j^*$ and γ_{jk}^* denote the true model parameters. In the same spirit of the binary group case, we propose the following identifiability multiple group minimal L_1 condition to find a sparse and unique representation of $\{\gamma_{jk}^* : j = 1, ...J, k =$

$1, ..., p\}$,

$$\sum_{k=1}^{p}\sum_{j=1}^{J}|\gamma_{jk}^{*}| < \sum_{k=1}^{p}\sum_{j=1}^{J}|\gamma_{jk}^{*} - a_{j}^{*}c_{k}|, \qquad \forall c_{k} \neq 0. \tag{4.10}$$

Note that we can view (4.10) separately for each $k = 1, ..., p$ dimension. Condition (4.10) then implies that each $\boldsymbol{\gamma}_{k}^{*} = \{\gamma_{jk}^{*}, j = 1, ..., J\}$ will have the smallest L_1 norm. Condition (4.10) tends to hold when $\boldsymbol{\gamma}_{k}^{*}$ is sparse since L_1 norm measures the sparsity level of a vector (*Tibshirani*, 1996). Similar to Proposition IV.1, we also provide the necessary and sufficient condition in Proposition IV.5 below for Condition (4.10) to hold. The proof for Proposition IV.5 is given in Appendix C.

Proposition IV.5. *Condition* (4.10) *holds if and only if for any* $k = 1, ..., p$,

$$\sum_{j=1}^{J}|a_{j}^{*}|\left(-I\big(\frac{\gamma_{jk}^{*}}{a_{j}^{*}} \geq 0\big) + I\big(\frac{\gamma_{jk}^{*}}{a_{j}^{*}} < 0\big)\right) < 0$$

and

$$\sum_{j=1}^{J}|a_{j}^{*}|\left(-I\big(\frac{\gamma_{jk}^{*}}{a_{j}^{*}} > 0\big) + I\big(\frac{\gamma_{jk}^{*}}{a_{j}^{*}} \leq 0\big)\right) > 0,$$

where $I(\cdot)$ *is the indicator function.*

The following corollary can be derived directly from Proposition IV.5 that provides a sufficient condition for the ML1 condition to hold. We remark that the ML1 condition will hold when $\gamma_{j,k}^{*} = 0$ for a sufficient number of items for each $k = 1, ..., p$.

Corollary IV.6. *Condition* (4.10) *holds if for* $k = 1, ..., p$,

$$\sum_{j=1}^{J}|a_{j}^{*}|I(\gamma_{jk}^{*} = 0) > \max\left\{\sum_{j=1}^{J}|a_{j}^{*}|I(\gamma_{jk}^{*}/a_{j}^{*} < 0), \sum_{j=1}^{J}|a_{j}^{*}|I(\gamma_{jk}^{*}/a_{j}^{*} > 0)\right\}.$$

Estimation

We give a procedure for parameter estimation under Condition (4.10). The procedure consists of two steps and is detailed in Algorithm 5.

Algorithm 5: ML1 Algorithm for Multiple Group DIF Detection

Step 1: Maximizing marginal log-likelihood

$$\tilde{\Psi} = \arg\max_{\Psi} \log L(\Psi), \quad s.t. \quad \{\gamma_{1k} : k = 1, ..., p\} = \mathbf{0}, \tag{4.11}$$

where $\Psi = \{\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{a}, \boldsymbol{d}, \boldsymbol{\gamma}\}$.

Step 2: For each $k = 1, ..., p$, solve the optimization problem

$$\hat{c}_k = \arg\min_{c_k} \sum_{j=1}^{J} |\tilde{\gamma}_{jk} - \tilde{a}_j c_k| \tag{4.12}$$

Output: The ML1 estimates $\hat{\gamma}_{jk} = \tilde{\gamma}_{jk} - \tilde{a}_j \hat{c}_k$, $\hat{\mu}_k = \tilde{\mu}_k + \hat{c}_k$, $\hat{a}_j = \tilde{a}_j$, $\hat{d}_j = \tilde{d}_j$ and $\hat{\sigma}_k = \tilde{\sigma}_k$.

Similarly, we point out that $\boldsymbol{\gamma}_1 = \mathbf{0}$ constraint is completely arbitrary, irregardless of whether item 1 exhibits DIF. It simply provides a mathematically convenient way to ensure estimability of parameters. Selecting any item $j \in \{1, ..., J\}$ as the anchor item will not affect the final output of Algorithm 5. The optimization problem (4.12) can be solved using standard median regression approaches.

Similar to Theorem IV.3, Condition (4.10), together with some additional regularity conditions, guarantees the consistency of the estimators returned by Algorithm 5. Consequently, DIF detection consistency can be guaranteed. This is formalized in Proposition IV.7.

Proposition IV.7. *Let* $\Psi^* = \{\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*, \boldsymbol{a}^*, \boldsymbol{d}^*, \boldsymbol{\gamma}^*\}$ *be the true model parameters and let* $\hat{\Psi} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{d}}, \hat{\boldsymbol{\gamma}}\}$ *be estimators returned by Algorithm 5. Assume the model satisfies the standard regularity conditions in Theorem 5.14 of van der Vaart (2000) and that Condition (4.10) holds. Then for* $j = 1, .., J, k = 1, ..., p$, $|\hat{\mu}_k - \mu_k^*| = o_P(1), |\hat{\sigma}_k - \sigma_k^*| = o_P(1)$, $|\hat{a}_j - a_j^*| = o_P(1), |\hat{d}_j - d_j^*| = o_P(1)$, $|\hat{\gamma}_{jk} - \gamma_{jk}^*| = o_P(1)$, *as* $N \to \infty$. *Furthermore, for any* δ *satisfying* $0 < \delta < \min\{|\gamma_{j,k}^*| : \gamma_{j,k}^* \neq 0\}$, *the probability*

$$P(1_{\{|\hat{\gamma}_{jk}| \leq \delta\}} = 1_{\{\gamma_{jk}^* = 0\}})$$

converges to 1, as $N \to \infty$, *for all* $j = 1, ..., J$ *and* $k = 1, ..., p$.

Clustering

When dimension p is high and when $\boldsymbol{\gamma}_j$ is dense, it is difficult to understand the underlying sources driving DIF effects based on the estimated $\hat{\boldsymbol{\gamma}}_j$ alone. Another potential interest, when performing high dimensional DIF analysis, is to study the relationships amongst the groups. To achieve these goals, we propose to apply hierarchical clustering on $(p+1)$ groups based on the estimated $\hat{\boldsymbol{\gamma}}$ values. The explicit procedure is summarized in Algorithm 6 below.

We remark that hierarchical clustering is particularly helpful in understanding the relationships amongst the groups here because it displays clustering of all levels of granularity. Moreover, groups classified into the same cluster will have similar DIF patterns. Hence, DIF effects tend to vanish amongst groups within each clusters. This can help reduce bias brought by DIF effects for those cross-group studies that do not take DIF into account, if group comparisons are made within each clusters

Algorithm 6: Group Clustering Algorithm

Input: The estimated $\hat{\boldsymbol{\gamma}}$ from Algorithm 5.

Step 1: Assign each $k = 1, ..., p+1$ into a separate cluster.

Step 2: Evaluate the distance matrix $D \in \mathbb{R}^{(p+1)\times(p+1)}$ with

$$D_{i,k} = \Big\{ \sum_{j=1}^{J} (\hat{\gamma}_{ji} - \hat{\gamma}_{jk})^2 \Big\}^{\frac{1}{2}}.$$

Step 3: Group (i, k) into the same cluster for $(i, k) = \arg\min_{i=1,...,p+1, k=1,...,p+1} D_{i,k}$.

Step 4: Update the distance matrix D:

Step 4.1: Index (i, k) as i and remove the k'th column and row from D.

Step 4.2: Update $D_{i,l}$ for $l \neq i$ with the average Euclidean distance between clusters i and l.

Step 4: Repeat Steps 3-4 until a single cluster remains.

Output: A dendrogram, or a tree-like diagram recording the process of merging.

only. This can be a useful tool because many practitioners will simply ignore the DIF effects due to its additional complexity. We will also demonstrate its utility in our application study on the PISA 2018 data in Section 4.3.4.

Inference

For DIF detection, whether a particular item exhibits DIF is often of interest. For an item j, similar to the binary group case, the uncertainty of $\hat{\boldsymbol{\gamma}}_j$ stems from $\tilde{\Psi}$. Let $\Psi^\dagger = \{\boldsymbol{\mu}^\dagger, \boldsymbol{\sigma}^\dagger, \boldsymbol{a}^\dagger, \boldsymbol{d}^\dagger, \boldsymbol{\gamma}^\dagger\}$ denote the model parameters such that $\boldsymbol{\gamma}_1^\dagger = \mathbf{0}$. Note by large-

sample theory for maximum likelihood estimation, $\sqrt{N}(\tilde{\Psi} - \Psi^{\dagger}) \to N(\mathbf{0}, \Sigma^{*})$ as $N \to \infty$. In practice, we can apply the Louis identity (*Louis*, 1982) to obtain a consistent estimator $\hat{\Sigma}_N$, for Σ^{*}. Further note that by Condition (4.10) and Algorithm 5, we can explicitly express $\hat{\boldsymbol{\gamma}}_j$ as

$$\hat{\boldsymbol{\gamma}}_j = \tilde{\boldsymbol{\gamma}}_j - \tilde{a}_j \underset{\{c_1,\ldots,c_p\}}{\arg\min} \sum_{k=1}^{p} \sum_{l=1}^{J} \left| \tilde{\gamma}_{lk} - \tilde{a}_l c_k \right|.$$

Define a function

$$G_j(\Psi) = \boldsymbol{\gamma}_j - a_j \underset{\{c_1,\ldots,c_p\}}{\arg\min} \sum_{k=1}^{p} \sum_{l=1}^{J} \left| \gamma_{lk} - a_l c_k \right|.$$

To draw statistical inference for $\boldsymbol{\gamma}_j$, we need the distribution of

$$\hat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^{*} = G_j(\tilde{\Psi}) - G_j(\Psi^{\dagger}).$$

By the asymptotic distribution of $\sqrt{N}(\tilde{\Psi} - \Psi^{\dagger})$, we know that the distribution of $G_j(\tilde{\Psi}) - G_j(\Psi^{\dagger})$ can be approximated by that of $G_j(\Psi^{\dagger} + \mathbf{Z}/\sqrt{N}) - G_j(\Psi^{\dagger})$, and the latter can be further approximated by $G_j(\tilde{\Psi} + \mathbf{Z}/\sqrt{N}) - G_j(\tilde{\Psi})$, where $\mathbf{Z}$ follows a multivariate Gaussian distribution $N(\mathbf{0}, \Sigma^{*})$.

To test $H_0 : \boldsymbol{\gamma}_j = \mathbf{0}$, we suggest to use Hotelling test statistic T_j, specified as follows,

$$T_j = \hat{\boldsymbol{\gamma}}_j^{T} \hat{\Sigma}_{HMj}^{-1} \hat{\boldsymbol{\gamma}}_j.$$

where $\hat{\Sigma}_{HMj}$ is the sample covariance matrix of $\hat{\boldsymbol{\gamma}}_j$. Though $\hat{\boldsymbol{\gamma}}_j$ may not be strictly Gaussian due to the non-linear transformation G_j, Monte Carlo method easily applies and empirical null distribution of T_j can be generated, from which a p-value can be returned. The exact procedure is summarized in Algorithm 7 below.

Algorithm 7: Inference for Multiple Group DIF Detection

Input: The number of Monte Carlo samples M.

Step 1: Evaluate $\hat{\boldsymbol{\gamma}}_j = G_j(\tilde{\Psi})$.

Step 2: Generate M i.i.d. samples from $N(\mathbf{0}, \hat{\Sigma}_N)$. We denote these samples as $\mathbf{Z}_1$, ..., $\mathbf{Z}_M$.

Step 3: Obtain $\boldsymbol{e}_{jm} = G_j(\tilde{\Psi} + \mathbf{Z}_m/\sqrt{N}) - G_j(\tilde{\Psi})$, for $m = 1, ..., M$.

Step 4: Evaluate sample covariance matrix for $\boldsymbol{e}_j$, denote it by $\hat{\Sigma}_{HMj} \in \mathbb{R}^{p\times p}$.

Step 5: Compute the test statistic,

$$\hat{t}_j = \hat{\boldsymbol{\gamma}}_j^T \hat{\Sigma}_{HMj}^{-1} \hat{\boldsymbol{\gamma}}_j.$$

Step 6: Evaluate empirical null distribution for T_j consisting of $\{E_{jm} : E_{jm} = \boldsymbol{e}_{jm}^T \hat{\Sigma}_{HMj}^{-1} \boldsymbol{e}_{jm}, m = 1, ..., M\}$.

Step 7: Obtain the p-value,

$$p_j = \frac{\sum_{m=1}^{M} 1_{\{E_{jm} > \hat{t}_j\}}}{M}.$$

Output: The p-value p_j.

We remark that similar to the binary group case, Algorithm 7 only involves sampling from a multivariate normal distribution and solving a convex LAD objective function. Both steps are computationally efficient which enables fast inference. Fur-

thermore, p-values can be used to control the type-I error rate, and more strictly, the False Discovery Rate (FDR) for DIF detection (*Bauer et al.*, 2020) to account for multiple comparisons when jointly testing for multiple items. Benjamini-Hochberg (B-H) procedure (*Benjamini and Hochberg*, 1995) can also be applied here for better FDR control.

4.3.3 Simulation Study

Simulation studies are carried out to compare the performance of the proposed method and the traditional LRT method (*Thissen et al.*, 1986; *Thissen*, 1988; *Thissen et al.*, 1993). Correctly specified anchor sets with sizes one, three and five are considered for the LRT method.

For both methods, the total number of test items is set to be $J = 15$. Two sample size settings are considered with $N = 1000$ and 2000. The discrimination parameters $\boldsymbol{a}$ are set between 1 and 2, and the easiness parameters $\boldsymbol{d}$ are set between -1 and 1. The exact values are summarized in Table 4.9 below. A p-dimensional variable $\boldsymbol{x}$ is used to split the observations into $(p+1)$ groups of equal sizes, captured by $\boldsymbol{x}_i \in \{(0,0,...,0),(1,0,...,0),(0,1,...,0),...,(0,0...,0,1)\}$ for $i = 1,...,N$. Two group sizes, either five groups, i.e. $p = 4$, or ten groups, i.e. $p = 9$, are considered. The latent traits θ_i are sampled from the Gaussian distribution with the corresponding mean parameters $\boldsymbol{\mu} = \{0.5, 1, 0.75, 1.2\}$ and $\{0.5, 0.6, 0.7, 0.8, 0.9, 1, 0.75, 0.85, 1.1\}$, respectively, and standard deviation parameters $\boldsymbol{\sigma} = \{0.5, 0.7, 0.9, 1.1\}$ and $\{0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9, 1.1\}$, respectively. We consider a small DIF setting and a large DIF setting. In both settings, only the last three items are designed to exhibit DIF. In the

Table 4.9: True discrimination and easiness parameter values used in the simulation study for multiple group DIF analysis.

Item	a_j	d_j
1	1.30	0.80
2	1.40	0.20
3	1.50	-0.40
4	1.70	-1.00
5	1.60	1.00
6	1.30	0.80
7	1.40	0.20
8	1.50	-0.40
9	1.70	-1.00
10	1.60	1.00
11	1.30	0.80
12	1.40	0.20
13	1.50	-0.40
14	1.70	-1.00
15	1.60	1.00

small DIF setting, $\boldsymbol{\gamma}_{13} = \boldsymbol{\gamma}_{14} = \boldsymbol{\gamma}_{15} = \{-0.4, 0.4, -0.5, 0.4\}$ and $\{-0.4, 0.4, -0.5, 0.5, 0.35, -0.55, -0.45, 0.45, 0.55\}$, respectively. In the large DIF setting, $\boldsymbol{\gamma}_{13} = \boldsymbol{\gamma}_{14} = \boldsymbol{\gamma}_{15} = \{-1.2, 1, -0.8, 1.1\}$ and $\{-1.2, 1, -0.8, 1.1, 0.7, -0.7, -1, -0.9, 1\}$, respectively. This leads to eight different settings in total. 100 independent data sets are generated for each setting.

We first evaluate the estimation accuracy of the proposed estimators in Algorithm 5. Table 4.10 below summarizes the results. Mean squared errors (MSEs) are first evaluated by averaging out of 100 replications for all a_j, d_j, $\boldsymbol{\gamma}_j$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Then average MSEs are evaluated over the J items for a_j, d_j and $\boldsymbol{\gamma}_j$. We observe that these MSEs and average MSEs are small in magnitude, suggesting our proposed estimators work well. We also point out that estimation accuracy tends to decrease

as dimension p increases and sample size decreases.

Table 4.10: Average mean squared errors of the estimated parameters. Mean squared errors are first evaluated by averaging out of 100 replications and then averaged across 15 items to obtain the average mean squared errors for $\boldsymbol{a}$, $\boldsymbol{d}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Small and Large denote the small DIF setting and the large DIF setting, respectively.

	$p = 4$				$p = 9$			
	Small		Large		Small		Large	
	$N = 1000$	$N = 2000$	$N = 1000$	$N = 2000$	$N = 1000$	$N = 2000$	$N = 1000$	$N = 2000$
$\boldsymbol{a}$	0.030	0.013	0.031	0.013	0.043	0.019	0.041	0.020
$\boldsymbol{d}$	0.039	0.017	0.039	0.017	0.073	0.036	0.074	0.036
$\boldsymbol{\gamma}$	0.079	0.041	0.085	0.043	0.147	0.074	0.157	0.079
$\boldsymbol{\mu}$	0.015	0.006	0.015	0.006	0.039	0.015	0.038	0.015
$\boldsymbol{\sigma}$	0.006	0.003	0.007	0.003	0.010	0.005	0.010	0.005

We then compare the proposed method and the LRT method in terms of their performance on statistical inference. In particular, we compare the performance of the two methods in overall DIF detection for each items. For the proposed method, as detailed in Algorithm 7, a Hotelling test statistic is introduced to test whether a given item j exhibits DIF. For the LRT method, to test whether item j displays DIF, $\gamma_{jk} = 0$ is imposed for $k = 1, ..., p$ in the restricted model. While for the augmented model, all DIF parameters, except for those in the anchor set, are freely estimated. A Chi-squared distribution with p degrees of freedom is used to perform the LRT test. We first explore whether item-wise type I error rate can be controlled and compare the power of the two methods. The empirical type I error rates and the empirical powers are summarized in Tables 4.11 and 4.12 below. Theoretical significance level is set to be 5%. From Table 4.11, we first note that the type I errors can be controlled well for the proposed method across all settings. While for the LRT method, the type I errors are controlled well in most settings, except for the $N = 1000$ and $p = 9$

case, in which the error rates are nearly 7%. This might be due to convergence issue to asymptotic Chi-square distribution because of the smaller sample size and the larger number of estimates in higher dimensional p. From Table 4.12, we note the empirical powers of the proposed method outperform the LRT method with one anchor item and are slightly lower than those with three anchor items. The results are noteworthy because the proposed method does not require any prior knowledge on anchor items. For the proposed method, we note that statistical powers tend to drop when the dimension p increases, sample size decreases and the DIF parameters (signals) are small. We also remark that in practice, it is increasingly more difficult to locate an anchor item as dimension p becomes larger. Our proposed method is particularly useful in these cases. Additional inference results on individual γ_{jk} and the control for the false discovery rate (FDR) using the Benjamini-Hochberg procedure for multiple testing on all items can be found in Appendix C.

Table 4.11: Average empirical type I errors of the proposed method and the LRT method with 1, 3 and 5 anchor items respectively at significance level of 5%. Small and Large denote the small DIF setting and the large DIF setting, respectively.

	$p = 4$				$p = 9$			
	$N = 1000$		$N = 2000$		$N = 1000$		$N = 2000$	
	Small	Large	Small	Large	Small	Large	Small	Large
Proposed method	0.030	0.031	0.039	0.034	0.015	0.018	0.023	0.026
LRT (1 anchor item)	0.040	0.040	0.035	0.036	0.065	0.069	0.039	0.041
LRT (3 anchor items)	0.038	0.038	0.049	0.050	0.067	0.062	0.037	0.036
LRT (5 anchor items)	0.060	0.057	0.040	0.037	0.069	0.069	0.041	0.041

Finally, we compare the detection power of the two methods using the receiver operating characteristic (ROC) curves. An ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings. More specifically, ROC curves are constructed by thresholding the

Table 4.12: Average empirical powers of the proposed method and the LRT method with 1, 3 and 5 anchor items respectively at significance level of 5%. Small and Large denote the small DIF setting and the large DIF setting, respectively.

	$p = 4$				$p = 9$			
	$N = 1000$		$N = 2000$		$N = 1000$		$N = 2000$	
	Small	Large	Small	Large	Small	Large	Small	Large
Proposed method	0.730	1.000	0.953	1.000	0.680	1.000	0.950	1.000
LRT (1 anchor item)	0.500	1.000	0.817	1.000	0.467	0.987	0.890	1.000
LRT (3 anchor items)	0.853	1.000	0.990	1.000	0.880	1.000	0.987	1.000
LRT (5 anchor items)	0.883	1.000	0.997	1.000	0.933	1.000	1.000	1.000

p-values, from zero to one with a step-size of 0.01, for the proposed method and the LRT method, respectively. Note that for the LRT method, the TPR and FPR are calculated based on the non-anchor items only. For each method, an average ROC curve is obtained first based on the 100 replications, using which the area under the ROC curve (AUC) is calculated. A larger AUC value indicates better detection power. The AUC values for different methods across our simulation settings are given in Table 4.13. We observe that the proposed method and the LRT method with three anchor items have similar detection power and perform much better than knowing only one anchor item. This result aligns with the empirical power results in Table 4.12.

Table 4.13: Comparison of AUC of the proposed method and the LRT method with 1, 3 and 5 anchor items respectively. Small and Large denote the small DIF setting and the large DIF setting, respectively.

	$p = 4$				$p = 9$			
	$N = 1000$		$N = 2000$		$N = 1000$		$N = 2000$	
	Small	Large	Small	Large	Small	Large	Small	Large
Proposed method	0.951	0.997	0.989	0.998	0.965	1.000	0.994	0.999
LRT (1 anchor item)	0.846	0.984	0.949	0.974	0.828	0.989	0.976	0.996
LRT (3 anchor items)	0.971	0.993	0.995	0.994	0.978	0.996	0.996	0.996
LRT (5 anchor items)	0.969	0.994	0.992	0.994	0.964	0.991	0.998	0.998

4.3.4 Application to PISA 2018 Data

We apply our proposed method to PISA 2018 data. PISA is a worldwide study by the OECD intended to evaluate the participating country's educational systems by measuring 15-year-old school students' scholastic performance on mathematics, science and reading (*Araújo et al.*, 2021). It measures participants' problem solving and cognitive skills.

In this application, we explore the DIF effects brought by country factors in PISA 2018 reading, mathematics and science items separately. We focus on analysing the 37 OECD member countries. Only the computer-based and binary-scored items are considered. Furthermore, only items with responses from at least one subject from each of the 37 OECD countries are retained for valid $\hat{\gamma}_{jk}$. For reading items in specific, the fluency items are not included as they are designed to provide additional information about students' reading ability towards the lower end of the reading proficiency scale and has very small discriminative power and very low difficulty levels (see Chapter 2 and Annex A of *PISA* (2020)). After data pre-processing, the reading data set, the math data set and the science data set comprise 63, 43 and 76 cognitive items, respectively, and 290202, 145159 and 145228 observations respectively. We regard **country** as a multi-level categorical variable, with each category representing belonging to one of these 37 countries. To apply our proposed method, countries are encoded by $\boldsymbol{x}_i = (x_{i1}, ..., x_{i36})$. In specific, Australia is set to be the baseline for which $\boldsymbol{x}_i = \mathbf{0}$, and each of the remaining countries is encoded by $\boldsymbol{x}_i = \{0, , ..., 0, x_{ik} = 1, 0, ..., 0\}$. We point out that the results would be invariant of which baseline country is used, due to the symmetry in $\boldsymbol{x}_i$. The list of the OECD

countries considered in the study is summarized in Table 4.14 where country index equals k. We employ Algorithm 5 to learn the parameters in the proposed model (4.9). All the parameter estimates can be found in the Appendix C

Table 4.14: List of OECD countries considered in the PISA study.

Index	1	2	3	4	5
Country	Australia	Austria	Belgium	Canada	Chile
Index	6	7	8	9	10
Country	Colombia	Czech Republic	Denmark	Estonia	Finland
Index	11	12	13	14	15
Country	France	Germany	Greece	Hungary	Iceland
Index	16	17	18	19	20
Country	Ireland	Israel	Italy	Japan	Korea
Index	21	22	23	24	25
Country	Latvia	Lithuania	Luxembourg	Mexico	Netherlands
Index	26	27	28	29	30
Country	New Zealand	Norway	Poland	Portugal	Slovak Republic
Index	31	32	33	34	35
Country	Slovenia	Spain	Sweden	Switzerland	Turkey
Index	36	37			
Country	United Kingdom	United States			

We first explore the overall DIF effects for each item brought by country factors. As detailed in Section 4.3.2, we use the Hotelling statistic applied on $\hat{\boldsymbol{\gamma}}_j = \{\hat{\gamma}_{jk} : k = 1, ..., 36\}$ to measure the overall DIF exhibited by item j. To quantify the detection uncertainty, we employ Algorithm 7 in Section 4.3.2 to obtain a valid p-value for each item j. Using $M = 10000$ Monte Carlo samples, negligible p-values are obtained for all the items in the three assessment domains. Therefore, we conclude that DIF is prevalent in PISA 2018 data where most items are expected to suffer from DIF brought by country factors. This is not surprising because PISA 2018 item design puts more emphasis on comprehension skills (*Araújo et al.*, 2021), which is more prone to cultural and language impacts that are closely related to country factors. This underscores the need for DIF analysis to ensure fair and accurate assessment of

the participating country's educational systems. The ranking of items based on the overall DIF effect size (as measured by the Hotelling statistic, arranged in decreasing orders) considered in the three assessment domains are summarized in Tables 4.15, 4.16 and 4.17. Greater attention may be needed for the high ranking items for any future cross-country studies.

Table 4.15: Ranking of 63 PISA 2018 reading items in decreasing severity of DIF effects brought by country factors, as measured by the Hotelling statistics.

Ranking	1	2	3	4	5	6	7
Item	CR220Q04S	CR220Q06S	DR420Q02C	DR406Q01C	CR424Q02S	DR219Q01EC	DR219Q01C
Ranking	8	9	10	11	12	13	14
Item	DR432Q05C	CR227Q01S	DR455Q03C	DR453Q04C	DR437Q07C	DR455Q02C	DR420Q06C
Ranking	15	16	17	18	19	20	21
Item	CR412Q01S	CR220Q01S	DR406Q02C	DR446Q06C	DR453Q06C	CR404Q07S	DR466Q02C
Ranking	22	23	24	25	26	27	28
Item	DR102Q05C	CR437Q06S	CR455Q04S	CR466Q03S	DR412Q08C	CR412Q05S	DR456Q02C
Ranking	29	30	31	32	33	34	35
Item	CR455Q05S	CR424Q03S	CR220Q05S	CR432Q06S	DR432Q01C	CR104Q02S	DR460Q01C
Ranking	36	37	38	39	40	41	42
Item	DR219Q02C	DR102Q04C	DR227Q03C	CR111Q01S	CR102Q07S	CR437Q01S	DR456Q06C
Ranking	43	44	45	46	47	48	49
Item	CR460Q06S	CR424Q07S	CR456Q01S	CR404Q06S	DR227Q06C	CR055Q01S	DR055Q02C
Ranking	50	51	52	53	54	55	56
Item	DR406Q05C	DR055Q05C	DR404Q10BC	CR067Q01S	CR104Q01S	CR453Q05S	CR453Q01S
Ranking	57	58	59	60	61	62	63
Item	DR420Q09C	CR404Q03S	DR404Q10AC	CR466Q06S	CR460Q05S	CR220Q02S	CR446Q03S

Table 4.16: Ranking of 43 PISA 2018 math items in decreasing severity of DIF effects brought by country factors, as measured by the Hotelling statistics.

Ranking	1	2	3	4	5	6	7
Item	CM800Q01S	CM420Q01S	CM949Q02S	CM915Q02S	CM408Q01S	CM828Q03S	CM982Q03S
Ranking	8	9	10	11	12	13	14
Item	CM915Q01S	CM423Q01S	CM982Q04S	CM564Q01S	CM998Q04S	CM982Q02S	CM803Q01S
Ranking	15	16	17	18	19	20	21
Item	CM982Q01S	CM411Q02S	CM909Q01S	CM906Q01S	CM474Q01S	CM446Q01S	CM155Q04S
Ranking	22	23	24	25	26	27	28
Item	CM603Q01S	CM192Q01S	CM559Q01S	CM992Q02S	CM033Q01S	CM992Q01S	CM447Q01S
Ranking	29	30	31	32	33	34	35
Item	CM411Q01S	CM464Q01S	CM909Q02S	CM564Q02S	CM496Q01S	CM155Q01S	CM949Q01S
Ranking	36	37	38	39	40	41	42
Item	CM00GQ01S	CM305Q01S	CM571Q01S	CM034Q01S	CM442Q02S	CM496Q02S	CM273Q01S
Ranking	43						
Item	CM909Q03S						

Table 4.17: Ranking of 76 PISA 2018 science items in decreasing severity of DIF effects brought by country factors, as measured by the Hotelling statistics.

Ranking	1	2	3	4	5	6	7
Item	CS408Q04S	CS408Q05S	CS657Q01S	CS428Q01S	CS607Q02S	CS627Q01S	CS625Q02S
Ranking	8	9	10	11	12	13	14
Item	CS602Q04S	CS608Q01S	CS527Q03S	CS415Q08S	CS603Q03S	CS602Q02S	CS629Q02S
Ranking	15	16	17	18	19	20	21
Item	CS413Q04S	CS326Q04S	CS527Q04S	CS256Q01S	CS645Q03S	CS478Q03S	CS415Q07S
Ranking	22	23	24	25	26	27	28
Item	CS638Q04S	CS603Q04S	CS657Q02S	CS438Q01S	CS602Q01S	CS607Q01S	CS648Q03S
Ranking	29	30	31	32	33	34	35
Item	CS413Q06S	CS629Q04S	CS615Q07S	CS527Q01S	CS627Q03S	CS413Q05S	CS648Q02S
Ranking	36	37	38	39	40	41	42
Item	CS415Q02S	CS478Q02S	CS438Q02S	CS604Q02S	CS326Q03S	CS408Q01S	CS605Q03S
Ranking	43	44	45	46	47	48	49
Item	CS634Q01S	CS408Q01S	CS326Q03S	CS605Q03S	CS610Q02S	CS626Q02S	CS466Q01S
Ranking	50	51	52	53	54	55	56
Item	CS498Q02S	CS626Q01S	CS620Q02S	CS627Q04S	CS625Q03S	CS603Q05S	CS626Q03S
Ranking	57	58	59	60	61	62	63
Item	CS634Q04S	CS605Q02S	CS608Q03S	CS615Q05S	CS643Q04S	CS425Q05S	CS638Q01S
Ranking	64	65	66	67	68	69	70
Item	CS615Q01S	CS657Q03S	CS608Q02S	CS646Q02S	CS615Q02S	CS620Q01S	CS646Q03S
Ranking	71	72	73	74	75	76	
Item	CS635Q02S	CS603Q01S	CS646Q01S	CS643Q01S	CS428Q03S	CS425Q02S	

Next, we apply Algorithm 6 to group the countries so that DIF effects will vanish within each clusters, and we also seek to study the relationships amongst these 37 OECD countries and discuss the potential sources driving the DIF effects based on the learned $\hat{\boldsymbol{\gamma}}$ patterns. Essentially, countries are grouped into the same cluster if they have similar DIF pattern exhibited by the items, i.e. similar $(\gamma_{jk}, j = 1, ...J)$, and the dissimilarity is measured by the Euclidean distance between the DIF patterns. Furthermore, to remove noise in the trained $\hat{\boldsymbol{\gamma}}$, we also perform a simple principal component analysis, where each country is represented in a Cartesian coordinate system by projecting the corresponding $\hat{\boldsymbol{\gamma}}_k$ onto the first two principal components. Similarity matrices based on the Euclidean distances depicting the closeness in relationship amongst these countries are also presented.

For the reading data, the results are summarized in Figures 4.5, 4.6 and 4.7. We find that countries that are close in geographical locations and share similar speaking languages/cultures/history tend to have similar DIF patterns.

From Figure 4.5, we observe that Japan and Korea are the two most dissimilar countries compared to other OECD countries. We point out that both Japan and Korea are east Asian countries, which have very different cultures and spoken languages compared to other OECD countries. A full hierarchical clustering for countries based on the reading data is presented in Figure 4.6. We note the former countries that belong to the British Empire, including the Ireland, United States, Canada, Australia, New Zealand and the United Kingdom (*Marshall*, 2001), are clustered together in the dendrogram. Those countries speak English as their official language and share a deep-rooted British culture. Norway and Sweden, adjacent in geographical location, are clustered together, both are not too far away from other Nordic countries such as Iceland and Denmark. We point out that these countries share a common linguistic heritage (*Hovdhaugen et al.*, 2000). In particular, Danish, Icelandic, Norwegian and Swedish belong to the North Germanic branch of the Indo-European languages. The languages originated from a common Nordic language. Though having moved away from each other during the past 1000 years, it is still possible for Danish, Norwegian and Swedish speakers to understand each other. And these languages are taught in school throughout the Nordic countries. We notice that western European countries stay close in the dendrogram. Western European countries speak similar languages which mostly fall within two Indo-European language families: the Romance languages, descended from the Latin of the Roman Empire and and the Germanic lan-

guages (*Renfrew*, 1989). For example, Switzerland have German, French and Italian as its official languages, while Luxembourg uses Luxembourgish, French and German as its administrative languages. The central and eastern European countries, consisting of Poland, Czech Republic, Slovak Republic, Hungary and Slovenia are segregated closely to one another. Those countries also share similar history origins where the central Europe comprises most of the territories of the Holy Roman Empire (*Wilson*, 2011). Besides, these countries also speak similar languages. For example, in Slovenia, Hungarian is co-official with Slovene in 30 settlements in 5 municipalities, and Italian is co-official with Slovene in 25 settlements in 4 municipalities (*languages Slovene and Italians*, 2002). Moreover, we note that the Baltic states stay close in the dendrogram. The term includes Latvia, Estonia and Lithuania. These Baltic states are not only close in geographical locations, but they also share similar languages, cultures and history. For example, both the Latvian and Lithuanian languages belong to the Indo-European language family (*Subačius and Tekorienė*, 2002). All the three Baltic states gained independence from the Russian Empire in between the two world wars (*Lehti and Smith*, 2003) and eventually from the Stalinist Soviet Union in 1991. Finland is adjacent to Estonia in geographical location, that might explain why it is clustered together with the three Baltic countries. While south American countries, comprising of Chile, Colombia and Mexico are grouped closely together. We note that all the three nations speak Spanish, and they are also the founding members of Pacific Alliance in 2011, which is an initiative of regional integration to move progressively towards free mobility of goods, resources and people. Similar patterns can be observed in Figure 4.7 where countries are represented by projecting

the corresponding $\hat{\boldsymbol{\gamma}}_k$ onto the first two principal components.

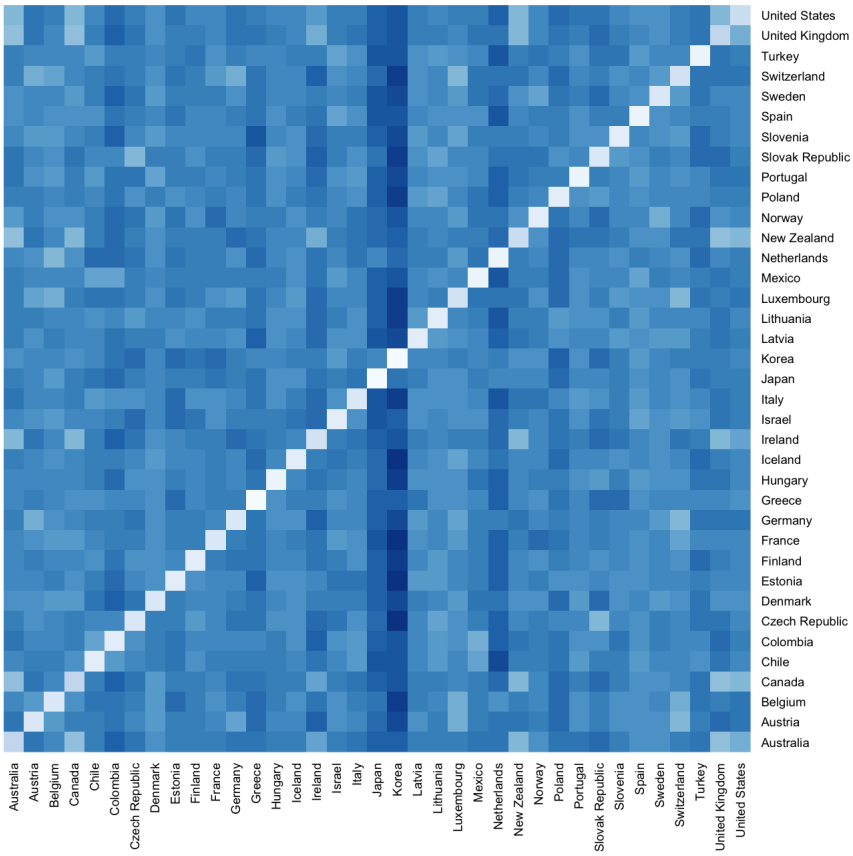

Figure 4.5: Plots of the similarity matrix of all 37 OECD countries based on the estimated DIF values for the PISA 2018 reading data. The level of dissimilarity between two countries is proportional to the degree of darkness of the block.

Similar analyses are performed on the math data and the science data respectively. In particular, results for math data set are summarized in Figures 4.8, 4.9 and 4.10. Results for the science data set are summarized in Figures 4.11, 4.12 and 4.13.

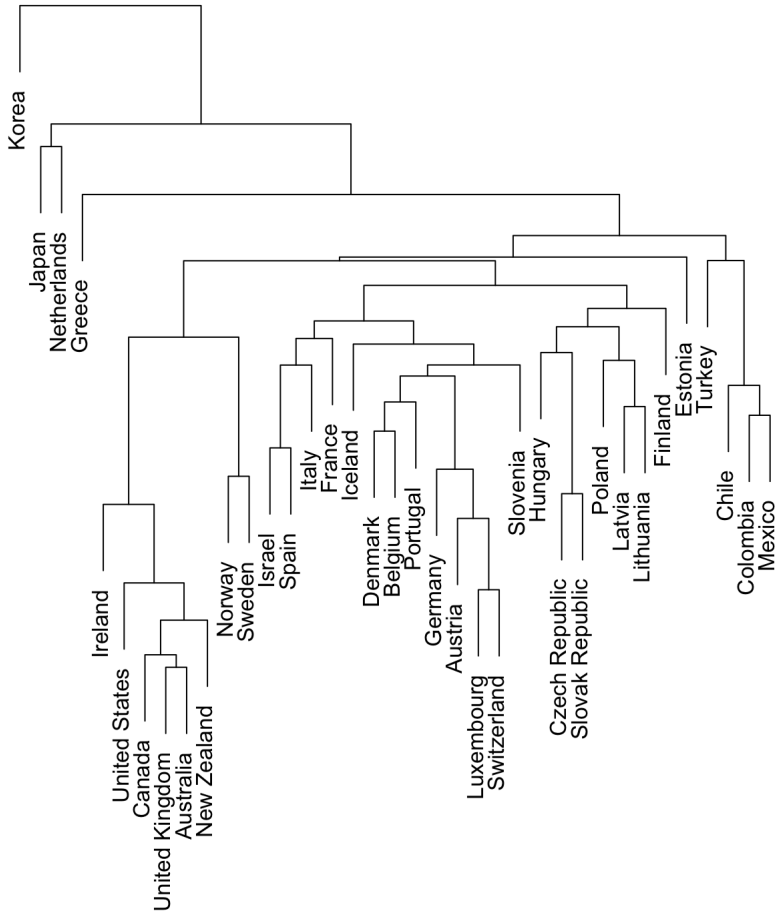

Figure 4.6: Hierarchical clustering of countries based on $\hat{\gamma}_{jk}$ values for the PISA 2018 reading data.

Again, we find that most of the OECD countries are very different from the two Asian countries, Japan and Korea, and the transcontinental country, Turkey. Furthermore, we notice some differences in clustering patterns on the math and the science data compared with the reading data. Due to the generality and conciseness of mathematical symbols, we conjecture that DIF on the math items stems also from the level

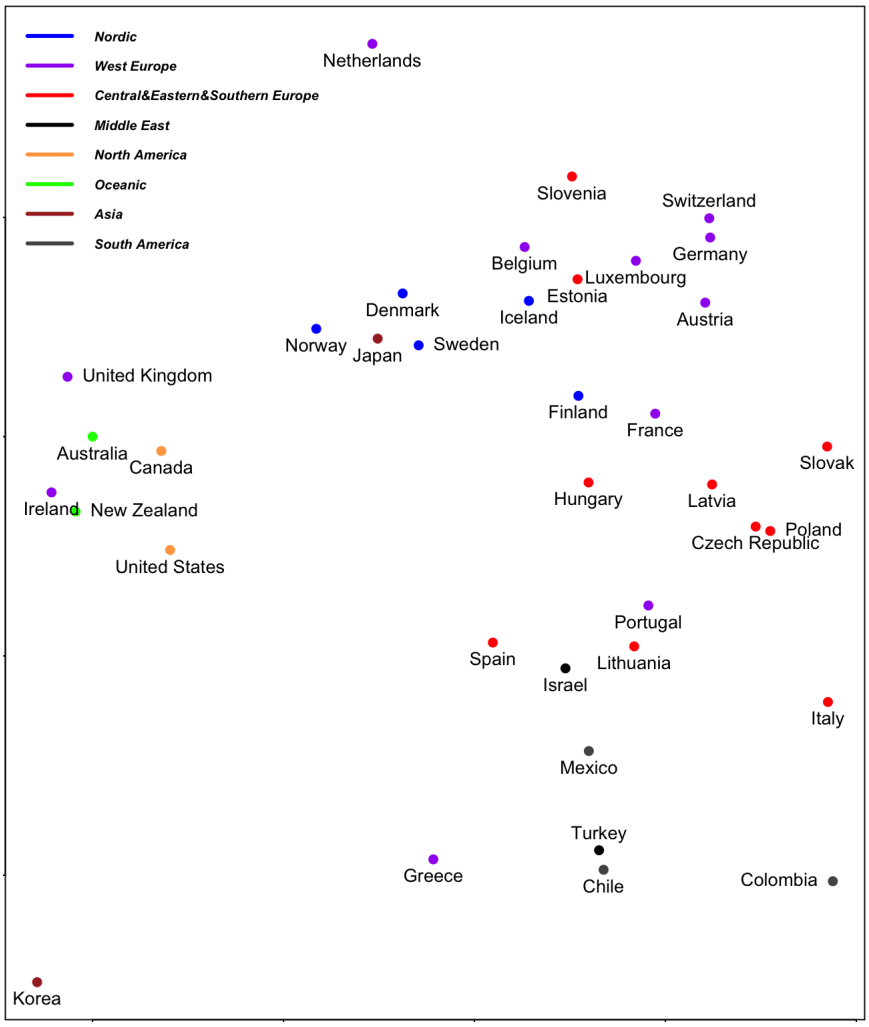

Figure 4.7: Plot of the countries represented by the projection of $\hat{\gamma}_k$ onto the first two principal component based on the PISA 2018 reading data.

of economic development and math literacy rates, in addition to locations, languages and cultures. This can be observed from the clustering patterns. For example, in Figure 4.9, we first note that western European countries, north American countries and Oceanic countries are close to each other in the dendrogram. These countries are very well developed, have high math literacy rates and share inter-twined history

origins of mathematical education. Again, the Nordic countries have very similar DIF patterns and are grouped together. This makes sense because Nordic countries share very similar history and origin in math education (*Lingefjärd*, 2018). Eastern European countries and the south American countries, on the other hand, stay closer in the clustering. These countries are generally less developed and have lower math literacy rate. While the two Asian countries, Korea and Japan, which have the most dissimilar educational system to other OECD nations, stay alone in a separate cluster.

For the science data, similar patterns can be observed to those in the math data sets. One difference we would like to point out is that for the Nordic countries, Denmark, Norway and Sweden are close to one another in the dendrogram, but are relatively separated from Finland and Iceland. This makes sense because Denmark, Norway and Sweden, often classified as the Scandinavian countries, are close to one another in geographic locations. They are often viewed as forming a homogenous unit in cultrual, social, economic and political matters. The two other countries within the group of the Nordic countries, Finland and Iceland, are separated both geographically, culturally and by language (*Lingefjärd*, 2018).

We remark that the clustering results can also be useful to facilitate analyses that do not take DIF effects into account. Many practitioners do not consider DIF when performing cross-country assessment analyses due to the additional complexity involved (*McConney et al.*, 2014; *D'Agostino*, 2016; *Cordero et al.*, 2018, 2022; *Joo et al.*, 2021; *Mazurek et al.*, 2021). The clustering results can help reduce the bias introduced by the DIF effects in these cross-country analyses even if DIF is not taken

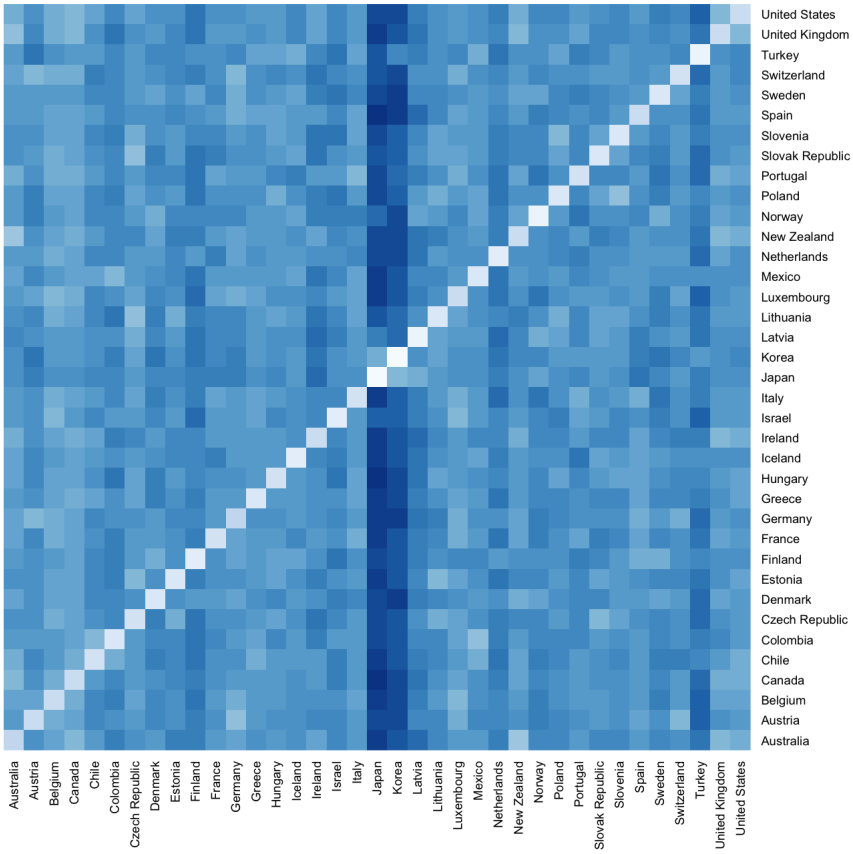

Figure 4.8: Plots of the similarity matrix of all 37 OECD countries based on the estimated DIF values for the PISA 2018 math data. The level of dissimilarity between two countries is proportional to the degree of darkness of the block.

into account. To achieve so, practitioners just need to perform analyses cluster-wise. Since DIF effects tend to be small within clusters; such analyses would be less prone to biases brought by DIF with respect to country factors. To illustrate how ignorance of the DIF effects can skew the analysis results, we compare country-wise

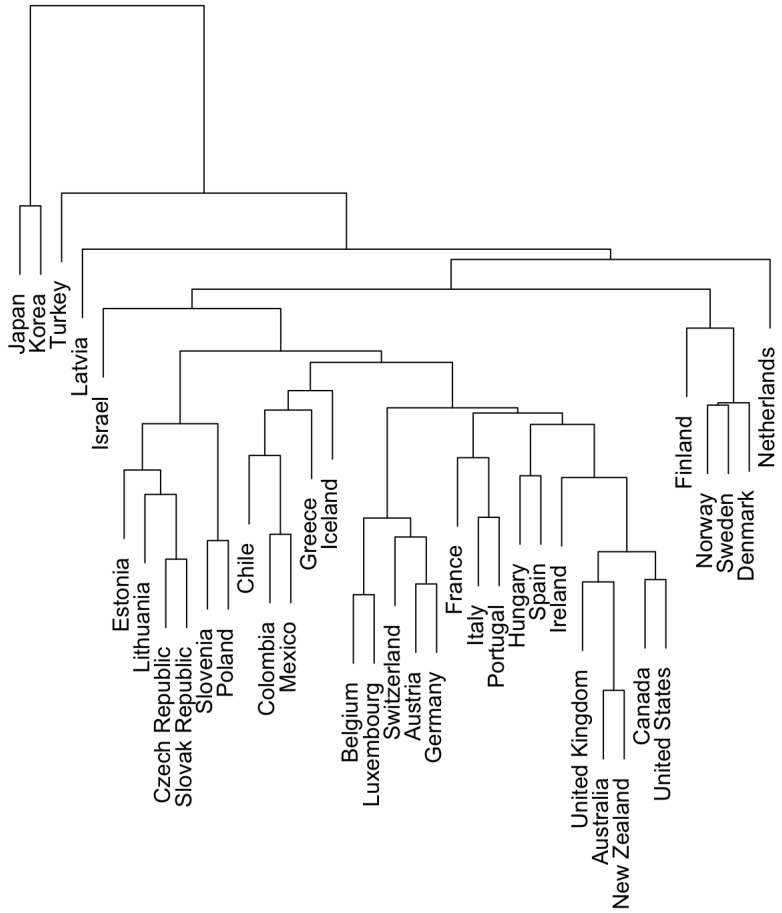

Figure 4.9: Hierarchical clustering of countries based on $\hat{\gamma}_{jk}$ values for the PISA 2018 math data.

average latent scores μ_k (level of examinees' competency) using a 2PL model with and without taking DIF into account for all the three assessment domains. We give a ranking of countries, from the highest to the lowest, based on the their mean latent proficiency scores in Tables 4.18, 4.19 and 4.20, on the reading, the math and the science data respectively. We observe that on the reading data, the order of the

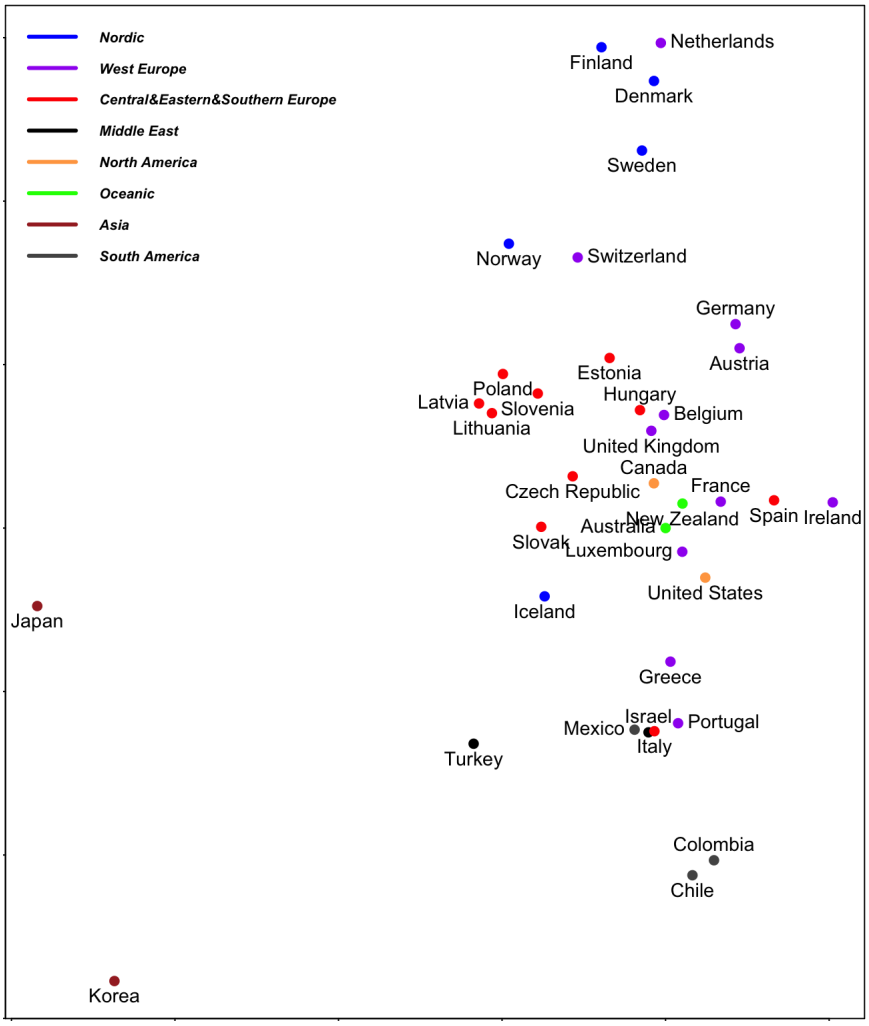

Figure 4.10: Plot of the countries represented by the projection of $\hat{\gamma}_k$ onto the first two principal component based on the PISA 2018 math data.

rankings change significantly. This suggests DIF brought by country factors would have massive impact on analysis results if they are not taken into account. However, for the math and the science data, though the rankings are still visibly different, they do not change as much as the reading data, suggesting DIF by country factors would have relatively smaller impact on these analysis results. This observation is

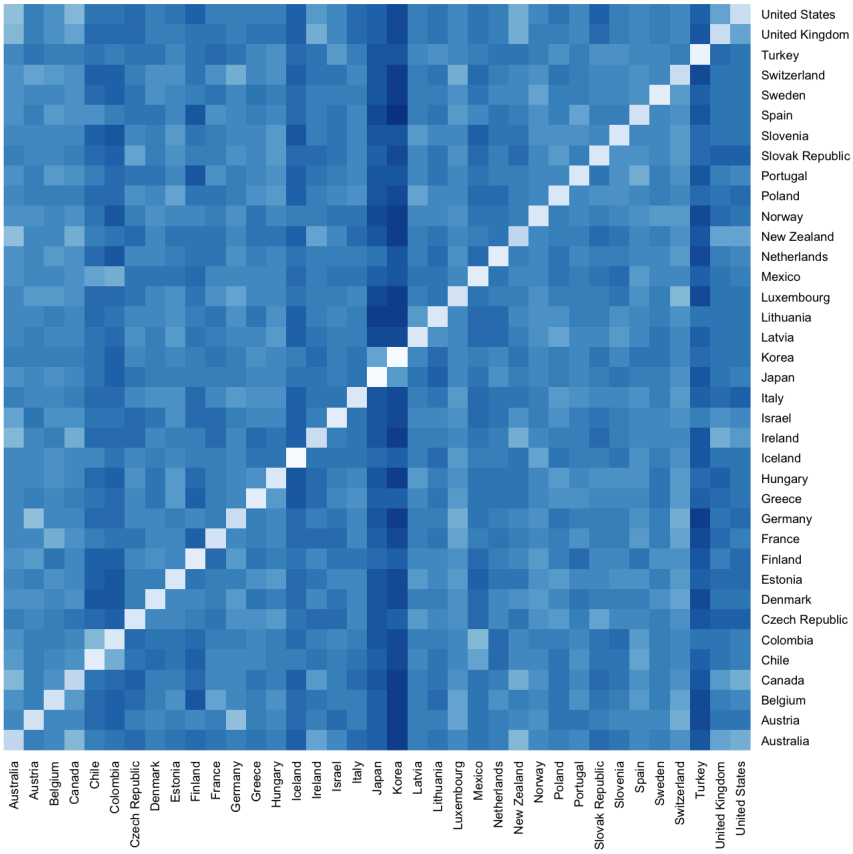

Figure 4.11: Plots of the similarity matrix of all 37 OECD countries based on the estimated DIF values for the PISA 2018 science data. The level of dissimilarity between two countries is proportional to the degree of darkness of the block.

confirmed by evaluating the Kendall's rank correlations. We note a kendall's rank correlation of 0.841 for rankings based on the reading data, lower than that of 0.886 and 0.886 on the math and the science data respectively. The observation is in line with our expectation since the design of the reading items involves a lot of wording

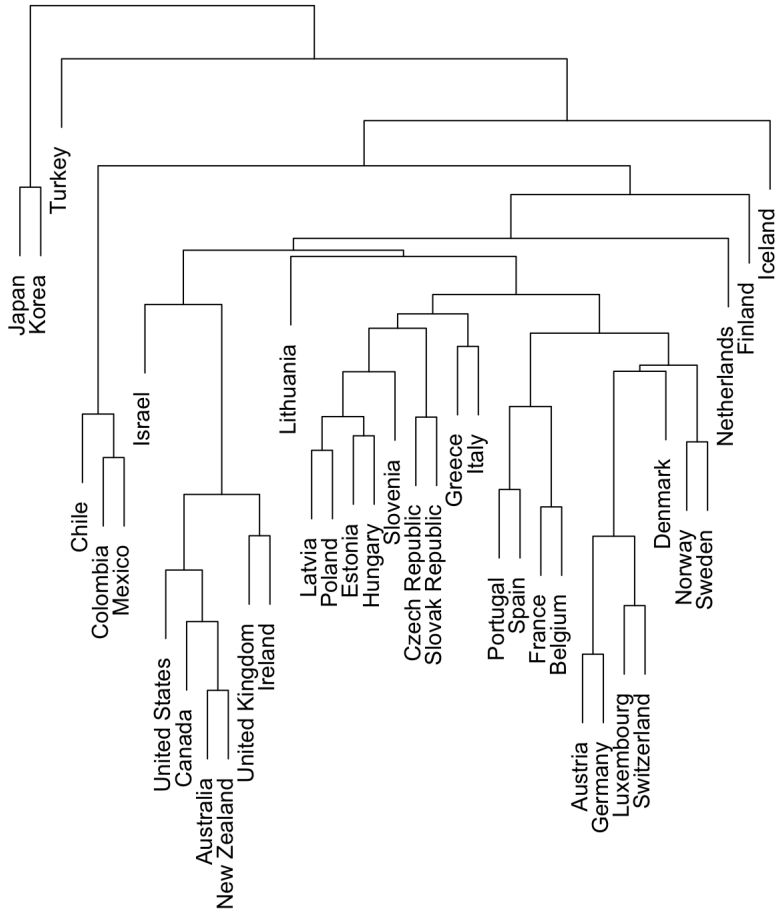

Figure 4.12: Hierarchical clustering of countries based on $\hat{\gamma}_{jk}$ values for the PISA 2018 science data.

description and requires more comprehension, which is more sensitive to country related factors such as language and cultural effects. This suggests why DIF is more severe in the reading items.

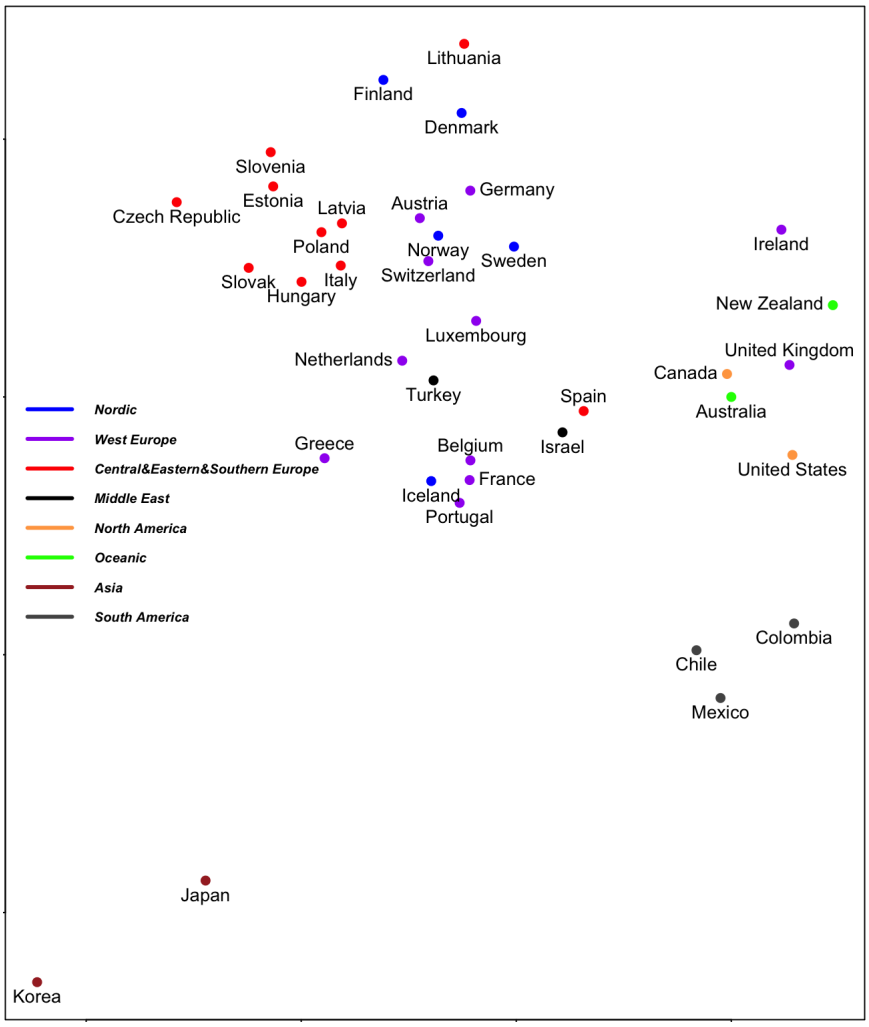

Figure 4.13: Plot of the countries represented by the projection of $\hat{\boldsymbol{\gamma}}_k$ onto the first two principal component based on the PISA 2018 science data.

4.4 Conclusion

We propose an ML1 condition for simultaneous DIF detection and model identifiability for 2PL models. The proposed method works for DIF analysis with high dimensional features. Consistency results for the proposed estimators and DIF de-

Table 4.18: Comparison of rankings of country-wise average latent skill levels between 2PL model without DIF and with DIF for the PISA 2018 reading data.

Ranking	Without DIF	With DIF	Ranking	Without DIF	With DIF
1	Finland	Estonia	20	Switzerland	Austria
2	Estonia	Finland	21	Austria	Iceland
3	Poland	Canada	22	United States	Netherlands
4	Canada	Poland	23	Latvia	Latvia
5	Korea	Ireland	24	Luxembourg	Switzerland
6	Japan	Sweden	25	Hungary	Italy
7	Sweden	Korea	26	Portugal	Hungary
8	Norway	Japan	27	Iceland	Lithuania
9	Ireland	Norway	28	Australia	Spain
10	France	New Zealand	29	Spain	Luxembourg
11	Germany	United Kingdom	30	Lithuania	Portugal
12	Belgium	Slovenia	31	Israel	Israel
13	New Zealand	Germany	32	Turkey	Turkey
14	Slovenia	France	33	Greece	Greece
15	United Kingdom	Belgium	34	Slovak Republic	Slovak Republic
16	Czech Republic	Denmark	35	Chile	Chile
17	Denmark	Czech Republic	36	Mexico	Mexico
18	Italy	United States	37	Colombia	Colombia
19	Netherlands	Australia			

Table 4.19: Comparison of rankings of country-wise average latent skill levels between 2PL model without DIF and with DIF for the PISA 2018 math data.

Ranking	Without DIF	With DIF	Ranking	Without DIF	With DIF
1	Japan	Japan	20	Ireland	Netherlands
2	Korea	Estonia	21	Austria	Italy
3	Estonia	Korea	22	Iceland	United Kingdom
4	Switzerland	Switzerland	23	Portugal	Ireland
5	Poland	Poland	24	Italy	Latvia
6	Australia	Slovenia	25	Latvia	Portugal
7	Slovenia	Canada	26	Luxembourg	Slovak Republic
8	Canada	Denmark	27	Lithuania	Luxembourg
9	Belgium	Australia	28	Spain	Spain
10	Denmark	Belgium	29	Hungary	Lithuania
11	Finland	Finland	30	Slovak Republic	United States
12	Netherlands	Germany	31	United States	Hungary
13	Czech Republic	Austria	32	Israel	Israel
14	Germany	Czech Republic	33	Greece	Greece
15	Norway	Norway	34	Turkey	Turkey
16	Sweden	New Zealand	35	Chile	Chile
17	United Kingdom	France	36	Mexico	Mexico
18	France	Sweden	37	Colombia	Colombia
19	New Zealand	Iceland			

tection accuracy are provided. Inference procedures to quantify the uncertainty in item-wise DIF detection is developed. Our proposed method does not require any prior knowledge about anchor items, permits easy statistical inference and enjoys

Table 4.20: Comparison of rankings of country-wise average latent skill levels between 2PL model without DIF and with DIF for the PISA 2018 science data set.

Ranking	Without DIF	With DIF	Ranking	Without DIF	With DIF
1	Estonia	Korea	20	France	France
2	Japan	Estonia	21	Switzerland	Latvia
3	Finland	Japan	22	Norway	Switzerland
4	Korea	Finland	23	Portugal	Portugal
5	Poland	Canada	24	Latvia	Norway
6	Canada	Poland	25	Lithuania	Lithuania
7	Slovenia	Australia	26	Hungary	Iceland
8	New Zealand	United Kingdom	27	Luxembourg	Luxembourg
9	United States	New Zealand	28	Spain	Hungary
10	United Kingdom	United States	29	Iceland	Italy
11	Germany	Slovenia	30	Italy	Spain
12	Sweden	Germany	31	Israel	Israel
13	Austria	Czech Republic	32	Slovak Republic	Slovak Republic
14	Australia	Denmark	33	Turkey	Turkey
15	Czech Republic	Netherlands	34	Greece	Greece
16	Ireland	Austria	35	Chile	Chile
17	Belgium	Sweden	36	Mexico	Mexico
18	Netherlands	Belgium	37	Colombia	Colombia
19	Denmark	Ireland			

fast computation.

There are numerous future works worth pursuing. First is to extend the current setting to take potential DIF effects in slope parameters a_j in 2PL model. The current setting only allows detection of uniform DIF while rules out the potential existence of non-uniform DIF effects. Considering DIF in slope parameters would make it feasible to capture both uniform and non-uniform DIF effects. However, we also remark on the potential drawbacks that the addition of high dimensional DIF parameters in slope would compound the difficulty in parameter estimation, and based on our initial exploration, much larger sample sizes are needed to ensure accurate estimation. Another interesting direction is to extend current setting to include high dimensional continuous features. One potential way is to employ feature partitioning so that population can be segregated into different groups. However, an appropriate objective function maximising potential DIF effects will be required to

perform reasonable partitioning of the features first, which can be complicated on its own. These are left for future exploration.

CHAPTER V

Inference for Optimal Differential Privacy Procedures for Frequency Tables

5.1 Introduction

When releasing data to the public, a critical concern is the risk of exposing individual information in the data set. Law enforcement, such as the European General Data Protection Regulation, has made de-identification compulsory before releasing data, i.e. removing personal identity from the record. However, even with such measures, data adversaries may still be able to infer about an individual's identity. Some examples on privacy breach include the Netflix prize (*Narayanan and Shmatikov*, 2008), the Washington State health record identification (*Sweeney*, 2013), recovery of the anonymous location data (*Golle and Partridge*, 2009) and privacy loss in genomic data (*Wang et al.*, 2009a).

In the past two decades, a concept known as the data differential privacy (DP) has been developed for the purpose of protecting against the risk of privacy loss.

Dwork et al. (*Dwork et al.*, 2006b) define the first formal definition of DP, which relates the risk for privacy loss to how much the answer to a query would change given the presence or absence of the most extreme person who is prone to privacy breach. Machine-learning types of development quickly adopt the DP-concept. A non-exhaustive list includes streaming (*Dwork et al.*, 2010), data mining (*Mohammed et al.*, 2011), dimension reduction (*Chaudhuri, Sarwate, and Sinha*, 2012), genome-wide association tests (*Yu et al.*, 2014), Bayesian learning (*Wang, Fienberg, and Smola*, 2015b) and recommender systems (*Friedman, Berkovsky, and Kaafar*, 2016). In a separate front, efforts have been made to incorporate the DP framework into the traditional disclosure-risk control data-synthesis approaches in which synthetic datasets are generated to represent the original observed data (*Rubin*, 1993; *Little*, 1993; *Raghunathan et al.*, 2003; *Reiter*, 2005; *Drechsler*, 2011; *Raab et al.*, 2016). In this direction, applications include creating private versions of discrete and categorical data (*Charest*, 2011; *McClure and Reiter*, 2012; *Abowd and Vilhuber*, 2008; *Hay et al.*, 2016; *Quick*, 2021), continuous data (*Wasserman and Zhou*, 2010; *Snoke et al.*, 2018), and network data (*Karwa, Krivitsky, and Slavković*, 2017; *Karwa and Slavković*, 2016). *Bowen and Liu* (2020) provide a comprehensive review of the private data synthesis methods.

The research on performing hypothesis tests within the framework of DP has gained momentum lately and consists of diverse directions. For the traditional tests applying to normal data, *Sheffet* (2017) considers the DP hypothesis testing and confidence interval construction for ordinary least squares and ridge estimators in linear regression. *Barrientos et al.* (2019) propose a differentially private mechanism

to release the test statistic and p-value from testing a regression coefficient against 0 from a linear regression model. In anomaly detection, *Degue and Le Ny* (2018) design a DP generalized likelihood ratio method to decide if data modeled as a sequence of independent and identically distributed (i.i.d.) Gaussian random variables has a given mean value. *Ding et al.* (2018) study how to conduct hypothesis tests on two population means while preserving privacy under the more restrictive requirement of local differential privacy. *Campbell et al.* (2018) provide a private analogue of the ANOVA test. *Task and Clifton* (2016) and *Couch et al.* (2019) study non-parametric DP rank-based tests. Generalizing a concept given by *Wasserman and Zhou* (2010), *Liu et al.* (2019) investigate the relationship between differential privacy framework and hypothesis testing with the goal of using testing analogue to further refine optimal DP regime. *Ferrando et al.* (2022) consider using parametric bootstrap to construct private confidence intervals and establish a consistency result for the proposed intervals.

Most of the existing DP mechanisms are developed on releasing summary statistics of the data set or responding to queries. On the testing front, the development also mostly focuses on 'perturbing' summary statistics, e.g., test statistics or sufficient statistics at parametric setting. *Avella-Medina* (2021) extends this direction by working on the influence-function structure directly for M-estimators' robust DP inferences. *Rogers et al.* (2016) investigate the use of adaptive hypothesis testing for p-value corrections and derive valid testing procedures under the challenging (ϵ, δ)-DP scenarios. When it comes to releasing the dataset itself, which is essential in the current data-sharing climate, data synthesis methods constructed under the Bayesian

framework remain to be the dominating trend.

As for releasing tabular data, the summary cell counts are also observations. For hypothesis testing using the DP-private tabular data, there are many research issues. For example, suppose the cell counts are provided according to different genders and races but the users are only interested at the variable race when the distributions of cell counts do not differ for different genders. For the original data, users can simply combine the categories from different genders and conduct testing on the merged data directly. For the private data on the other hand, to the best of our knowledge, an investigation of testing procedures has not been presented in the literature for this simple and commonly used operation. The same statement applies when one combine the observations collected from different locations or from different years in an analysis.

There have also been some works dedicated to developing hypothesis-testing procedures on private tabular data sets. When adding noise to each cell of a contingency table, *Johnson and Shmatikov* (2013) justify the practice of using classical statistical tests on the private tabular data theoretically by showing that the test statistic computed from a noisy table still asymptotically has the same chi-squared distribution as using the classical method. In an earlier and pioneer work, *Vu and Slavkovic* (2009) investigate the sample-size determination so that Chi-squared tests using either the private tabular data or the original data can achieve the same power. That is, the additional variation that is considered negligible asymptotically is not always truly negligible – a statement supported by our numerical investigation. *Wang et al.* (2015a) add Laplace errors to cell counts to create private tabular data and consider

the additional variation in the testing procedures. They resort to Monte Carlo methods to ensure the validity of the testing methods. Besides creating their own version of Monte-Carlo based methods, *Gaboardi et al.* (2016) add Laplace or normal errors to create their private tabular data and use the structure to derive the corresponding asymptotic distributions of the test statistics. Focusing on a simple Binomial setting and using the positive count of the original data, *Awan and Slavković* (2018) extend the use of the Neyman-Pearson lemma to construct the most powerful test under DP. For frequency tables, these existing methods could not avoid the scenarios of having negative cell counts.

As the cell counts in a frequency table can be presented as a one-way frequency, we concentrate on this setting and recommend an optimal mechanism that satisfies the standard ϵ-DP. Differing from most of the existing literature, the optimal procedure does not add errors with an explicit form of distribution to the test statistics or cell counts and only allows realization from the non-negative discrete values as entries of private cell counts. The proposed procedures naturally avoids having negative cell counts without further truncating the private versions of the observations at zero, thus are not subject to the induced biases or loss of utilities discussed in *Rinott et al.* (2018). Valid procedures for carrying out goodness-of-fit tests on the private tabular data are developed for the associated procedures. In particular, a de-biased test statistic for the optimal procedure is proposed and its asymptotic distributions are derived. Finite sample approximating distributions for the Chi-square goodness-of-fit test statistics on the commonly used Laplace and Gaussian mechanisms (with/without post-processing of converting negative cells to zero) are also

provided to adjust for the additional privacy-related noises injected so that valid inference can be made in even relatively small sample settings. As far as we know, our work is among the first to deal with statistical inference for DP mechanisms with such post-processing. Furthermore, we identify an explicit rate requirement for privacy regimes ϵ under which the inference procedures are valid. Moreover, we derive valid procedures for goodness-of-fit tests on private data after performing some common operations in practice, including inter-table merging (combine multiple tables) and intra-table merging (combine interior categories within a table).

We organize the remaining of the chapter as follows. Section 5.2 reviews the basic concepts of DP and some field standard DP mechanisms. Section 5.3 presents the optimal mechanism. Inference procedures for goodness-of-fit tests are included in Section 5.4, where both the inter-table and intra-table merging operations are considered. Section 5.5 consists of simulation studies to compare the performance of the optimal mechanism with the field standards. In Section 5.6, we apply our proposed methods to NCEDL's multi-state study data set to demonstrate the utility of our proposed method. Section 5.7 concludes with discussions and some potential future directions. The proofs for the developed theoretical results and some additional simulation results can be found in Appendix D. The materials of this chapter are mainly based on *Li et al.* (2022b).

5.2 Review of DP Fundamentals

In this section, we review the basics of DP and some of the most commonly used mechanisms in the literature. Differential privacy (DP) quantifies the degree

of privacy protection in terms of privacy budget ϵ. Importantly, DP is a property of the algorithms that produce the privacy-protected data and the algorithms are often created according to a given utility function. Algorithms that satisfy the DP criteria are referred to as differentially private algorithms. Before giving the formal definition of DP, we first introduce some notations. We denote the original observed count data cell and its private version as Y and Y^* respectively. We use Y' to denote the neighbor of Y. Here neighboring data means Y and Y' only differ by one individual. We state the formal definition of ϵ-DP.

Definition V.1 (ϵ-Differential Privacy)**.** An algorithm $\mathcal{M}$, is ϵ-DP if for all subsets $S \subset \text{Range}(M)$ and for all Y, Y' such that $d(Y, Y') = 1$, $P(\mathcal{M}(Y) \in S)/P(\mathcal{M}(Y') \in S) \leq \exp(\epsilon)$.

In the definition above, $\epsilon > 0$ is the privacy budget and $d(Y, Y') = 1$ means that Y and Y' differ by one record, making them being the so-called neighbors. One concern about algorithms that satisfy ϵ-DP is that they may inject large amount of noise to statistical query results for the reason of attaining a strong privacy guarantee. The practice could result in poor data utility. Several relaxations have been developed. Examples include the (ϵ, δ)-DP (*Dwork et al.*, 2006a) and probabilistic DP (*Machanavajjhala et al.*, 2008). These are considered as relaxations because, while still being 'formal', they offer slightly weaker privacy guarantees. Below we give the formal definition of (ϵ, δ)-DP, which is commonly used in the literature.

Definition V.2 ((ϵ, δ)-Differential Privacy)**.** An algorithm $\mathcal{M}$, is (ϵ, δ)-DP if for all subsets $S \subset \text{Range}(M)$ and for all Y, Y' such that $d(Y, Y') = 1$, $P(\mathcal{M}(Y) \in S) \leq \exp(\epsilon)P(\mathcal{M}(Y') \in S) + \delta$, where $\delta \in [0, 1]$.

Note ϵ-DP is a special case of (ϵ, δ)-DP when $\delta = 0$. The parameter δ adds a small probability when the bound given in Definition V.1 does not hold. Next we review some DP mechanisms that are considered as field standards, for the purpose of comparison. Since we focus on studying frequency tables, we formalize these algorithms using one-way frequency tables as observed data set $D = (Y_1, ..., Y_K)$. We assume $D \sim \text{Multinomial}(n, P_1, P_2, ..., P_K)$ here.

Laplace Mechanism (Lap). The Laplace mechanism can be applied to each of the K cells independently via $Y_k^* \sim \text{Lap}(y_k, 1/\epsilon)$ independently for $k = 1, ..., K$, where $Y_k = y_k$ is the actual observation made. Since we are dealing with count data, we use a discretized version of Laplace distribution with probability mass function, $P(Y_k^* = y_k^* \mid Y_k = y_k) = (1/C_1)\exp(-\epsilon|y_k - y_k^*|)$, for any integer y_k^*, where $C_1 = \sum_{l \in \mathbb{Z}} \exp\{-\epsilon|y_k - l|\} = 1 + 2\exp(-\epsilon)/\{1 - \exp(\epsilon)\}$.

It is well-known that the above procedure satisfies ϵ-DP. Since frequency tables contain non-negative cell counts only, it is natural to perform post-processing to ensure all the private data entries are non-negative. Here we denote the Lap procedure with the post-processing of converting all negative private entries to zero as the the truncated Laplace (TLap) mechanism.

Gaussian Mechanism (GDP). Similar to the Laplace mechanism, the Gaussian mechanism perturbs each of the K cells independently via $Y_k^* \sim \text{N}(y_k, 2\log(1.25/\delta)/\epsilon^2)$ independently for $k = 1, ..., K$. It has been shown in *Dwork and Roth* (2014) that this procedure satisfies (ϵ, δ)-DP whenever $0 < \epsilon, \delta < 1$. Again, we use a discretized

version here with $P(Y_k^* = y_k^* \mid Y_k = y_k) = (1/C_2)\exp\{-(y_k - y_k^*)^2/(2\sigma^2)\}$, for any integer y_k^*, where $C_2 = \sum_{m\in\mathbb{Z}} \exp\{-(y_k - m)^2/(2\sigma^2)\}$.

It has been shown in *Canonne et al.* (2020) that this discrete version has approximately the same privacy guarantee as the continuous Gaussian mechanism. For comparison purpose only, we adopt this discrete version of the Gaussian mechanism. Similarly, we consider the GDP with the post-processing of converting all negative private entries to zero as the the truncated GDP (TGDP) mechanism.

Binomial-Beta McClure-Reiter Mechanism (MR). *McClure and Reiter* (2012) proposed an approach to synthesize count data using $Y_k^* \mid D = (y_1, ..., y_K) \sim \text{Bin}(n, (Y_k + \alpha_k)/(n + \alpha_k + \beta_k))$ independently for each cell Y_k, where $\alpha_k = \beta_k = 1/\{\exp(\epsilon/n) - 1\}$ makes this procedure satisfy ϵ-DP.

This is among the most commonly used data synthesis method, adapted to satisfy the DP requirement. Its advantage is that it preserves the underlying data structure in that, marginally, each Y_k follows a Binomial distribution. However, this procedure completely ruins the cell-wise information due to large α and β values and will in general yield deteriorated utilities.

5.3 Optimal Mechanism

From the previous section, we know some existing privacy mechanisms have been developed for releasing the frequency table data. However, these mechanisms are not optimal and have other shortcomings. Take the most commonly used Laplace and Gaussian mechanisms as examples; one of the main concerns is that negative count

data are easily generated, which does not make any practical sense in the frequency table setting. The popular existing methods that overcome this shortcoming often bear a large amount of variation. Taking the MR mechanism as an example, we note that, although the negative count issue is overcome, the damages to the utilities are not always well controlled under the targeted privacy constraints. Furthermore, in real applications, practitioners may face too many choices of mechanisms, often making it difficult for them to pick up the "best" one to use. We also note that optimality of DP algorithms in terms of utility maximization have been discussed by several authors. For example, *Ghosh, Roughgarden, and Sundararajan* (2012) studies the optimality of ϵ-differentially private mechanisms under a Bayesian framework. *Geng and Viswanath* (2015) derives that the optimal ϵ-differentially private mechanism for real-valued query functions takes the staircase-shaped probability densities that are geometrically decaying. While *Kairouz, Bonawitz, and Ramage* (2016) proves the optimality of the randomized aggregatable privacy-preserving ordinal response algorithm and the k-ary randomized response algorithm, under the local differential privacy framework. In this section, we seek to extend the universal optimality idea from *Ghosh et al.* (2012), in which the mechanism allows a flexible design of loss functions to measure utility, and the corresponding expected utilities are maximized under any given privacy requirements, and we recommend an optimal mechanism for the practitioners when applied to releasing the frequency-table type of data.

Before introducing the optimal mechanism, we first define some notations. Denote the observed data as $D = (y_1, ..., y_K)$, generated from Multinomial$(n, P_1, ..., P_K)$. The corresponding private data after DP procedures is denoted as $D^* = (y_1^*, ..., y_K^*)$.

For notation simplicity, we denote $i \in \{0, 1, ..., n\}$ as inputs (i.e. the values for y_k). Furthermore, we denote r as the private responses (i.e. the values for y_k^*) where $r \in \{0, 1, ..., n\}$. Let $p = \{p_{ir} : i = 0, 1, ..., n, r = 0, 1, ..., n\} \in \mathbb{R}^{(n+1)\times(n+1)}$ with p_{ir} denoting the probability of mapping an input i to r. Then the optimal p, denoted as $p^* \in \mathbb{R}^{(n+1)\times(n+1)}$, minimizes the expected loss (i.e. maximizes the expected utility), such that

$$p^* = \arg\min_p \sum_{i=0}^{n} \sum_{r=0}^{n} p_{ir} L(i, r), \tag{5.1}$$

where $L(i, r)$ can be any arbitrary loss function, subject only to the constraints that $L(i, r)$ are non-negative, and non-decreasing in $|i - r|$ for each fixed $i = 0, ..., n$. Note that $p^* = (p_{ir}^*)$ defines a stochastic mechanism that maps an input $i = 0, ..., n$ to an output $r = 0, ..., n$. The commonly used loss functions include L_1 and L_2 losses.

Algorithm 8: Optimal Mechanism

Input: Observed data $D = (y_1, ..., y_K)$, privacy regime ϵ.

Output: Optimal perturbation matrix p^*, private data D^*.

Set $\alpha = \exp(-\epsilon)$; Initialize $g, h, x \in \mathbb{R}^{(n+1)\times(n+1)}$.

Step 1. Evaluate g corresponding to the truncated and discretized Laplace Mechanism.

for $i = 0, 1, 2, ..., n$ **do**
  for $r = 0, 1, 2, ..., n$ **do**
    if $r = 0$ *or* 1 **then**
      $g_{ir} = \alpha^{|i-r|}/(1+\alpha)$;
    else
      $g_{ir} = \alpha^{|i-r|}(1-\alpha)/(1+\alpha)$.
  end
end

Step 2. Evaluate the "posterior" probabilities h as if using uniform prior on i.

for $r = 0, 1, 2, ..., n$ **do**
  $s = \left(\sum_{i'=0}^{n} g_{i'r}\right)$;
  for $i = 0, 1, 2, ..., n$ **do**
    $h_{ir} = g_{ir}/s$.
  end
end

Step 3. Compute the optimal remap matrix x.

for $r = 0, 1, 2, ..., n$ **do**
  $r^* = \arg\min_{j\in\{0,...,n\}} \sum_{i=1}^{n} h_{ir} L(i, j)$.
  for $k = 0, 1, ..., n$ **do**
    if $k = r^*$ **then**
      $x_{rk} = 1$.
    else
      $x_{rk} = 0$.
  end
end

Step 4. Evaluate the optimal perturbation matrix p^*.

for $i = 0, 1, 2, ..., n$ **do**
  for $r = 0, 1, 2, ..., n$ **do**
    $p^*_{ir} = \sum_{r'=0}^{n} g_{ir'} x_{r'r}$.
  end
end

Step 5. Generate private frequency table.

for $k = 1, 2, ..., K$ **do**
  Sample $y^*_k \sim \{0, 1, ..., n\}$ according to $\{p^*_{y_k r} : r = 0, 1, ..., n\}$;
  $D^*[k] = y^*_k$.
end

The optimal mechanism is detailed in Algorithm 8. Step 1 in Algorithm 8 evaluates a perturbation matrix g corresponding exactly to the discretized Laplace Mechanism, but truncated at 0 and n. The tail probabilities beyond 0 and n are all accumulated as the boundary probabilities. Its optimality has been demonstrated in *Geng and Viswanath* (2014) and *Ghosh et al.* (2012). If we fix response r in Step 2, we note that the vector $h_{0r}, ..., h_{nr}$ can be interpreted as a list of posterior probabilities conditioning on the response r with a uniform prior on the inputs $i = 0, ..., n$. We will use h to evaluate an optimal remap specific to the loss function $L(i, r)$, which is presented in Step 3. Its main goal is to achieve the best balance between the bias and variance so that the expected loss can be minimized. For a general loss function $L(i, r)$, step 3 seeks to find an optimal remap index r^*, for each response $r = 0, ..., n$, such that

$$r^* = \arg \min_{j \in \{0,...,n\}} \sum_{i=1}^{n} h_{ir} L(i, j).$$

Note that it is computed as the minimizer of the weighted expected loss. In reality, this optimal remap often brings in some bias into the random error added, but the output variance is significantly reduced which more than compensates for the bias. Then the optimal remapping matrix $x \in \mathbb{R}^{(n+1)\times(n+1)}$ is set to be $x_{rk} = 1$ if $k = r^*$ and $x_{rk} = 0$ if $k \neq r^*$. Finally in step 4, the optimal perturbation matrix p^* can be obtained by combining g and the optimal remap matrix x with $p^* = g \times x$, where $\times$ here denotes the matrix multiplication. Lastly, to find the private data cell y_k^*, we can simply sample using $r \in \{0, 1, ..., n\}$ with probability distribution $\{p^*_{y_k r} : r = 0, ..., n\}$.

Remark V.3. In step 3, when $L(i, r) = |i - r|$ is the L_1 loss, optimal remap index is

simply r^*=min$\{k = 0, 1, ..., n : \sum_{i=0}^{k} h_{ir} \geq 0.5\}$. Step 3 in Algorithm 8 simply returns r^* as the ceiling function of the conditional median of $\{0, ..., n\}$ with probabilities h_{ir} for $i = 0, ..., n$ in this case. Note that if we take $L(i, j) = (i - j)^2$ as the squared loss, then optimal r^* can be evaluated as the ceiling function of the conditional mean of $\{0, ..., n\}$ with probabilities h_{ir} for $i = 0, ..., n$.

Remark V.4. Algorithm 8 has time complexity of $O(n^3)$ and space complexity of $O(n^2)$ in general. Note that the time complexity is dominated by Step 3. In the most commonly used L_1 and L_2 losses, short-cuts in Remark 1 can be used, in which cases the time complexity can be reduced to $O(n^2)$, comparable to the field standards the Laplace mechanism and the Gaussian mechanism.

Following from Theorem 3.1 in *Ghosh et al.* (2012) by taking a uniform prior on the input $\{i = 0, 1, ..., n\}$ with probability mass function $P(i) = 1/(1 + n)$, it can be shown that the p^* obtained from the above steps solves the objective function (5.1) while satisfying the ϵ-DP framework. This is formalized in the proposition below.

Proposition V.5. *The perturbation matrix* $p^* \in \mathbb{R}^{(n+1)\times(n+1)}$ *obtained through Steps 1 to 4 in Algorithm 8 solves the problem* (5.1) *with loss function* $L(i, r)$ *that is non-negative and non-decreasing in* $|i - r|$*, satisfying the following constraints: for any* $0 < \epsilon < \infty$*, (1)* $p^*_{ir} - \exp(\epsilon)p^*_{(i+1)r} \geq 0$ *for* $i = 0, ..., n - 1, r = 0, ..., n$ *and (2)* $\exp(\epsilon)p^*_{ir} - p^*_{(i+1)r} \leq 0$ *for* $i = 0, ..., n - 1, r = 0, ..., n$*. Therefore, the mechanism described in Algorithm 8 satisfies* ϵ*-DP.*

Note that here p^* gives a perturbation matrix that is optimal in that it minimizes the overall expected losses while satisfying the ϵ-DP framework. In the following sections, we will work with the optimal mechanism that minimizes the most commonly

used expected L_1 loss and develop inference procedures for it. At the same time, the derivation applies to other losses as well.

5.4 Goodness-of-Fit Test

In this section, we develop procedures for conducting goodness-of-fit tests on private data. Furthermore, we also consider common operations including both inter- and intra-table mergings.

Assume the true data is $D = (Y_1, Y_2, ..., Y_K) \sim \text{Multinomial}(n, P_1, P_2, ..., P_K)$. Following a common practice, we release both $D^* = \left(Y_1^*, Y_2^*, ..., Y_K^*\right)$, the private tabular data, and the private mechanism used to generate D^*. Suppose we are interested in the goodness-of-fit test $H_0 : P_1 = p_1, P_2 = p_2, ..., P_K = p_K$ against $H_1 : P_1 \neq p_1$ or $P_2 \neq p_2$, ..., or $P_K \neq p_k$. Note that unlike the Gaussian or the Laplace mechanisms that inject a mean zero noise into each tabular cell, the boundary truncation and the optimal remapping step in the optimal mechanism will introduce some biases into the outputs to reach optimality. We propose a debiased goodness-of-fit test statistic on the private data generated from the optimal procedures described in Section 5.3. Consider the test statistic T^*_{opt} with

$$T^*_{opt} = \sum_{k=1}^{K} \left(\frac{y_k^* - np_k - b(y_k^*)}{\sqrt{np_k}} \right)^2 = \sum_{k=1}^{K} {T'_k}^2,$$

where $b(y_k^*)$ is the bias estimate stemming from the injected noise which can be evaluated using Algorithm 9 below.

Algorithm 9: Evaluation of Bias

Input: Private data D^* and optimal perturbation matrix p^*.

Output: Bias terms $b(y_k^*)$ for $k = 1, ..., K$.

for $k = 1, \ldots, K$ **do**

1. $f_{y_k^*} = (p^*_{0y_k^*}, p^*_{1y_k^*}, ..., p^*_{ny_k^*})^T / \Big(\sum_{i=0}^n p^*_{iy_k^*} \Big)$;
2. Evaluate $b = (b_0, b_1, ..., b_n)$:

for $i = 0, 1, ..., n$ **do**

$b_i = \sum_{j=0}^n p^*_{ij}(j - i)$;

end

3. $b(y_k^*) = \sum_{i=0}^n f_{iy_k^*} b_i$.

end

Remark V.6. Step 1 of Algorithm 9 seeks to find the list of probabilities of input values (denoted as $f_{y_k^*}$) from which the observed private y_k^* is likely to be sampled from. While step 2 computes the list of expected biases if the input values are $0, 1, ..., n$. Step 3 computes a weighted average of the expected biases to give the final bias estimate at y_k^*.

In order to take the second moment of the injected noise into account, we give an estimate for the variance, $v(y_k^*)$, to approximate the variance of the injected noise added to y_k using Algorithm 10 below.

Algorithm 10: Evaluation of Variance

Input: Private data D^* and optimal perturbation matrix p^*.

Output: Variance terms $v(y_k^*)$ for $k = 1, ..., K$.

for $k = 1, \ldots, K$ **do**

1. $f_{y_k^*} = (p^*_{0y_k^*}, p^*_{1y_k^*}, ..., p^*_{ny_k^*})^T / \Big(\sum_{i=0}^n p^*_{iy_k^*} \Big)$;
2. Evaluate $v = (v_0, v_1, ..., v_n)$:

for $i = 0, 1, ..., n$ **do**

$v_i = \sum_{j=0}^n p^*_{ij} (j - \sum_{j=0}^n j p^*_{ij})^2$;

end

3. $v(y_k^*) = \sum_{i=1}^n f_{iy_k^*} v_i$.

end

Remark V.7. Step 1 of Algorithm 10 is exactly the same as in Algorithm 9. Step 2 of Algorithm 10 computes the list of expected variances given the possible original observations of $0, 1, ..., n$ (we denote it as $v = (v_0, v_1, ..., v_n)$). Step 3 computes a weighted average of the expected variances to give the final estimate for the variance term at y_k^*.

We characterize the asymptotic null distribution of T^*_{opt} in the following theorem.

Theorem V.8. *Assume the private data are generated from the optimal procedure with privacy regime ϵ_n satisfying $\epsilon_n^{-1} n^{-1/2} \to 0$ as $n \to \infty$. Then under the null hypothesis $H_0 : P_1 = p_1, ..., P_K = p_K$, for some $1 < K < \infty$, $T^*_{opt} \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ^2_1 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K \times K}$ where $\Sigma_{kk} = 1 - p_k + v(y_k^*)/(np_k)$ for $k = 1, ..., K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \le k \ne j \le K$.*

Remark V.9. We can decompose $y_k^* = y_k + err_k$, and Theorem V.8 takes the second moment of err_k into account so that the asymptotic null distribution can have better finite sample properties when the sample size n is small and the privacy-protection requirement is high (small ϵ_n). Furthermore, we state that the rate of decrease of privacy regime ϵ_n cannot be faster than $n^{-1/2}$ for the asymptotics to work. For much perturbed outputs with small n and ϵ_n, inference procedures have low powers in general. Under such scenarios, as shown in the numerical outcomes, our proposed optimal procedure outperforms others. When an even smaller ϵ_n is required so that the asymptotic fails, one perhaps should carefully consider whether it is meaningful to release such a deteriorated data set.

Theorem V.8 can be generalized easily to any DP mechanisms whose injected noises are additive to the true cell counts. Below we take the most commonly used Laplace and Gaussian mechanisms (and their corresponding post-processing versions, TLap and TGDP) as examples and derive their asymptotic distributions. We use a standard Pearson Chi-square test statistic in the literature which is given as follows,

$$T^* = \sum_{k=1}^{K} \frac{(y_k^* - np_k)^2}{np_k} = \sum_{k=1}^{K} \Big(\frac{y_k^* - np_k}{\sqrt{np_k}} \Big)^2 = \sum_{k=1}^{K} T_k^2.$$

Theorem V.10. *Assume the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$. Under the null hypothesis $H_0 : P_1 = p_1, ..., P_K = p_K$, for some $1 < K < \infty$, the following results hold.*

(a). When the private data are generated from the ϵ_n-DP Laplace mechanism or the ϵ_n-DP truncated Laplace mechanism (at zero). $T^* \to \sum_{k=1}^{K} \Lambda_k Z_k$ in distribu-

tion, where Z_k are i.i.d. χ^2_1 random variables and Λ_k are eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K\times K}$ where $\Sigma_{kk} = 1 - p_k + 2/(np_k\epsilon_n^2)$ for $k = 1, ..., K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \le k \ne j \le K$.

(b). When the private data are generated from the (ϵ_n, δ)-DP Gaussian mechanism or the truncated (ϵ_n, δ)-DP Gaussian mechanism (at zero) for some $0 < \delta < 1$. $T^* \to \sum_{k=1}^K \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ^2_1 random variables and Λ_k are eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K\times K}$ where $\Sigma_{kk} = 1-p_k+(2\log(1.25/\delta)-1)/(np_k\epsilon_n^2)$ for $k = 1, ..., K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \le k \ne j \le K$.

5.4.1 Merging Multiple Frequency Tables

The data users may often encounter the need to merge different private tabular datasets. For example, the users may want to merge multiple data sets across different time-periods or regions before performing statistical analysis. Merging multiple frequency lists can increase sample size and therefore improve confidence when performing statistical inference. In this section, we develop inference procedures that can be applied to the merged private frequency tables.

Suppose the users are interested in merging C data lists $j = 1, ..., C$ together, with the j'th private data list denoted as $D_j^* = \{Y_{j1}^*, Y_{j2}^*, ..., Y_{jK}^*\}$ with sample size n_j. Further assume $n = \sum_{j=1}^C n_j$. The merged data set can then be denoted as $D_m^* = \{\sum_{j=1}^C Y_{j1}^*, ..., \sum_{j=1}^C Y_{jK}^*\}$. Furthermore, suppose the user knows the DP procedure used to create each of the private data lists D_j^*. To test $H_0 : P_1 = p_1, P_2 = p_2, ..., P_K = p_K$ against $H_1 : H_0$ does not hold on the merged data, We consider the following test

statistic

$$T_M^* = \sum_{k=1}^{K} \Big(\frac{\sum_{j=1}^{C} Y_{jk}^* - np_k - b_M(\{y_{jk}^*\}_{j=1}^{C})}{\sqrt{np_k}} \Big)^2 = \sum_{k=1}^{K} T_{mk}^2,$$

where $b_M(\{y_{jk}^*\}_{j=1}^{C}) = \sum_{j=1}^{C} b(y_{jk}^*)$. Then Theorem V.11 characterizes the asymptotics of the T_M^* under the null hypothesis. We give the results for both the recommended optimal procedure and the commonly used mechanisms in the literature.

Theorem V.11. *Assume $\epsilon_n^{-1} n^{-1/2} \to 0$ as $n \to \infty$. Under the null hypothesis H_0 : $P_1 = p_1, ..., P_K = p_K$, for some $1 < K, C < \infty$, the following results hold.*

(a) If D_j^* are obtained from the optimal procedure with privacy regime ϵ_n, then $T_M^* \to \sum_{k=1}^{K} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K \times K}$ where $\Sigma_{kk} = 1 - p_k + v_M(\{y_{jk}\}_{j=1}^{C})/(np_k)$ for $k = 1, ..., K$ with $v_M(\{y_{jk}\}_{j=1}^{C}) = \sum_{j=1}^{C} v(y_{jk}^*)$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \le k \ne j \le K$.

(b) If D_j^* are obtained from the ϵ_n-DP Laplace mechanism or the truncated ϵ_n-DP Laplace mechanism (at zero), then we set $b_M(\{y_{jk}^*\}_{j=1}^{C}) = 0$ in T_M^*. We have $T_M^* \to \sum_{k=1}^{K} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K \times K}$ where $\Sigma_{kk} = 1 - p_k + 2C/(\epsilon_n^2 np_k)$ for $k = 1, ..., K$ and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \le k \ne j \le K$.

(c) If D_j^* are obtained from the (ϵ_n, δ)-Gaussian mechanism or the truncated (ϵ_n, δ)-Gaussian mechanism (at zero) for some $0 < \delta < 1$, then we set $b_M(\{y_{jk}^*\}_{j=1}^{C}) = 0$ in T_M^*. We have $T_M^* \to \sum_{k=1}^{K} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{K \times K}$ where $\Sigma_{kk} = 1 - p_k +$

Cv_M/np_k *for* $k = 1, ..., K$ *with* $v_M = (2\log(1.25/\delta) - 1)/\epsilon_n^2$, *and* $\Sigma_{kj} = -\sqrt{p_k p_j}$ *for* $1 \leq k \neq j \leq K$.

5.4.2 Merging Cells Within a Frequency Table

Data users may also be interested in combining entries within a frequency table, either because they are interested in a more general group of classes or because the sample sizes of some cells are too small to carry out valid analysis. Similar to inter-table merging, intra-table merging directly on the private tabular data accumulates random noises in the merged cells, resulting in invalid analysis results if these noises are not taken into account separately. In this section, we provide goodness-of-fit test procedures that can be applied to the intra-table merged private data sets.

Without loss of generality, suppose the users are interested in merging the first M cells of the private list $D^* = \{Y_1^*, Y_2^*, ..., Y_K^*\}$ for some $M < K$. Denote the resulting merged data set as $D_m^* = \{\sum_{k=1}^{M} Y_k^*, Y_{M+1}^*, ..., Y_K^*\} = \{Y_{m1}^*, Y_{m2}^*, ..., Y_{m(K-M+1)}^*\}$. To test $H_0 : P_{m1} = p_1, P_{m2} = p_2, ..., P_{m(K-M+1)} = p_{K-M+1}$ against $H_1 : H_0$ does not hold on the merged data set D_m^*. We consider

$$T_M^* = \sum_{k=1}^{K-M+1} \Big(\frac{Y_{mk}^* - np_k - b_M(y_{mk}^*)}{\sqrt{np_k}} \Big)^2 = \sum_{k=1}^{K-M+1} T_{mk}^2,$$

where $b_M(y_{m1}^*) = \sum_{i=1}^{M} b(y_i^*)$ and $b_M(y_{mk}^*) = b(y_{M+k-1}^*)$ for $k = 2, ..., K - M + 1$. The following theorem characterizes the asymptotic null distribution of T_M^*. Again, we give the results for both the recommended optimal procedure and the commonly used mechanisms in the literature.

Theorem V.12. *Assume the privacy regime ϵ_n satisfies $\epsilon_n^{-1}n^{-1/2} \to 0$ as $n \to \infty$. Under the null hypothesis $P_{m1} = p_1, P_{m2} = p_2, ..., P_{m(K-M+1)} = p_{K-M+1}$, for some $1 < K < \infty$ and $1 \leq M < K$, the following results hold.*

(a) If D^* are from the ϵ_n-DP optimal procedure, then $T_M^* \to \sum_{k=1}^{K-M+1} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the matrix $\Sigma \in \mathbb{R}^{(K-M+1)\times(K-M+1)}$ where $\Sigma_{kk} = 1 - p_k + v_M(y_{mk}^*)/(np_k)$ for $k = 1, ..., K-M+1$, with $v_M(y_{m1}^*) = \sum_{i=1}^{M} v(y_i^*)$ and $v_M(y_{mk}^*) = v(y_{M+k-1}^*)$ for $k = 2, ..., K-M+1$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leq k \neq j \leq K-M+1$.

(b) If D^* are from the ϵ_n-DP Laplace mechanism or the truncated ϵ_n-DP Laplace mechanism (at zero), we set $b_M(y_{mk}^*) = 0$ in T_M^*. then $T_M^* \to \sum_{k=1}^{K-M+1} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{(K-M+1)\times(K-M+1)}$ where $\Sigma_{11} = 1 - p_1 + Mv_M/np_k$ for $k = 1, ..., K$ and $\Sigma_{kk} = 1 - p_k + v_M/np_k$ for $k = 2, ..., K-M+1$, with $v_M = 2/\epsilon_n^2$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leq k \neq j \leq K-M+1$.

(c) If D^* are from the (ϵ_n, δ)-DP Gaussian mechanism or the truncated (ϵ_n, δ)-DP Gaussian mechanism (at zero), we set $b_M(y_{mk}^*) = 0$ in T_M^*. Then we have $T_M^* \to \sum_{k=1}^{K-M+1} \Lambda_k Z_k$ in distribution, where Z_k are i.i.d. χ_1^2 random variables and Λ_k are the eigenvalues of the matrix $\Sigma \in \mathbb{R}^{(K-M+1)\times(K-M+1)}$ where $\Sigma_{11} = 1 - p_1 + Mv_M/np_k$, $\Sigma_{kk} = 1 - p_k + v_M/np_k$ for $k = 2, ..., K-M+1$ with $v_M = (2\log(1.25/\delta) - 1)/\epsilon_n^2$, and $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $1 \leq k \neq j \leq K-M+1$.

5.5 Simulation Study

Various simulation studies are designed to examine and compare the effectiveness of the recommended Opt procedure in Section 5.3 with the five methods, Lap, TLap, GDP, TGDP and MR reviewed in Section 5.2; the latter five are commonly used algorithms in the literature. Throughout the section, we consider three targeted privacy regimes of $\epsilon = 0.25, 0.5, 0.75$. L_1 loss is used to compare the performance of different procedures. We also numerically validate the inference procedures given in Section 5.4.

5.5.1 Utility

Here, sample size $n = 500$ and observed counts of $i = 5, 200, 450$ are considered. 5000 Monte Carlo samples are generated for each setting and L_1 losses are evaluated. For the MR mechanism, $\alpha = \beta = 1999.5, 999.5, 666.2$ are required to achieve $\epsilon = 0.25, 0.5, 0.75$ respectively. For the GDP/TGDP, since the ϵ-DP does not exist, we relax it to (ϵ, δ)-DP instead. It is a common practice to take $\delta \leq 1/n$ and we set $\delta = 1/500 = 0.002$ in this case. For Lap, TLap and Opt, procedures are implemented as described in Section 5.2 and Section 5.3 respectively. The distributions of the losses across different settings are summarized and compared using box-plots in Figure 5.1. From Figure 5.1, we note that the Opt mechanism improves utilities significantly in comparison to the traditional data synthesis MR mechanism. We also observe that the MR method's utility varies with the observed values of i (5, 200, 450). The closer the observed i-value is to $n/2 = 250$, the better the MR method's utility. Regardless of the values of i, MR mechanism's performance is much worse than

the other procedures. GDP/TGDP, on the other hand, though requiring further relaxation on privacy regime with an additional δ of 0.002, its utility is still worse than that of the Lap/TLap and the Opt across all scenarios. The truncated mechanisms have similar performances as their counterparts, though they tend to give results with lower variations as indicated by slightly smaller box sizes.

Overall, the Opt mechanism appears to achieve comparable utilities to the Lap/T-Lap mechanisms. To examine closely, we present the Monte Carlo means and variances of the L_1 losses in Table 5.1. Compared to the Lap/TLap mechanisms, we observe that the Opt mechanism achieves smaller Monte Carlo means and similar variances of the L_1 losses under the same privacy constraints. The major improvement of Opt is achieved under the most restrictive privacy regime, $\epsilon = .25$ and for the observed $i = 5$, most distant from the center $n/2$ and therefore more prone to privacy risks. Moreover, we also point out that the truncated mechanisms, including the TLap and the TGDP, are expected to perform no worse than their non-truncated counterparts because converting negative cells to zero will produce private data closer to their true underlying values which are always non-negative. Indeed, when the true value is close to zero at $i = 5$ and when the privacy regime is high at $\epsilon = 0.25$ (so the injected noises are large and more negative cells would be converted to zero), we observe significant improvements on the mean and variance of L_1 losses for the truncated versions over their non-truncated counterparts. While when the underlying value is large and privacy regime is small, the improvement is not so obvious. Some counter-intuitive observations are due to Monte Carlo errors.

In the last part of Table 5.1 under 'Negative Proportion', we report the proportion

of Monte Carlo samples with at least one negative count. In reality, a negative cell count will not be observed. Releasing a private table with a negative cell count will likely reduce the users' confidence in the quality of the released tables. We notice that when $\epsilon = 0.25$, the proportion of negative counts yielded from the Lap and GDP versions of private tables are 12.3% and 36.8% respectively. In contrast, the Opt mechanism ensures that no negative count will be generated. This characteristic demonstrates the benefits of releasing Opt versions of private frequency tables as compared to the traditional Lap or GDP mechanism without any post-processing procedures.

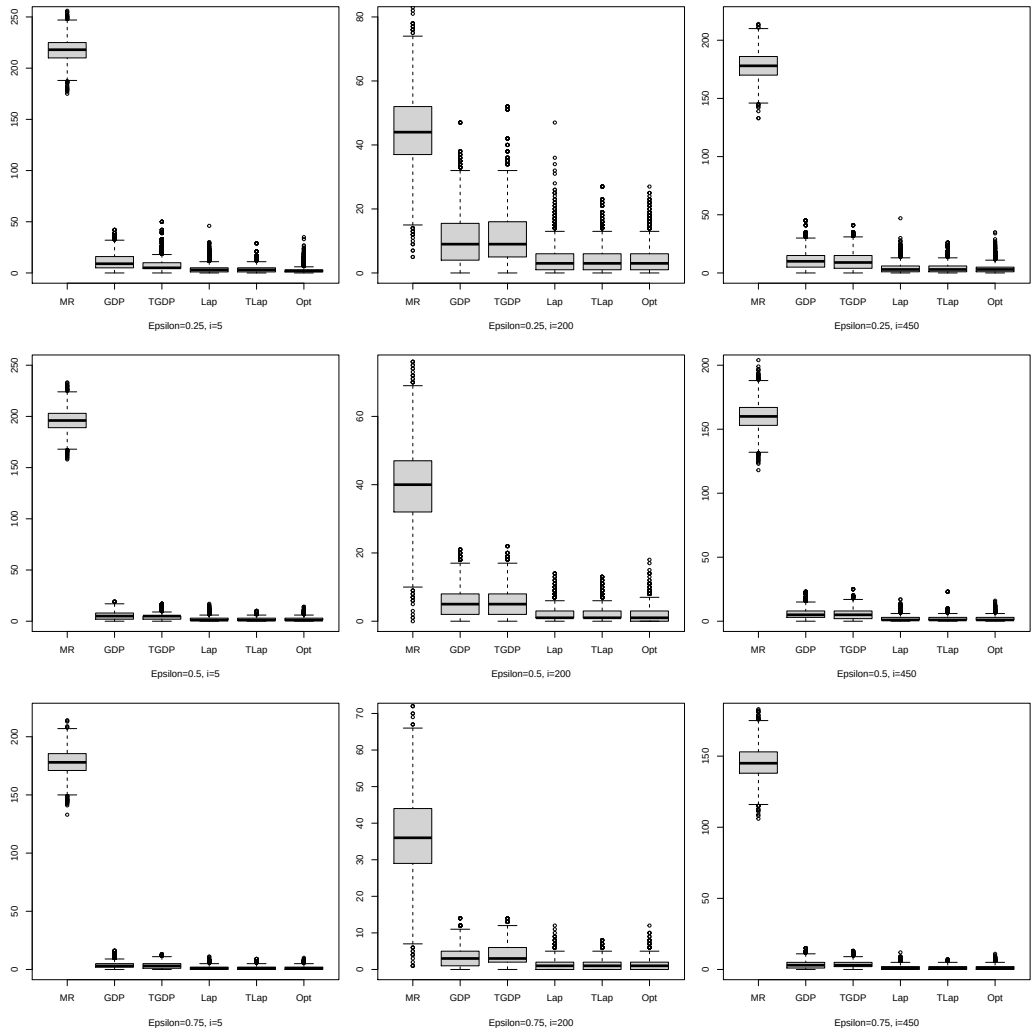

Figure 5.1: Utility comparison (L_1 loss) amongst the Opt, Lap, TLap (truncated Laplace), GDP, TGDP (truncated GDP) and MR mechanisms across different privacy regimes $\epsilon = 0.25, 0.5, 0.75$ and observed data counts $i = 5, 200, 450$.

	$\epsilon = 0.25$			$\epsilon = 0.5$			$\epsilon = 0.75$		
	i=5	i=200	i=450	i=5	i=200	i=450	i=5	i=200	i=450
	Mean								
Opt	2.94	3.98	3.93	1.79	1.90	1.88	1.23	1.20	1.24
Lap	3.97	4.04	4.08	2.03	1.96	1.96	1.31	1.27	1.33
TLap	3.44	3.99	4.15	1.84	2.04	1.94	1.31	1.31	1.34
GDP	11.08	11.04	10.98	5.41	5.61	5.76	3.75	3.42	3.63
TGDP	7.82	11.19	10.40	4.47	5.64	5.67	3.49	3.74	3.67
MR	217.55	44.20	177.89	195.79	39.80	160.05	178.20	36.45	145.29
	Variance								
Opt	10.03	15.53	15.60	3.06	4.11	3.94	1.84	1.86	1.96
Lap	16.68	16.67	16.60	4.58	3.97	4.17	1.96	1.83	1.95
TLap	9.69	17.29	17.26	2.98	4.04	4.35	1.66	1.80	1.87
GDP	68.61	71.93	69.33	15.87	18.90	17.84	8.40	7.73	7.56
TGDP	52.56	72.54	63.63	9.41	16.87	18.64	6.03	7.38	7.33
MR	121.85	126.44	125.23	118.62	129.83	123.29	115.95	122.04	118.34
	Negative Proportion								
Opt	0.000			0.000			0.000		
Lap	0.123			0.037			0.009		
TLap	0.000			0.000			0.000		
GDP	0.368			0.210			0.104		
TGDP	0.000			0.000			0.000		
MR	0.000			0.000			0.000		

Table 5.1: Mean, variance of the L_1 losses and proportion of negative counts out of 5000 Monte Carlo samples for different privacy regimes, $\epsilon = 0.25, 0.5, 0.75$, and observed counts, $i = 5, 200, 450$.

5.5.2 Goodness-of-fit Test

In this sub-section, we seek to numerically validate the inference procedures in Section 5.4. Considering the frequency data $D \sim \text{Multinomial}(n, P_1 = 0.1, P_2 = 0.1, P_3 = 0.8)$, we are interested in testing $H_0 : P_1 = 0.1, P_2 = 0.1, P_3 = 0.8$. We consider differing sample sizes $n = 100, 1000$, under three privacy targets $\epsilon = 0.25, 0.5, 0.75$. Setting the significance level to be 0.05, we evaluate 500 empirical test statistics. First, we examine the use of the traditional Chi-square distribution with $K - 1$ degrees of freedom as if the data is not perturbed. We check whether the empirical type I errors could be controlled using this naive asymptotic null distribution for the private data sets produced by the five mechanisms, Opt, Lap, TLap, GDP and TGDP. The resulting average empirical type I errors are provided under the "Naive

Method" scenario in Table 5.2. Except for the setups when the sample size is large at $n = 1000$ and the privacy control is not strict with $\epsilon > 0.5$, the average empirical type I error rates are way above the targeted value of 0.05 for all the mechanisms. That is, the use of the naive χ^2_{K-1} null distribution cannot control the type I error in such cases. The situation is worse when the GDP or TGDP mechanisms are used. When sample sizes are small at $n = 100$ or when the privacy regime is strict with $\epsilon = 0.25$, the naive method performs poorly and valid inference is impossible.

Next, we test the results in Theorems V.8 and V.10 under the same simulation settings as above. The goal is to check whether the new test statistics and null distributions derived can control the type I error rate well so that valid hypothesis testing can be implemented. Under the "Proposed Method" scenario of Table 5.2, we report the mean empirical type I errors obtained from the 500 simulated samples. In contrast to the "Naive Method" scenario, the empirical type I errors are controlled fairly well at around 5%. We point out that truncation tends to reduce the type I error rate, especially when sample size is small and privacy regime is high.

		$n = 100$			$n = 1000$		
		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
Naive Method	Opt	0.392	0.180	0.108	0.102	0.063	0.054
	Lap	0.449	0.188	0.112	0.102	0.064	0.056
	TLap	0.448	0.187	0.113	0.101	0.062	0.055
	GDP	0.865	0.549	0.327	0.401	0.141	0.088
	TGDP	0.865	0.549	0.327	0.401	0.141	0.088
Proposed Method	Opt	0.037	0.047	0.054	0.052	0.050	0.052
	Lap	0.066	0.058	0.055	0.055	0.051	0.051
	TLap	0.034	0.053	0.055	0.055	0.051	0.051
	GDP	0.052	0.051	0.051	0.051	0.052	0.051
	TGDP	0.023	0.028	0.038	0.051	0.052	0.051

Table 5.2: Mean empirical type I errors out of 500 simulated samples under two study scenarios across different sample sizes $n = 100, 1000$ and different privacy regimes $\epsilon = 0.25, 0.5, 0.75$. The two scenarios are "Naive Method": Chi-squre tests for original data, "Proposed Method": procedure proposed in Section 5.4.

To compare the statistical powers, we consider the alternative hypotheses H_1 : $P_1 = p_1, P_2 = p_1, P_3 = 1 - 2p_1$, and explore the cases with $p_1 = 0.1, 0.11, 0.12, ..., 0.25$ for $n = 100$ and $p_1 = 0.1000, 0.1025, 0.1050, ..., 0.1350$ for $n = 1000$. The results are summarized in Figure 5.2. In Figure 5.2, as the H_1 hypothesized values depart further from that in the H_0, all mechanisms have the powers rising to 1. The Lap/T-Lap mechanisms and the Opt procedures have significantly higher power than the GDP/TGDP mechanisms, especially when the sample size and the ϵ-value is small ($n = 100$ or $\epsilon = 0.25$). The Lap/TLap and the Opt procedures perform similarly with the Opt outperforming Lap/TLap slightly when the sample size is small. As n and ϵ get larger, the differences in power amongst the mechanisms diminish. This is reasonable because the random noise injected is inversely proportional to ϵ and is scaled by $\sqrt{n}$ in the test statistic. So, when the sample size increases, the noises become increasingly more negligible. That makes the statistical powers for the mechanisms merge to the same performance level when the sample size becomes large. We also note that post-processing of truncating at zero does not have much impact on the statistical power.

Furthermore, we compare the power performance of the non-debiased and the de-biased test statistics for the Opt under the same setting as above. The results are summarized in Figure 5.3. We observe that the bias correction step can help improve statistical power slightly when ϵ and sample size are small. On the other hand, the simpler version without the bias-correction component is as competitive in all settings.

Moreover, we also conduct simulation studies to validate the results given in The-

orems V.11 and V.12 for inter- and intra-table merging. Simulation results show that empirical type I errors can be controlled well using the approximate null distribution developed, suggesting Theorems V.11 and V.12 provide valid inference procedures. Our simulation results also show that the recommended Opt mechanism perform the best in terms of the empirical powers. These additional simulation results are included in Appendix D

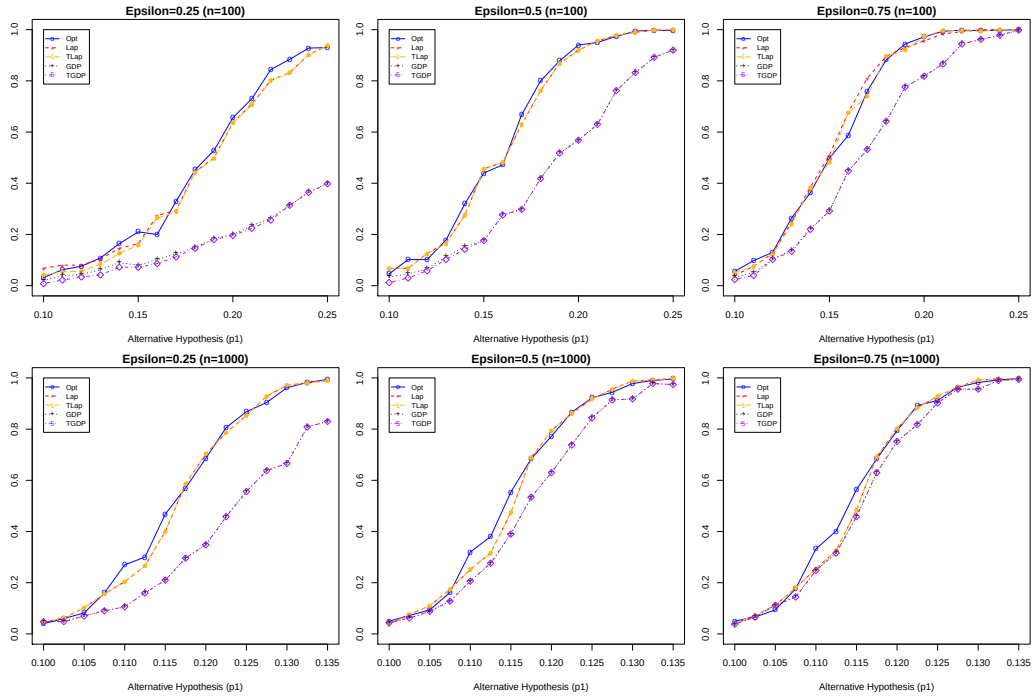

Figure 5.2: Empirical power comparison for five privacy procedures: Opt, Lap, TLap, GDP and TGDP.

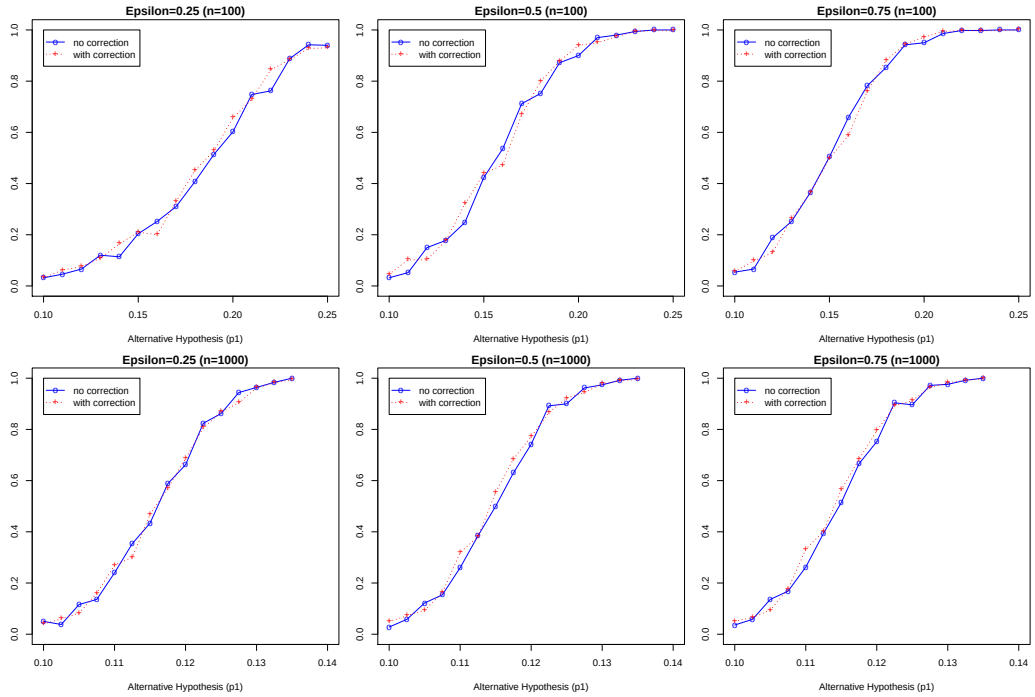

Figure 5.3: Empirical power comparisons for the Opt procedures with and without bias-correction.

5.6 Analysis of Children's Early Development and Learning Data

In this section, we consider an application to the data from the NCEDL's multi-state study (*M. Clifford et al.*, 2017). The data set consists of 2982 records of 308 variables collected from pre-kindergarten children in 11 states of the United States of America. We focus on one-way frequency tables and select two categorical variables *household type* and *family income* to investigate the differences between the three east coast states, Massachusetts, New York and New Jersey, and the other states.

We start with the variable, *household type.* This is a variable with five categories describing different household types: (I) single mom or dad, (II) mom and dad both in home, (III) w/o dad, (IV) multiple adults, but parents/step-parents not both in home, and (V) single adult, not mom or dad. After removing the instances with missing outcomes and collapsing into frequency lists, the values, including those from the three east coast states and the remaining eight states, are summarized in Table 5.3.

	(I)	(II)	(III)	(IV)	(V)	n
Massachusetts	85	237	9	36	5	372
New York	48	83	4	24	3	162
New Jersey	66	174	18	51	4	313
Others	403	1241	143	251	21	2059

Table 5.3: One-way frequency table of the variable, *household type,* in Massachusetts, New York, New Jersey and the other eight states (denoted as Others) in the NCEDL study.

State by state for the three east coast states, we apply the three mechanisms considered in Section 5.5 with $\epsilon = 0.25, 0.5, 0.75$, and $\delta = 1/n$ (only for GDP and TGDP) to these frequency tables. In particular, the Opt mechanism used here optimizes against L_1 losses and is implemented according to Algorithm 8. For each case, 500 private samples are simulated. Using the New York state as an example, we report the summarized utilities for the five DP-mechanisms. In Table 5.4, we present the mean values of the private entries under each of the five categories, the average entry-wise Monte Carlo standard errors (Ave. SD), the average entry-wise L_1 loss (Mean L_1 loss), and the proportions of private datasets generated with at least one negative entry out of 500 simulated samples ('< 0' Proportion). We note that the Opt procedure will always produce private data with the smallest mean

L_1 losses as compared to the other methods, in line with the theoretical results in Proposition V.5. The results also suggest that the improvement in the utility is a result of the reduced uncertainty as illustrated by the smallest average Monte Carlo SD's achieved by the Opt mechanism. Furthermore, we observe that categories (III) and (V) have close to 0 entries; thus their corresponding private versions for Lap and GDP procedures might be negative, especially in high privacy regimes where noise injected is large. Indeed, when $\epsilon = 0.25$, the proportion of private frequency tables produced by the Lap and GDP mechanisms with negative entries are quite high. In reality, releasing such frequency tables could cause confusion and doubts amongst users about the usefulness of the data sets.

		(I)	(II)	(III)	(IV)	(V)	Ave. SD	Ave. L_1 loss	'< 0' Proportion
$\epsilon = 0.25$	OPT	48.00	82.70	5.37	24.12	4.89	4.87	3.44	0.00
	Lap	47.59	82.99	3.76	23.88	3.07	5.74	4.01	0.35
	TLap	47.87	83.10	4.76	24.13	3.96	5.23	3.70	0.00
	GDP	48.29	82.45	3.96	23.64	3.12	12.87	10.26	0.63
	TGDP	48.29	82.45	7.16	23.84	6.79	11.20	8.84	0.00
$\epsilon = 0.5$	OPT	47.79	82.92	4.25	23.77	3.32	2.59	1.79	0.00
	Lap	47.79	82.98	3.87	23.95	3.01	2.88	1.99	0.12
	TLap	47.79	82.98	4.04	23.95	3.21	2.72	1.92	0.00
	GDP	48.14	82.69	4.01	23.84	3.05	6.45	5.12	0.45
	TGDP	48.14	82.69	4.95	23.84	4.35	5.86	4.67	0.00
$\epsilon = 0.75$	OPT	48.02	83.04	3.94	23.91	2.98	1.80	1.20	0.00
	Lap	47.88	83.01	3.92	23.96	3.00	1.93	1.31	0.06
	TLap	47.88	83.01	3.97	23.96	3.08	1.86	1.28	0.00
	GDP	48.10	82.82	3.99	23.87	3.05	4.30	3.42	0.31
	TGDP	48.10	82.82	4.33	23.87	3.63	4.03	3.23	0.00

Table 5.4: Properties of private one-way frequency tables of the variable *household type* in New York state. The summary statistics reported are: mean values of the private entries under each of the five categories (I) to (V), average Monte Carlo standard deviations of all the private data entries (Ave. SD), mean L_1 loss with respect to true values in Table 5.3 (Ave. L_1 loss) and the proportion of 500 private tables with at least one negative entry ('< 0' Proportion).

For the variable *household type* and using the entries from the other eight states, we obtain the proportions under the five categories, $p_0 = (p_{01}, \cdots, p_{05})^T$, where

$p_{01} = 0.196$, $p_{02} = 0.603$, $p_{03} = 0.069$, $p_{04} = 0.122$, $p_{05} = 0.010$ and use them as the null-hypothesized values. Let $P_{MA,\ell}$, $P_{NY,\ell}$, and $P_{NJ,\ell}$, $\ell = 1, 2, \cdots, 5$, denote the population proportions of the five *household type* categories for the states of Massachusetts, New York and New Jersey respectively. Here, we are interested in testing whether the distribution(s) of the variable *household type* for the three east coast states equal the null hypothesized value p_0. Specifically, we have the null hypothesis $H_0 : P_{MA,\ell} = p_{0\ell}, P_{NY,\ell} = p_{0\ell}, P_{NJ,\ell} = p_{0\ell}$, for $\ell = 1, \cdots, 5$. The goodness-of-fit testing procedures proposed in Section 5.4 are directly applicable here for individual state and for the complex H_0 for all three states as stated above. We apply the test statistics constructed using results from Theorems V.8 and V.10. For the Opt mechanism, bias correction using Algorithm 9 is applied. For the all the private mechanisms tested, we add the three test statistics constructed using the private data for each of the east coast states. From Theorems V.8 and V.10, we know that asymptotically this combined test statistic is equivalent to the sum of 15 weighted Chi-square random variables with one degree of freedom, where the weights can be evaluated according to our results. To verify the effectiveness of our proposed procedure under this finite-sample setting, we conduct a parallel simulation study. First, we generate new data sets assuming that the data generating process is as in the H_0 and the sample sizes are the same as those of the three east coast states. To explore the statistical powers, we simulate data sets according to the alternative hypotheses H_1: $P_{NJ,2} = 0.603, 0.603 + 0.005, ..., 0.603 + 0.05$, $P_{MA,1} = 0.196, 0.196 - 0.005, ..., 0.196 - 0.05$, and all the other terms in the H_1 are kept to be the same as in the H_0. We set the level of significance to be 0.05. The

empirical type I errors and powers are summarized in Table 5.5 and in Figure 5.4 respectively. We observe that the empirical type I errors are well controlled across all three mechanisms and for all privacy regimes. The Opt mechanism attains the highest power compared to the other methods, especially when the level of privacy-protection is high.

		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
Household type	Opt	0.055	0.052	0.053
	Lap	0.062	0.055	0.054
	TLap	0.058	0.055	0.054
	GDP	0.050	0.051	0.053
	TGDP	0.033	0.050	0.053
Income level	Opt	0.052	0.051	0.050
	Lap	0.052	0.050	0.052
	TLap	0.051	0.050	0.052
	GDP	0.051	0.051	0.052
	TGDP	0.044	0.050	0.051

Table 5.5: Mean empirical type I errors of the goodness-of-fit test using the private data sets generated when H_0 is true. The reported values are calculated using 500 simulated samples generated according to the NCEDL's settings. Two variables considered are *household type* and *income level*.

Next we compare the p-values obtained using the true data and using the private data. In the simulation, p-values are evaluated on each of the 500 simulated private data tables and the average is reported in Table 5.6. Using the true data, we obtain a p-value of 0.0006, suggesting that the distributions of the variable *household type* differ between the three east coast states and the other eight states considered in the NCEDL study. However, in Table 5.6, we observe that all the p-values yielded from the private data are inflated to some extent, due to the information loss as a result of the random noises injected. When the privacy requirement is high at $\epsilon = 0.25$, none of the methods correctly rejects the potentially wrong H_0. When $\epsilon = 0.5$ and if the level of significance is set to be 0.05, only the private data yielded from the Opt

mechanism can correctly reject H_0. When the privacy regime is set at $\epsilon = 0.75$, the Lap, TLap and Opt procedures produce satisfactory p-values with the Lap and Opt mechanisms yielding a slightly lower average p-value than TLap. All the p-values yielded from private data generated from GDP or TGDP mechanisms do not give correct inference results. The numerical results here suggest that the chances of the testing signals being undetermined increase with the levels of privacy requirements (decrease with ϵ). The Opt mechanism tend to give the smaller deviations from the truth, thus is more preferred to conduct private inferences.

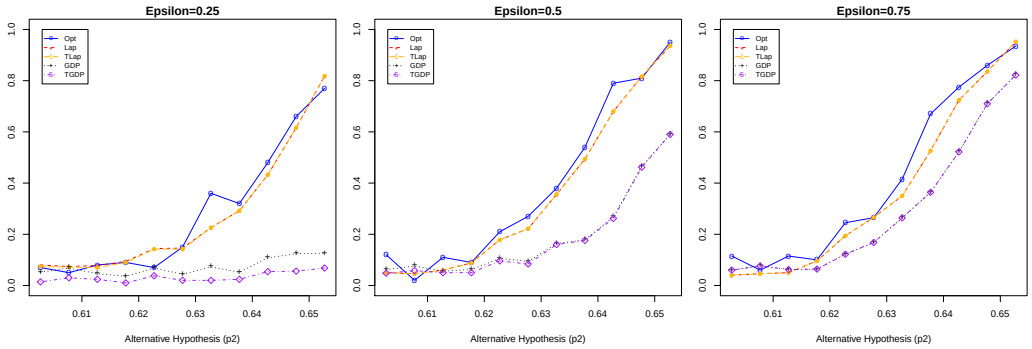

Figure 5.4: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP on household type data.

		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
Household type	Opt	0.372	0.042	0.006
	Lap	0.326	0.068	0.006
	TLap	0.394	0.054	0.010
	GDP	0.450	0.297	0.157
	TGDP	0.673	0.422	0.189
Income level	Opt	0.000	0.000	0.000
	Lap	0.000	0.000	0.000
	TLap	0.000	0.000	0.000
	GDP	0.038	0.000	0.000
	TGDP	0.022	0.000	0.000

Table 5.6: Average p-values of the goodness-of-fit tests using the private data sets in the NCEDL studies. The reported values are calculated using 500 simulated samples. Two variables considered are *household type* and *income level*.

Hereafter, we consider another categorical variable *family income* which has 18 categories describing the income levels of the population. The categories are (1) $\leq \$5,000$, (2) \$5,001 - \$10,000, (3) \$10,001 - \$15,000,..., (18) $\geq \$85,001$. There are 354, 155 and 286 observations for the three east coast states, Massachusetts, New York and New Jersey, respectively. It is a common operation to reduce the total number of categories in a variable when the number of categories is large. The results reported in Theorem V.12 provide theoretical support for reducing the number of categories via combining multiple cells into one. Here, for each state, the data is collapsed into frequency tables with three categories: low income ($\leq \$20,000$), middle income ($\$20,001 - \$50,000$) and high income ($\geq \$50,001$), and denote them as D_{MA}, D_{NY} and D_{NJ} respectively. The three *family income* categories are constructed by respectively merging 4, 6, and 8 out of the original 18 categories into three based on the corresponding income levels.

Goodness-of-fit tests are conducted to check whether the distributions of the variable *family income* for the three east coast states differ from the targeted distribution built from the other states. Following a similar operation as earlier for the variable *household type,* for the 18 categories from the eight other states, the corresponding proportions are $p_0 = (p_{0,1}, \cdots, p_{0,18})^T$, with the value of $p_0 =$ (0.099, 0.101, 0.119, 0.118, 0.106, 0.103, 0.074, 0.043, 0.036, 0.031, 0.020, 0.026, 0.022, 0.013, 0.012, 0.014, 0.013, 0.050)^T. The act of combining categories leads to the null-hypothesized value of $p_0^c = (0.437, 0.393, 0.170)^T$ and the composite $H_0 : P_{MA,\ell} = p_{0,\ell}^c, P_{NY,\ell} = p_{0,\ell}^c, P_{NJ,\ell} = p_{0,\ell}^c$, for $\ell = 1, 2, 3$. For the power evaluation, we construct the alternatives from the original 18 categories. Specifically, we set $p_{1,6} = p_{0,6} - k\triangle$, and $p_{1,4} = p_{0,4} + k\triangle$,

where $\triangle = 0.0025$ and $k = 0, 1, ..., 14$, while keeping the rest of the H_1 the same as the H_0. Note that $p_{0,4}$ has the largest value amongst the 15 cell probabilities. The significance level is set to be 0.05. The empirical type I errors under the H_0 and the statistical powers are shown in Table 5.5 and Figure 5.5, respectively. We observe that the empirical type I errors are well controlled across all different mechanisms and for all privacy regimes. The Opt mechanism attains comparable power to the Lap/TLap method, but much larger than those from the GDP/TGDP mechanisms.

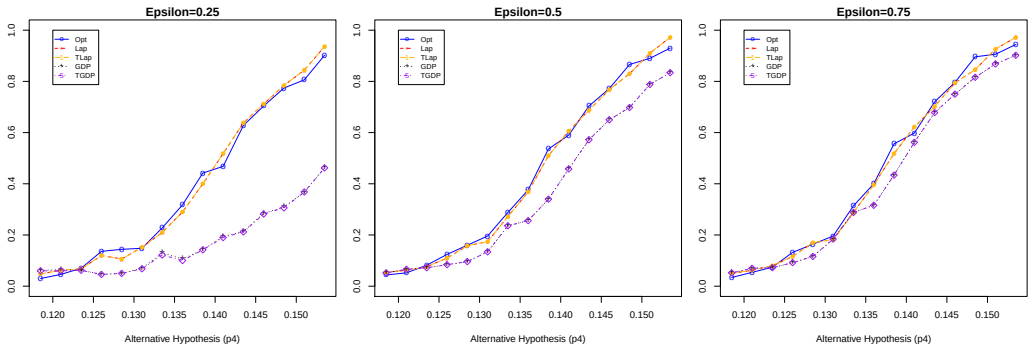

Figure 5.5: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP on the intra-table merged income level data.

Furthermore, we compare the p-values obtained from the true observations and from the private data sets. A p-value of zero is obtained using the true data, suggesting that the distribution for the variable *income level* of the east coast states differs from that of the other eight states in the NCEDL study. For the private data, p-values are evaluated on each of the 500 Monte Carlo samples and the averages are reported in Table 5.6. In this case, the signal of true H_1 is strong enough to be detected for all the mechanisms using a level of significance of 0.05, and the Opt approach attains one of the smallest p-values amongst the five mechanisms.

5.7 Conclusion

In this chapter, we recommend a mechanism satisfying ϵ-DP, specifically applicable to one-way frequency tables, which is optimal in the sense that the expected losses are minimized under a given privacy constraint, where the losses are flexible. Furthermore, we develop valid inference procedures for goodness-of-fit tests for the private data, not only for the optimal mechanism, but also for the Laplace mechanism and the Gaussian mechanism (with/without post-processing of converting negative cells to zero). In fact, the inference procedures developed work for general mechanisms with additive noises. Everyday operations in practice, including merging multiple frequency tables and combining categories within a table, are also considered. The valid inference procedures applicable to the private frequency tables are derived. However, ϵ-DP procedures can be too noisy in practice and it might be desirable to extend the current results to the (ϵ, δ)-DP framework. Currently, the developments of the (ϵ, δ)-DP mainly focus on the Gaussian mechanism, under which the numerical properties are vastly inferior to other mechanisms. The investigation and developments of alternative (ϵ, δ)-DP mechanisms with satisfactory inference characteristics are left as future work to explore.

APPENDICES

APPENDIX A

Appendix of Chapter 2

A.1 Additional Simulation Studies

A.1.1 Estimating Randomly Sampled Q-Matrix

In this section, we consider randomly sampled Q-matrix in a way that can simulate potentially more challenging scenarios. In specific, we include the one-, two- and three-attribute item designs. The exact construction of the Q-matrix is as follows. Similar to the construction in the main article, we still fix the dimension of the Q-matrix to be $3K$ by K, i.e. $3K$ items with K attributes. For each row j, we first determine which item design it will take by a random sampling scheme. Let $M = \binom{K}{1} + \binom{K}{2} + \binom{K}{3}$. The number of required attributes (denoted by n) for each item is randomly sampled from $\{1, 2, 3\}$ with probabilities $\{\binom{K}{1}/M, \binom{K}{2}/M, \binom{K}{3}/M\}$. Then, n attributes are sampled without replacement from $\{1, 2, ..., K\}$ with equal

probabilities, the corresponding entries in $\mathbf{q}_j$ will be set to 1 and the rest to 0. Note that this random construction of the Q-matrix would somewhat simulate the extreme situations where the easiest learned one-attribute items will be sampled with the smallest probabilities. For example, when $K = 15$, the probability to select a one-attribute item is only 0.0261. Furthermore, we also point out that under this random design, there will be a high chance the sampled Q-matrix is not identifiable, making the estimation even more difficult. 100 replications for each of $K = 5, 10, ..., 25$ are considered and the average results are presented in Figure A.1. For illustration purpose, we only consider the settings when $N = 2000$ and when the attributes are independent, for the DINA, the ACDM, and a mixture of the DINA, ACDM, and DINO data. For the data from a mixture of three models, the data are generated from the DINA, ACDM and DINO models with proportions 0.35, 0.35, and 0.3, respectively. All other set-ups remain the same as in Chapter II.

From Figure A.1, we can observe that the OE's of our proposed method remain controlled for three types of data. However, we can also see that the OE's worsen and the OTP's become much more volatile compared to the fixed Q-matrix design in Chapter II. This is not surprising because of the increased difficulty in the design where the Q-matrices contain more two- and three-attribute items and the number of non-identifiable Q-matrices increases significantly. In line with our observations in the main article, we also observe the increased uncertainty level impact most negatively on the OTP. However, overall, the proposed method still possesses certain degrees of learning power of the Q-matrix even in such extreme situations.

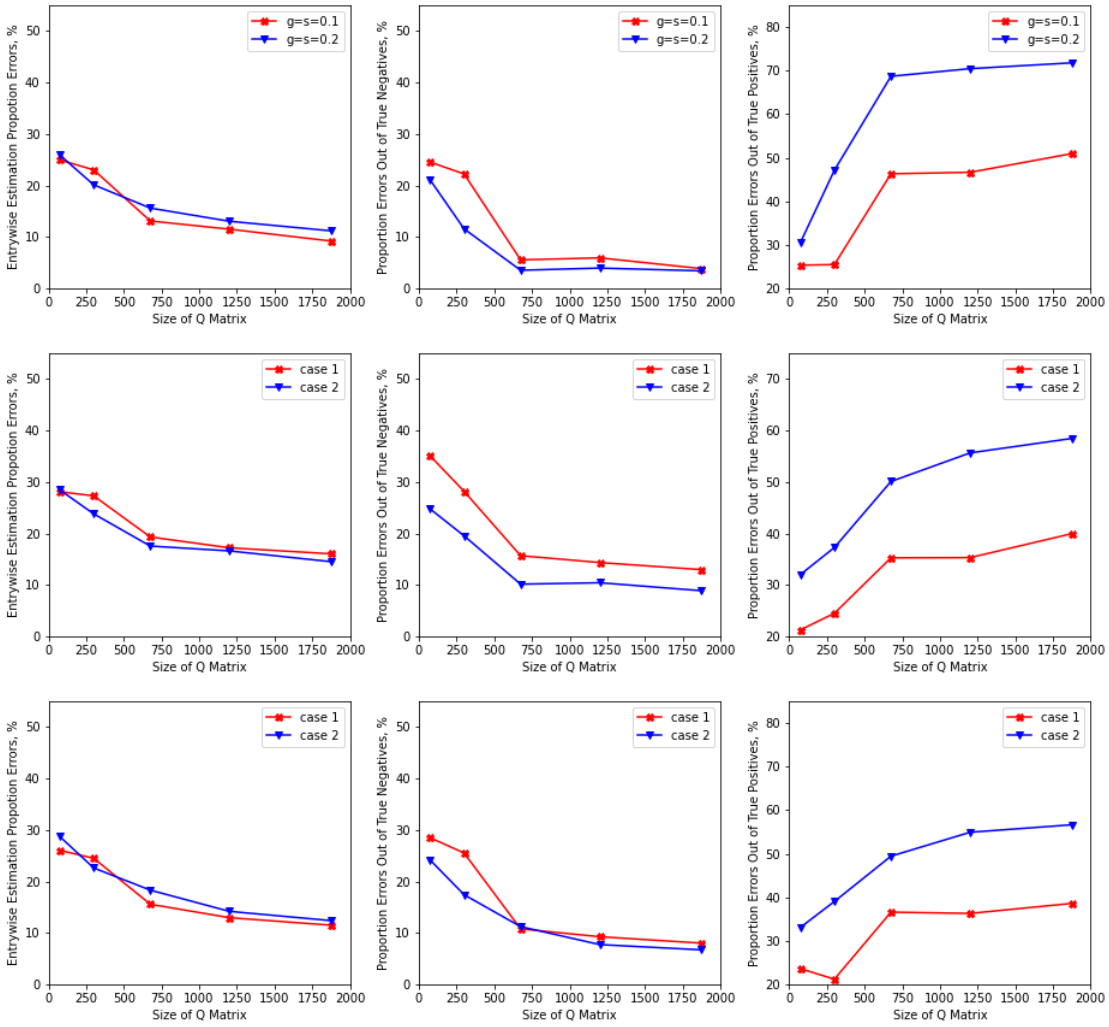

Figure A.1: Plots of different performance metrics against the sizes of the Q-matrix. Rows 1 to 3 correspond to the DINA data, the ACDM data and a mixture of the DINA, ACDM and DINO data, respectively. For the DINA and DINO data, two uncertainty levels are represented by $g_j = s_j = 0.1$ and $g_j = s_j = 0.2$ for all items j, where subscripts j are omitted in the legends. For both the ACDM data and the GDINA data, cases 1 and 2 represent the settings when $\delta_{j,0} = 0.1$, $p_j = 0.9$ and $\delta_{j,0} = 0.2$, $p_j = 0.8$ for all $j = 1, ..., J$ respectively.

A.1.2 Attribute Classifications in Correlated Settings

In this section, we explore the potential of our proposed method in learning the latent attribute patterns. As discussed in Chapter II, the marginal distributions of the latent attributes are mis-specified in RBMs. Therefore, we would like to explore to what extent our proposed method can perform latent attribute classifications directly when the conditional independence assumption is intensely violated. Similarly, ACC rate is used to assess the performance. Recall that the ACC of the k'th attribute is defined as

$$ACC(k) := \frac{1}{N}\sum_{i=1}^{N} |\hat{\alpha}_{ik} - \alpha_{ik}|,$$

where $\hat{\alpha}_{ik}$ and α_{ik} represent the estimated value and the true value respectively.

The simulation set-ups remain the same as in Chapter II. The recovered latent attribute matrix corresponding to the optimal estimated Q-matrix is returned. All the DINA, ACDM and GDINA data are considered. For each of the 100 replications, the ACC rate for every attribute in each of the settings with $K = 5, 10, ..., 25$ is evaluated. The setting-wise average ACC rate is evaluated by computing the average ACC for each attribute out of 100 repetitions first, and then averaging out of all the K latent attributes for each settings of $K = 5, 10, ..., 25$. The results are summarized in Table A.1.

Overall, we can see that the proposed method performs well in attribute classifications with all ACC rates above 0.85. Furthermore, we also observe that the ACC rates drop as the number of attributes increases in the model. The attribute patterns would increase as the number of attributes increments, making the estimation more

difficult. Similar to the observations made in the main article, we see the ACC rates are generally higher when the correlations amongst attributes are higher. We also point out that increasing sample size can in general improve ACC rates using the proposed method. The performance of the proposed method is better on the ACDM data and the GDINA data than on the DINA data. This is especially obvious when K is relatively small at 5 and 10.

$N = 2000$						$N = 10000$					
$\rho = 0.25$			$\rho = 0.75$			$\rho = 0.25$			$\rho = 0.75$		
DINA	ACDM	GDINA	DINA	ACDM	GDINA	DINA	ACDM	GDINA	DINA	ACDM	GDINA
0.898	0.916	0.916	0.917	0.927	0.924	0.903	0.916	0.917	0.918	0.932	0.931
0.897	0.896	0.900	0.888	0.902	0.903	0.901	0.907	0.911	0.885	0.911	0.912
0.878	0.876	0.880	0.880	0.888	0.893	0.891	0.887	0.893	0.880	0.897	0.900
0.875	0.863	0.869	0.879	0.885	0.889	0.883	0.879	0.882	0.874	0.894	0.893
0.866	0.853	0.857	0.875	0.883	0.887	0.877	0.868	0.874	0.874	0.887	0.890

Table A.1: Average ACC rates for using RBM on the DINA data, the ACDM data and the GDINA data. Rows 1 to 5 correspond to the settings with $K = 5, 10, ..., 25$ respectively.

A.2 Proofs of Lemmas and Propositions

Before proving our main Propositions II.4 and II.5, we first give a lemma which would be used in the proof of the main propositions.

Lemma A.1. *Assume $\boldsymbol{\alpha}$ are independent and $\alpha_k \sim Ber(p_k)$ for $k = 1, ..., K$. If true model with response Y satisfies either the GDINA model Equation (3) or the DINA model $P(Y = 1 \mid \boldsymbol{\alpha}) = g + (1 - s - g)\alpha_1\alpha_2...\alpha_{K^*}$ for some s, g satisfying $g < 1 - s$, then the mis-specified linear additive model of Y regressed on $(\alpha_1, \alpha_2, ..., \alpha_K)$ has the corresponding mean function in the form of $\mathbb{E}^*[Y \mid \boldsymbol{\alpha}] = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + ... + \beta_K\alpha_K$ with $\beta_k = 0$ for $k = K^* + 1, ..., K$.*

Proof of Lemma A.1. By the independence assumption and the linear regression theory, we have for $k = 1, \ldots, K$,

$$\begin{aligned}\beta_k &= \frac{1}{Var(\alpha_k)} Cov\big(\alpha_k, Y\big) \\ &= \frac{1}{p_k(1-p_k)} Cov\big(\alpha_k, Y\big).\end{aligned}$$

Denote $\alpha_{1,\ldots,K^*} := \{\alpha_1, ..., \alpha_{K^*}\}$, then by the Law of Total Covariance, we have for $k = K^* + 1, ..., K$,

$$Cov\big(\alpha_k, Y\big) = \mathbb{E}\left[Cov\big(\alpha_k, Y \mid \alpha_{1,\ldots,K^*}\big)\right] + Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,\ldots,K^*}\right], \mathbb{E}\left[Y \mid \alpha_{1,\ldots,K^*}\right]\big). \tag{A.1}$$

Applying the independence assumption again, we have

$$Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,\ldots,K^*}\right], \mathbb{E}\left[Y \mid \alpha_{1,\ldots,K^*}\right]\big) = Cov\big(p_k, \mathbb{E}[Y \mid \alpha_{1,\ldots,K^*}]\big) = 0.$$

Hence, we only need to consider the first term of (A.1). Referring to Figure A.2, we know that in both the DINA and the GDINA model setting, $Y \perp\!\!\!\perp \alpha_k \mid \alpha_{1,\ldots,K^*}$ for all $k = K^* + 1, ..., K$.

$$\mathbb{E}\left[Cov\big(\alpha_k, Y \mid \alpha_{1,\ldots,K^*}\big)\right] = 0.$$

Therefore,

$$\beta_k = \frac{0}{p_k(1-p_k)} = 0 \quad \forall k = K^* + 1, ..., K.$$

□

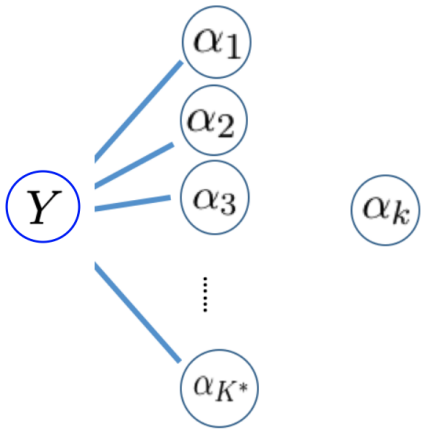

Figure A.2: Illustration of the conditional independence relationship between Y and α_k given $\alpha_1, ..., \alpha_{K^*}$ for all $k = K^* + 1, ..., K$

.

Next we give the proofs of our main propositions.

Proof of Proposition II.4. First note that by Lemma A.1, we have $\beta_k = 0$ for $k = K^* + 1, ..., K$.

In the DINA setting, we have

$$P(Y = 1 \mid \boldsymbol{\alpha}) = \begin{cases} 1 - s & \text{if } \boldsymbol{\alpha} \succcurlyeq \mathbf{1}_{K^*} \\ g & \text{otherwise,} \end{cases}$$

or,

$$Y \mid \boldsymbol{\alpha} \sim \begin{cases} \text{Ber}(1 - s) & \text{if } \boldsymbol{\alpha} \succcurlyeq \mathbf{1}_{K^*} \\ \text{Ber}(g) & \text{otherwise.} \end{cases} \tag{A.2}$$

Under the independence condition, for any $k = 1, ..., K^*$, we have

$$\beta_k = \frac{1}{Var(\alpha_k)} Cov\big(\alpha_k, Y\big) = \frac{1}{p_k(1-p_k)} Cov(\alpha_k, Y).$$

Consider the following two events which partition the sample space of $\boldsymbol{\alpha}$,

$E_{0,k} := \big\{\alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ..., \alpha_{K^*} \mid \prod_{i=1, i\neq k}^{K^*} \alpha_i = 0\big\}$ and

$E_{1,k} := \big\{\alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ..., \alpha_{K^*} \mid \prod_{i=1, i\neq k}^{K^*} \alpha_i = 1\big\}$.

Denote $\alpha_{1,...,K^*\backslash k} := \big\{\alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ..., \alpha_{K^*}\big\}$. By the Law of Total Covariance, we have

$$Cov\big(\alpha_k, Y\big) = \mathbb{E}\left[Cov\big(\alpha_k, Y \mid \alpha_{1,...,K^*\backslash k}\big)\right] + Cov\big(\mathbb{E}\left[\alpha_k | \alpha_{1,...,K^*\backslash k}\right], \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\big). \tag{A.3}$$

Applying the independence condition,

$$Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,...,K^*\backslash k}\right], \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\big) = Cov\big(p_k, \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\big) = 0.$$

Hence, we only need to consider the first term of (A.3),

$$\begin{aligned} &\mathbb{E}\left[Cov\big(\alpha_k, Y \mid \alpha_{1,...,K^*\backslash k}\big)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid \alpha_{1,...,K^*\backslash k}\right] - \mathbb{E}\left[\alpha_k \mid \alpha_{1,...,K^*\backslash k}\right] \cdot \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\right]. \end{aligned} \tag{A.4}$$

For a fixed k, define another two events: $E_{2,k} := \big\{\boldsymbol{\alpha} \mid \alpha_k = 0\big\}$ and $E_{3,k} := \big\{\boldsymbol{\alpha} \mid$

$\alpha_k = 1\big\}$. Then in the event of $E_{0,k}$,

$$\begin{aligned}
(\text{A.4}) =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E_{0,k}\right] - \mathbb{E}\left[\alpha_k \mid E_0\right]\mathbb{E}\left[Y \mid E_{0,k}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E_{0,k}, E_{3,k}\right]P(E_{3,k}) + \mathbb{E}\left[\alpha_k Y \mid E_{0,k}, E_{2,k}\right]P(E_{2,k})\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[Y \mid E_{0,k}\right]\right] \\
=& \mathbb{E}\left[g \cdot p_k - p_k \cdot g\right] \\
=& 0.
\end{aligned}$$

In the event of $E_{1,k}$,

$$\begin{aligned}
(\text{A.4}) =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E_{1,k}\right] - \mathbb{E}\left[\alpha_k \mid E_{1,k}\right]\mathbb{E}\left[Y \mid E_{1,k}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E_{1,k}, E_{3,k}\right]P(E_{3,k}) + \mathbb{E}\left[\alpha_k Y \mid E_{1,k}, E_{2,k}\right]P(E_{2,k})\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right] \cdot \mathbb{E}\left[Y \mid E_{1,k}, E_{3,k}\right] \cdot P(E_{3,k}) - \mathbb{E}\left[\alpha_k\right] \cdot \mathbb{E}\left[Y \mid E_{1,k}, E_{2,k}\right] \cdot P(E_{2,k})\right] \\
=& \mathbb{E}\left[(1-s)p_k + 0 - p_k(1-s)p_k - p_k g(1-p_k)\right] \\
=& p_k(1-p_k)(1-s-g).
\end{aligned}$$

Since the above reasoning works for any $k = 1, 2, ..., K^*$, we must have for each

$k = 1, 2, ..., K^*$,

$$\begin{aligned}
\beta_k &= \frac{1}{p_k(1-p_k)} Cov\big(\alpha_k, Y\big) \\
&= \frac{1}{p_k(1-p_k)} \big(0 \cdot P(E_{0,k}) + p_k(1-p_k)(1-s-g) \cdot P(E_{1,k})\big) \\
&= (1-s-g) \prod_{i=1, i\neq k}^{K^*} p_i \\
&\neq 0.
\end{aligned}$$

□

Proof of Proposition II.5. Note that by Lemma A.1, we have $\beta_k = 0$ for $k = K^* + 1, ..., K$.

Under the independence condition, for any $k = 1, ..., K^*$, we have

$$\begin{aligned}
\beta_k &= \frac{1}{Var(\alpha_k)} Cov\big(\alpha_k, Y\big) \\
&= \frac{1}{p_k(1-p_k)} Cov(\alpha_k, Y). \qquad \text{(A.5)}
\end{aligned}$$

Denote $S := \big\{1, 2, 3, ..., K^*\big\}$. We consider the following 2^{K^*} events: $E_0 := \big\{\boldsymbol{\alpha} \mid \alpha_l = 0, \forall l \in S\big\}$, $E_{1,i} := \big\{\boldsymbol{\alpha} \mid \alpha_i = 1, \alpha_j = 0, \forall j \neq i \in S\big\}$ for some $i \in S$ (i.e. events that only one of the required variables taking value of 1 and all others being 0), $E_{2,(i,j)} := \big\{\boldsymbol{\alpha} \mid \alpha_i = \alpha_j = 1, \alpha_k = 0, \forall k \neq i, j \in S\big\}$ for some $i \neq j \in S$ (i.e. events that any two of the required variables are 1 and all others being 0), ..., $E_{K^*} := \big\{\boldsymbol{\alpha} \mid \alpha_l = 1, \forall l \in S\big\}$. Note that $E_0, E_{1,i}$ for $i \in S$, $E_{2,(i,j)}$ for some $i \neq j \in S$, ..., E_{K^*} partition the sample space of $\boldsymbol{\alpha}$. The response R would have the following

distribution.

$$Y|\boldsymbol{\alpha} \sim \begin{cases} \text{Ber}(\delta_0) & \text{if } E_0 \\ \text{Ber}(\delta_0 + \delta_i) & \text{if } E_{1,i} \\ \text{Ber}(\delta_0 + \delta_i + \delta_j + \delta_{i,j}) & \text{if } E_{2,(i,j)} \\ ... & \\ \text{Ber}\big(\delta_0 + \sum_{k=1}^{K^*} \delta_k + ... + \delta_{12...K^*}\big) & \text{if } E_{K^*}. \end{cases} \tag{A.6}$$

By the Law of Total Covariance, we have

$$Cov\big(\alpha_k, Y\big) = \mathbb{E}\left[Cov\big(\alpha_k, Y \mid \alpha_{1,...,K^*\backslash k}\big)\right] + Cov\big(\mathbb{E}\left[\alpha_k|\alpha_{1,...,K^*\backslash k}\right], \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\big). \tag{A.7}$$

Similar to the DINA case, we also have

$$Cov\big(\mathbb{E}\left[\alpha_k \mid \alpha_{1,...,K^*\backslash k}\right], \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\big) = Cov\big(p_k, \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\big) = 0.$$

Hence, we only need to consider the first term of (A.7),

$$\begin{aligned} &\mathbb{E}\left[Cov\big(\alpha_k, Y \mid \alpha_{1,...,K^*\backslash k}\big)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid \alpha_{1,...,K^*\backslash k}\right] - \mathbb{E}\left[\alpha_k \mid \alpha_{1,...,K^*\backslash k}\right] \cdot \mathbb{E}\left[Y \mid \alpha_{1,...,K^*\backslash k}\right]\right]. \end{aligned} \tag{A.8}$$

Fix a $k \in S$. Let $S' := \big\{1, 2, ..., k-1, k+1, ..., K^*\big\}$. We can define new 2^{K^*-1} events: $E_0^* := \big\{\alpha_{1,...,K^*\backslash k} \mid \alpha_l = 0 \quad \forall l \in S'\big\}$, $E_{1,i}^* := \big\{\alpha_{1,...,K^*\backslash k} \mid \alpha_i = 1, \alpha_l = 0, \forall l \neq i \in S'\big\}$

for some $i \in S'$, $E^*_{2,(i,j)} := \left\{\alpha_{1,\ldots,K^*\setminus k} \mid \alpha_i = \alpha_j = 1, \alpha_l = 0, \forall l \neq i,j \in S'\right\}$ for some $i \neq j \in S'$,..., $E^*_{K^*-1} := \left\{\alpha_{1,\ldots,K^*\setminus k} \mid \alpha_l = 1 \quad \forall l \in S'\right\}$. And define $E'_0 := \left\{\boldsymbol{\alpha} \mid \alpha_k = 0\right\}$ and $E'_1 := \left\{\boldsymbol{\alpha} \mid \alpha_k = 1\right\}$.

In the event of E^*_0,

$$\begin{aligned}
\text{(A.8)} =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E^*_0\right] - \mathbb{E}\left[\alpha_k \mid E^*_0\right]\mathbb{E}\left[Y \mid E^*_0\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E^*_0, E'_1\right]P(E'_1) + \mathbb{E}\left[\alpha_k Y \mid E^*_0, E'_0\right]P(E'_0)\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[Y \mid E^*_0, E'_1\right]P(E'_1) - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[Y \mid E^*_0, E'_0\right]P(E'_0)\right. \\
=& \mathbb{E}\left[(\delta_0+\delta_k)p_k + (1-p_k)\cdot 0 - (\delta_0+\delta_k)p_k^2 - \delta_0(1-p_k)p_k\right] \\
=& p_k(1-p_k)\delta_k.
\end{aligned}$$

In the event of $E^*_{1,i}$ for some $i \in S'$,

$$\begin{aligned}
\text{(A.8)} =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E^*_{1,i}\right] - \mathbb{E}\left[\alpha_k \mid E^*_{1,i}\right]\mathbb{E}\left[Y \mid E^*_{1,i}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E^*_{1,i}, E'_1\right]P(E'_1) + \mathbb{E}\left[\alpha_k Y \mid E^*_{1,i}, E'_0\right]P(E'_0)\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[Y \mid E^*_{1,i}, E'_1\right]P(E'_1) - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[Y \mid E^*_{1,i}, E'_0\right]P(E'_0)\right] \\
=& \mathbb{E}\left[(\delta_0+\delta_i+\delta_k+\delta_{ik})p_k + (1-p_k)\cdot 0\right. \\
& \left. - (\delta_0+\delta_i+\delta_k+\delta_{ik})p_k^2 - (\delta_0+\delta_i)(1-p_k)p_k\right] \\
=& p_k(1-p_k)(\delta_k+\delta_{ik}).
\end{aligned}$$

In the event of $E^*_{2,(i,j)}$ for some $i \neq j \in S'$,

$$
\begin{aligned}
(\text{A.8}) =& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E^*_{2,(i,j)}\right] - \mathbb{E}\left[\alpha_k \mid E^*_{2,(i,j)}\right]\mathbb{E}\left[Y \mid E^*_{2,(i,j)}\right]\right] \\
=& \mathbb{E}\left[\mathbb{E}\left[\alpha_k Y \mid E^*_{2,(i,j)}, E'_1\right]P(E'_1) + \mathbb{E}\left[\alpha_k Y \mid E^*_{2,(i,j)}, E'_0\right]P(E'_0)\right. \\
& \left. - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[Y \mid E^*_{2,(i,j)}, E'_1\right]P(E'_1) - \mathbb{E}\left[\alpha_k\right]\mathbb{E}\left[Y \mid E^*_{2,(i,j)}, E'_0\right]P(E'_0)\right] \\
=& \mathbb{E}\left[(\delta_0 + \delta_i + \delta_j + \delta_k + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk})p_k + (1-p_k)\cdot 0\right. \\
& \left. - (\delta_0 + \delta_i + \delta_j + \delta_k + \delta_{ij} + \delta_{ik} + \delta_{jk} + \delta_{ijk})p_k^2 - (\delta_0 + \delta_i + \delta_j + \delta_{ij})(1-p_k)p_k\right] \\
=& p_k(1-p_k)(\delta_k + \delta_{ik} + \delta_{jk} + \delta_{ijk}).
\end{aligned}
$$

Continuing this process and substitute the relevant values into Equation (A.5), we can show that

$$
\beta_k = \begin{cases}
\delta_k & \text{if } E^*_0 \\
\delta_k + \delta_{ik} & \text{if } E^*_{1,i} \\
\delta_k + \delta_{ik} + \delta_{jk} + \delta_{ijk} & \text{if } E^*_{2,(i,j)} \\
... & \\
\delta_k + \sum_{i=1,i\neq k}^{K^*}\delta_{ik} + ... + \delta_{1...K^*} & \text{if } E^*_{K^*-1}.
\end{cases} \tag{A.9}
$$

Since the above holds for all $k = 1, 2, 3, ...K^*$, we have for each $k = 1, 2, 3, ...K^*$,

$$
\begin{aligned}
\beta_k =& \delta_k \cdot P(E^*_0) + \sum_{i\in S'}(\delta_k + \delta_{ik})\cdot P(E^*_{1,i}) + \sum_{i,j\in S', i\neq j}(\delta_k + \delta_{ik} + \delta_{jk} + \delta_{ijk})\cdot P(E^*_{2,(i,j)}) + \\
& ... + \left(\delta_k + \sum_{i=1,i\neq k}^{K^*}\delta_{ik} + ... + \delta_{1...K^*}\right)\cdot P(E^*_{K^*-1})
\end{aligned} \tag{A.10}
$$

Assuming monotonicity in acquiring an additional skill, we can show all the terms in (A.10) are greater than 0. The first term is positive as both δ_k and $P(E_0^*)$ are positive. To see why the second term is positive, consider two examinees, one with skill set $\boldsymbol{\alpha}_1 = \{\boldsymbol{\alpha} \mid \alpha_i = 1, \alpha_l = 0, \quad \forall l \neq i \in S\}$ while the other with skill set $\boldsymbol{\alpha}_2 = \{\boldsymbol{\alpha} \mid \alpha_i = \alpha_k = 1, \alpha_l = 0, \quad \forall l \neq i, k \in S\}$. Then we know according to Equation (3), $P(Y = 1 \mid \boldsymbol{\alpha_1}) = \delta_0 + \delta_i$ and $P(Y = 1 \mid \boldsymbol{\alpha_2}) = \delta_0 + \delta_i + \delta_k + \delta_{ik}$. The monotonicity assumption then implies $P(Y = 1 \mid \boldsymbol{\alpha_2}) - P(Y = 1 \mid \boldsymbol{\alpha_1}) = \delta_k + \delta_{ik} > 0$. Hence the second term is positive. We can use a similar strategy to show all the terms in (A.10) are positive and thus reach the conclusion that $\beta_k \neq 0$ for each $k = 1, 2, 3, ...K^*$. □

Discussion of Remark II.7. Conditional on $\alpha_1, \alpha_2, ..., \alpha_{K^*}$, consider adding one α_k, for any $k = K^* + 1, ..., K$, into the main effect regression model, then its coefficient can be expressed as

$$\beta_k = \frac{Cov\Big(Y - \mathbb{E}^*[Y \mid \alpha_1, ..., \alpha_{K^*}], \quad \alpha_k - \mathbb{E}^*[\alpha_k \mid \alpha_1, ..., \alpha_{K^*}]\Big)}{Var\Big(Y - \mathbb{E}^*[Y \mid \alpha_1, ..., \alpha_{K^*}]\Big)},$$

where $\mathbb{E}^*[A \mid B]$ is the the regression mean function of A on B. In the special case when $K^* = 1$, we seek to show $\beta_k = 0$. When $K^* = 1$, note that we must have $\mathbb{E}^*[Y \mid \alpha_1] = \mathbb{E}[Y \mid \alpha_1]$. This is because α_1 can only take values of 0 or 1. These two variability's can be modeled exhaustively by the free intercept and the only coefficient in the regression mean function. Note that when $K^* > 1$, this may

not hold in general. Note by the Law of Total Covariance,

$$
\begin{aligned}
& Cov\Big(Y - \mathbb{E}^*[Y \mid \alpha_1], \quad \alpha_k - \mathbb{E}^*[\alpha_k \mid \alpha_1]\Big) \\
=& \mathbb{E}\Big\{Cov\Big(Y - \mathbb{E}[Y \mid \alpha_1], \quad \alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\Big)\Big\} && \text{(A.11)} \\
& + Cov\Big\{\mathbb{E}\big(Y - \mathbb{E}[Y \mid \alpha_1] \mid \alpha_1\big), \quad \mathbb{E}\big(\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\big)\Big\}. && \text{(A.12)}
\end{aligned}
$$

Note (A.12) $= 0$ and

$$
\begin{aligned}
\text{(A.11)} =& \mathbb{E}\Big\{\mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid \alpha_1]\big)\big(\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1]\big) \mid \alpha_1\Big] \\
& + \mathbb{E}\Big[\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\Big]\mathbb{E}\Big[\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\Big]\Big\} \\
=& \mathbb{E}\Big\{\mathbb{E}\Big[(Y - \mathbb{E}[Y \mid \alpha_1])(\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1]) \mid \alpha_1\Big]\Big\} \\
=& \mathbb{E}\Big\{\mathbb{E}\Big[Y\alpha_k - Y\mathbb{E}(\alpha_k \mid \alpha_1) - \alpha_k\mathbb{E}(Y \mid \alpha_1) + \mathbb{E}(Y \mid \alpha_1)\mathbb{E}(\alpha_k \mid \alpha_1) \mid \alpha_1\Big]\Big\} \\
=& \mathbb{E}\Big\{\mathbb{E}[Y\alpha_k \mid \alpha_1] - \mathbb{E}[Y\alpha_k \mid \alpha_1] - \mathbb{E}[Y\alpha_k \mid \alpha_1] + \mathbb{E}[Y\alpha_k \mid \alpha_1]\Big\} \\
=& 0.
\end{aligned}
$$

Where the second line follows from $\mathbb{E}\big\{\alpha_k - \mathbb{E}[\alpha_k \mid \alpha_1] \mid \alpha_1\big\} = 0$ and the third line follows from the fact that $\mathbb{E}[Y \mid \alpha_1]\mathbb{E}[\alpha_k \mid \alpha_1] = \mathbb{E}[Y\alpha_k \mid \alpha_1]$ by the conditional independence between Y and α_k given α_1. Therefore, $\beta_k = 0$. □

APPENDIX B

Appendix of Chapter 3

Appendix B contains the proofs for theoretical results developed in Chapter III. In specific, Section B.1 contains proofs of theorems and proposition, Section B.2 contains the proofs of the supporting lemmas, and a full list of rankings of senators considered in real data application Section 3.5.2 according to their conservativeness scores is included in Section B.3.

B.1 Proof of Theorems and Proposition

Proof of Theorem III.10. We start with defining some notations that are needed in the proofs. Implicitly index J with N such that $J_N \to \infty$ as $N \to \infty$ for notation convenience. Note that this does not impose any rate requirement for N and J. Let $\Omega_N = \big\{x = (x_{ij} : z_{ij} = 1, i = 1, ..., N, j = 1, ..., J) : x_{ij} = \theta_i - \beta_j, \theta_i, \beta_j \in \mathbb{R}, \sum_{i=1}^N \theta_i = 0\big\}$ be a vector space. Define on Ω_N a variance weighted inner product $[\cdot, \cdot]_\sigma$ with $[x, y]_\sigma = \sum_{i=1}^N \sum_{j \in S_J(i)} x_{ij} \sigma_{ij}^2 y_{ij}$ for any $x, y \in \Omega_N$, where $S_J(i) = \{j =$

$1, ..., J : z_{ij} = 1\}$, $\sigma_{ij}^2 = \exp(m_{ij}^*)/\{1 + \exp(m_{ij}^*)\}^2$ and the subscript σ means the inner product depends on $\sigma_{ij}^2, i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$. Denote the associated norm as $\|\cdot\|_\sigma$ with $\|x\|_\sigma^2 = \sum_{i=1}^N \sum_{j \in S_J(i)} x_{ij}^2 \sigma_{ij}^2$ for $x \in \Omega_N$. Let $M_N = \big(m_{ij} : z_{ij} = 1, i = 1, ..., N, j = 1, ..., J, m_{ij} = \theta_i - \beta_j\big) \in \Omega_N$, $M_N^* = \big(m_{ij}^* : z_{ij} = 1, i = 1, ..., N, j = 1, ..., J, m_{ij}^* = \theta_i^* - \beta_j^*\big) \in \Omega_N$ and $\hat{M}_N = \big(\hat{m}_{ij} : z_{ij} = 1, i = 1, ..., N, j = 1, ..., J, \hat{m}_{ij} = \hat{\theta}_i - \hat{\beta}_j\big) \in \Omega_N$. Note that as a result of Proposition III.9, for any linear form g of M, $g(M)$ can be re-expressed as a linear form of $x \in \Omega_N$, with $g(x) = \sum_{i=1}^N \sum_{j \in S_J(i)} w_{ij} x_{ij}$, where we denote $w_{ij} = w_{ij}(g)$, which depends on g, for notation simplicity. Let Ω_N^* consist of all linear forms g on Ω_N such that $g(x) = 0$ if $x = 0$ and $x \in \Omega_N$. Without loss of generality, we will work with $g \in \Omega_N^*$ in the proofs. For any subset $A \subset \Omega_N^*$, define $\|\cdot\|_\sigma(A)$ to be the norm on Ω_N such that for any $x \in \Omega_N$, $\|x\|_\sigma(A)$ is the smallest non-negative number such that $|g(x)| \leq \|x\|_\sigma(A)\sigma(g)$ for any $g \in A$, where $\sigma(g) = \sup_{x \in \Omega_N}\{|g(x)| : \|x\|_\sigma \leq 1\}$. Let

$$E_N = \Big(E_{ij} : z_{ij} = 1, i = 1, ..., N, j = 1, ..., J\Big),$$

with $E_{ij} = \mathbb{E}[Y_{ij}] = e^{m_{ij}^*}/(1+e^{m_{ij}^*})$, be the vector of expected responses corresponding to the observed entries. Further define a residual alike vector $R_N \in \Omega_N$ satisfying

$$[x, R_N]_\sigma = \sum_{i=1}^N \sum_{j \in S_J(i)} x_{ij}(Y_{ij} - E_{ij}), \qquad x \in \Omega_N.$$

Define an evaluation measure $U_N(\cdot,\cdot)$ such that for any $y, v \in \Omega_N$, $U_N(y,v) \in \Omega_N$ is defined by the equation

$$[x, U_N(y,v)]_\sigma = \sum_{i=1}^{N} \sum_{j \in S_J(i)} x_{ij}\{\sigma^2(y_{ij}) - \sigma_{ij}^2\} v_{ij}, \qquad x \in \Omega_N,$$

where $\sigma^2(y_{ij}) = e^{y_{ij}}/(1+e^{y_{ij}})^2$. Note when y is equal to M_N^* or when v is a zero vector, then $U_N(y,v) = 0$. Further denote that $w_{i+} = \sum_{j \in S_J(i)} w_{ij}$, $w_{+j} = \sum_{i \in S_N(j)} w_{ij}$ and $w_{++} = \sum_{i=1}^{N} \sum_{j \in S_J(i)} w_{ij}$, where $S_N(j) = \{i = 1, ..., N : z_{ij} = 1\}$. We then extend the results in *Haberman* (1977) to prove the existence and consistency of the maximum likelihood estimator $\hat{M}_N$.

We first establish the existence of $\hat{M}_N$ by applying the fixed point theorems of *Kantorovich and Akilov* (1964). We start with constructing a function F_N on Ω_N with a fixed point $\hat{M}_N$. Consider $F_N(y) = y + r_N(y)$ for $y \in \Omega_N$, where $r_N : \Omega_N \mapsto \Omega_N$ is defined by the equation,

$$[x, r_N(y)]_\sigma = \sum_{i=1}^{N} \sum_{j \in S_J(i)} x_{ij}\{Y_{ij} - E(y_{ij})\}, \qquad x \in \Omega_N,$$

where $E(y_{ij}) = e^{y_{ij}}/(1 + e^{y_{ij}})$. Note that F_N has a fixed point $\omega \in \Omega_N$ if and only if

$$\sum_{i=1}^{N} \sum_{j \in S_J(i)} x_{ij}\{Y_{ij} - E(\omega_{ij})\} = 0, \qquad x \in \Omega_N.$$

Let P be the orthogonal projection onto Ω_N. Let $\hat{E} = \{E(\hat{m}_{ij}) : i = 1, ..., N, j = 1, ..., J, z_{ij} = 1\}$ and $Y_z = \{Y_{ij} : i = 1, ..., N, j = 1, ..., J, z_{ij} = 1\}$. Then following from

Berk (1972), $\hat{M}_N$ is a maximum likelihood estimator of M_N^* if and only if $P\hat{E} = PY_z$. Hence, $\hat{M}_N$ exists if and only if ω exists. Furthermore, since the log-likelihood $l(Y_z, \cdot)$ is strictly concave, if the maximum likelihood estimator $\hat{M}_N$ of M_N^* exists, then it must be unique. Therefore, if $\hat{M}_N$ exists, $\omega = \hat{M}_N$. So, we just need to verify the conditions of the fixed point theorem to show that the fixed point ω indeed exists.

The Kantorovich & Akilov's fixed point theorem requires construction of a sequence that converges to the fixed point. Consider the sequence $\{t_{Nk} : k = 0, 1, ...\}$, with $t_{N0} = M_N^*$ and $t_{N(k+1)} = F_N(t_{Nk})$ for $k = 0, 1, ...$ Note that $t_{N1} = M_N^* + R_N$. To check whether this sequence is well-defined and converges to $\hat{M}_N$, we need to examine the differential dF_{Ny} of F_N at $y \in \Omega_N$. Note that for $y + v \in \Omega_N$,

$$\begin{aligned}[x, F_N(y+v) - F_N(y)]_\sigma &= \sum_{i=1}^{N} \sum_{j \in S_J(i)} x_{ij}\sigma_{ij}^2 \Big[v_{ij} + (\sigma_{ij}^2)^{-1} \big\{ E(y_{ij}) - E(y_{ij} + v_{ij}) \big\} \Big] \\ &= -[x, U_N(y, v)]_\sigma + o(v),\end{aligned}$$

where $o(v)/\|v\|_\sigma \to 0$ as $\|v\|_\sigma \to 0$. It follows that $dF_{Ny}(v) = -U_N(y, v)$. Denote $\|dF_{Ny}\|_\sigma(A)$ to be the smallest nonnegative number such that

$$\|dF_{Ny}(v)\|_\sigma(A) \leq \|dF_{Ny}\|_\sigma(A)\|v\|_\sigma(A), \qquad v \in \Omega_N.$$

Let A_p be the set consisting of all the point maps f_{ij} on Ω_N, i.e. $f_{ij}(x) = x_{ij}$ for any $x \in \Omega_N$. By Lemma B.1(c) below, there exist sequences f_N and d_N such that

$$\|dF_{Ny}\|_\sigma(A_p) \leq d_N \|y - M_N^*\|_\sigma(A_p) \quad \text{whenever} \quad \|y - M_N^*\|_\sigma(A_p) \leq f_N, \quad y \in \Omega_N.$$

Lemma B.1. *Assume Conditions III.6, III.7 and III.8 hold. If* $A_p = \{f_{ij} : i = 1, ..., N, j = 1, ..., J, z_{ij} = 1\}$ *such that* $f_{ij}(x) = x_{ij}$ *for* $x \in \Omega_N$. *Let* $C_N = |A_p|$, *the cardinality of* A_p. *Then there exist sequences* $f_N > 0$ *and* $d_N \geq 0$ *satisfying the followings.*

(a). As $N \to \infty$, $f_N^2 / \log C_N \to \infty$.

(b). As $N \to \infty$, $f_N^2 (N_*^{-1} + J_*^{-1}) \to 0$.

(c). If $y, v \in \Omega_N$ *and* $\|y - M_N^*\|_\sigma(A_p) \leq f_N$, *then there exists* $n < \infty$ *such that for all* $N > n$, $\|U_N(y, v)\|_\sigma(A_p) \leq d_N \|y - M_N^*\|_\sigma(A_p) \|v\|_\sigma(A_p)$. *Furthermore,* $d_N f_N \to 0$ *as* $N \to \infty$.

As shown in Kantorovich & Akilov (1964, pages 695-711), if $\|R_N\|_\sigma(A_p) < \frac{1}{2} f_N$ and $d_N \|R_N\|_\sigma(A_p) < \frac{1}{2}$, then $\hat{M}_N$ exists. By Lemma B.2 below, we have $\mathrm{P}(\|R_N\|_\sigma(A_p) < \frac{1}{2} f_N) \to 1$ as $N \to \infty$. Therefore, it follows from Lemma B.1(c) that with probability tending to 1, $d_N \|R_N\|_\sigma(A_p) < \frac{1}{2} f_N d_N \to 0$.

Lemma B.2. *Let* $A \subset \Omega_N^*$. *Let* C_N *denote the cardinality of* A. *If there exist sequences* $f_N > 0$ *and* $d_N \geq 0$ *satisfying (a).* $0 < C_N < \infty$ *and* $f_N^2 / \log C_n \to \infty$ *as* $N \to \infty$, *(b). If* $y, v \in \Omega_N$ *and* $\|y - M_N^*\|_\sigma(A) \leq f_N$, *then there exists* $n < \infty$ *such that for all* $N > n$, $\|U_N(y, v)\|_\sigma(A) \leq d_N \|y - M_N^*\|_\sigma(A) \|v\|_\sigma(A)$, *(c).* $d_N f_N \to 0$ *as* $N \to \infty$. *Then* $P\big(\|R_N\|_\sigma(A) < \frac{1}{2} f_N\big) \to 1$ *as* $N \to \infty$.

Hence, the conditions of the fixed point theorem are satisfied with probability approaching 1. It then follows that the maximum likelihood estimators $\hat{M}_N$ exists with probability tending to 1. Since Condition III.8 holds, as a direct consequence of Proposition III.9, the corresponding maximum likelihood estimators $\hat{\theta}_i$, $i = 1, ..., N$

and $\hat{\beta}_j$, $j = 1, ..., J$ can be uniquely determined given $\hat{M}_N$. Therefore, with probability approaching 1 that they all exist, as $N \to \infty$. The first part of the theorem then follows.

Now we seek to prove the consistency results. Taking sequences f_N and d_N again as satisfying the results in Lemma B.1 and $A = A_p$. Then both Lemmas B.2 and B.3 hold. From the results of Lemmas B.2 and B.3, it can be implied that as $N \to \infty$, with probability tending to 1 that,

$$\|\hat{M}_N - M_N^*\|_\sigma(A_p) = O(f_N). \tag{B.1}$$

From *Haberman* (1977), $\sigma(g)$ is in fact the standard deviation of $g(\hat{M}_N)$. We further note by Lemma B.4 below,

$$\max_{g \in A_p} \sigma(g) \leq \tau_2^{-1}(N_*^{-1} + J_*^{-1})^{\frac{1}{2}}, \tag{B.2}$$

for some $0 < \tau_2 < \infty$.

Lemma B.3. *Assume Conditions III.6, III.7 and III.8 hold. Let $A \subset \Omega_N^*$. If there exist sequences $f_N > 0$ and $d_N \geq 0$ satisfying (a). $P\big(\|R_N\|_\sigma(A) < \frac{1}{2} f_N\big) \to 1$ as $N \to \infty$, (b). If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A) \leq f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y, v)\|_\sigma(A) \leq d_N \|y - M_N^*\|_\sigma(A) \|v\|_\sigma(A)$, (c). $d_N f_N \to 0$ as $N \to \infty$. Then, as $N \to \infty$, with probability approaching 1 that,*

$$\Big| \frac{\|\hat{M}_N - M_N^*\|_\sigma(A)}{\|R_N\|_\sigma(A)} - 1 \Big| \leq d_N^{\frac{1}{2}} \to 0 \quad \textit{and} \quad \|\hat{M}_N - M_N^* - R_N\|_\sigma(A) \leq d_N \|R_N\|_\sigma^2(A).$$

Lemma B.4. *Assume Conditions III.6, III.7 and III.8 hold and* $\sum_{i=1}^{N}\theta_i = 0$*, the asymptotic variance of the maximum likelihood estimator of* m_{ij}^**,* $var(\hat{m}_{ij})$*, for any* $i = 1, ..., N$ *and* $j = 1, ..., J$*, takes the form,*

$$var(\hat{m}_{ij}) = (\sigma_{i+}^2)^{-1} + (\sigma_{+j}^2)^{-1} + O(N_*^{-1}J_*^{-1}) \quad as \quad N \to \infty.$$

Then as $N \to \infty$, we have with probability approaching 1 that

$$\begin{aligned}
\max_{i,j,z_{ij}=1} |\hat{m}_{ij} - m_{ij}^*| &= \max_{i,j,z_{ij}=1} |f_{ij}(\hat{M}_N) - f_{ij}(M_N^*)| \\
&= \max_{i,j,z_{ij}=1} |f_{ij}(\hat{M}_N - M_N^*)| \\
&\leq \max_{i,j,z_{ij}=1} \sigma(f_{ij}) \|\hat{M}_N - M_N^*\|_\sigma(A_p) \\
&\leq \|\hat{M}_N - M_N^*\|_\sigma(A_p) \Big\{ \max_{g \in A_p} \sigma(g) \Big\} \\
&= O\Big\{ f_N \left(N_*^{-1} + J_*^{-1} \right)^{\frac{1}{2}} \Big\} \\
&\to 0. \qquad \text{(B.3)}
\end{aligned}$$

The second last line follows from (B.1) and (B.2) and the last line follows from Lemma B.1(b).

To derive explicit rates of convergence for $\|\hat{\theta} - \theta^*\|_\infty$ and $\|\hat{\beta} - \beta^*\|_\infty$, we adopt a similar approach as in the derivation of convergence of $\max_{i,j:z_{ij}=1} |\hat{m}_{ij} - m_{ij}^*|$. In particular, for the column parameters β_j, we consider linear functions $g_j \in \Omega_N^*$ such that $g_j(x) = \beta_j$. We can construct g_j as follows. The idea is to include all the row parameters θ_i so as to use the identifiability constraint $\sum_{i=1}^{N}\theta_i = 0$. For any

$i \in S_N(j)$, we use $m_{ij} = \theta_i - \beta_j$ in the construction. While for each $i \in S_{N_\phi}(j)$, where $S_{N_\phi}(j) = \{1, 2, ..., N\} \backslash S_N(j)$, by Condition III.8, there must exist $1 \leq i_{i1}, i_{i2}, ..., i_{ik} \leq N$ and $1 \leq j_{i1}, j_{i2}, ..., j_{ik} \leq J$ such that

$$z_{i,j_{i1}} = z_{i_{i1},j_{i1}} = z_{i_{i1},j_{i2}} = z_{i_{i2},j_{i2}} = ... = z_{i_{ik},j_{ik}} = z_{i_{ik},j} = 1.$$

Therefore, we can construct g_j as

$$\begin{aligned} g_j(x) =& -\frac{1}{N}\Big\{ \sum_{i \in S_N(j)} m_{ij} \\ &+ \sum_{i \in S_{N_\phi}(j)} \Big(m_{i,j_{i1}} - m_{i_{i1},j_{i1}} + m_{i_{i1},j_{i2}} - m_{i_{i2},j_{i2}} + ... - m_{i_{ik},j_{ik}} + m_{i_{ik},j} \Big) \Big\} \\ =& \beta_j. \end{aligned}$$

Let $A_\beta = \big\{ g_j : j = 1, ..., J \big\}$. Now consider a sequence f_N satisfying the rate requirements $f_N^2 / \log J \to \infty$ and $f_N^2 N_*^{-1/2} \to 0$ as $N \to \infty$. Then by Lemma B.5 below, we can pick a sequence d_N satisfying Lemma B.5(a) and Lemma B.5(b). Furthermore, by Lemma B.6 below, we know that $\sigma^2(g_j) = (\sigma_{+j}^2)^{-1} + O\big\{ (N_* J_*)^{-1} \big\}$ for any $g_j \in A_\beta$. Therefore, there exist positive $0 < c_1, c_2 < \infty$ and some n such that for all $N > n$,

$$c_1^{-1} N_*^{-\frac{1}{2}} < \max_{j=1,...,J} \sigma(g_j) < c_2^{-1} N_*^{-\frac{1}{2}}.$$

Lemma B.5. *Assume Conditions III.6, III.7 and III.8 hold. If $A_\beta = \{g_j : j = 1, ..., J\}$ such that $g_j \in \Omega_N^*$ and $g_j(x) = \beta_j$ for $x \in \Omega_N$. Let $C_N = |A_\beta| = J$ be*

the cardinality of A_β. *For any positive sequence* f_N *such that* $f_N^2/\log J \to \infty$ *and* $f_N^2 N_*^{-1/2} \to 0$ *as* $N \to \infty$, *there exists a sequence* $d_N \geq 0$ *satisfying the followings.*
(a). If $y, v \in \Omega_N$ *and* $\|y - M_N^*\|_\sigma(A_\beta) \leq f_N$, *then there exists* $n < \infty$ *such that for all* $N > n$, $\|U_N(y,v)\|_\sigma(A_\beta) \leq d_N \|y - M_N^*\|_\sigma(A_\beta)\|v\|_\sigma(A_\beta)$.
(b). $d_N f_N^2 \to 0$ *as* $N \to \infty$.

Lemma B.6. *Assume Conditions III.6, III.7 and III.8 hold and* $\sum_{i=1}^N \theta_i = 0$. *The asymptotic variance of the maximum likelihood estimator of an individual column parameter,* $var(\hat{\beta}_j)$, *asymptotically attains the oracle variance* $(\sigma_{+j}^2)^{-1}$ *in the sense that*

$$var(\hat{\beta}_j) = (\sigma_{+j}^2)^{-1} + O(N_*^{-1} J_*^{-1}) \qquad as \quad N \to \infty.$$

Note that by taking sequences f_N and d_N satisfying the conditions in Lemma B.5 and setting $A = A_\beta$, it can be shown easily that the results of Lemmas B.2 and B.3 still hold. Hence, it can be implied that as $N \to \infty$, with probability tending to 1,

$$\|\hat{M}_N - M_N^*\|_\sigma(A_\beta) = O(f_N).$$

Then as $N \to \infty$, we have with probability approaching 1 that,

$$\begin{aligned}
\max_{j=1,\ldots,J} |\hat{\beta}_j - \beta_j^*| &= \max_{j=1,\ldots,J} |g_j(\hat{M}_N) - g_j(M_N^*)| \\
&= \max_{j=1,\ldots,J} |g_j(\hat{M}_N - M_N^*)| \\
&\leq \|\hat{M}_N - M_N^*\|_\sigma(A_\beta) \max_{j=1,\ldots,J} \sigma(g_j) \\
&< c_2^{-1} N_*^{-\frac{1}{2}} \|\hat{M}_N - M_N^*\|_\sigma(A_\beta) \\
&= O\Big\{(\log J)^{\frac{1}{2}} N_*^{-\frac{1}{2}}\Big\} \quad \text{as} \quad N \to \infty,
\end{aligned}$$

where the last step can be implied from the results in Lemma B.7 below applied to the fact that $\|\hat{M}_N - M_N^*\|_\sigma(A_\beta) = O\big(f_N\big)$, and the rate requirement of f_N in Lemma B.5, where the minimum order of f_N is determined by $f_N^2/\log J \to \infty$ as $N \to \infty$. Therefore,

$$\|\hat{\beta} - \beta^*\|_\infty = O_p\Big\{(\log J)^{\frac{1}{2}} N_*^{-\frac{1}{2}}\Big\}. \tag{B.4}$$

Lemma B.7. *Let a_N and c_N be positive sequences. As $N \to \infty$, suppose that $a_N = O(b_N)$, for any sequence b_N satisfying $b_N/c_N \to \infty$. Then $a_N = O(c_N)$ as $N \to \infty$.*

Now for the row parameters θ_i, we adopt a similar strategy by constructing linear functions $g_i \in \Omega_N^*$ such that $g_i(x) = \theta_i$. In specific, we can construct the linear

function g_i as follows.

$$g_i(x) = \frac{1}{|S_J(i)|} \sum_{j \in S_J(i)} \{m_{ij} + g_j(x)\} = \frac{1}{|S_J(i)|} \sum_{j \in S_J(i)} (\theta_i - \beta_j + \beta_j) = \theta_i,$$

where $|S_J(i)|$ denotes the cardinality of $S_J(i)$. Let A_θ consist of $g_i, i = 1, ..., N$, i.e. $A_\theta = \{g_i : i = 1, ..., N\}$. Take a positive sequence f_N satisfying the rate requirements $f_N^2 / \log N \to \infty$ and $f_N^2 J_*^{-1} \to 0$ as $N \to \infty$, then by Lemma B.8 below, we can pick a sequence d_N satisfying Lemma B.8(a) and Lemma B.8(b). Furthermore, by Lemma B.9 below, we know that $\sigma^2(g_i) = (\sigma_{i+}^2)^{-1} + O\{(N_* J_*)^{-1}\}$ for any $g_i \in A_\theta$. Hence, there exist positive $0 < \gamma_1, \gamma_2 < \infty$ and such that

$$\gamma_1^{-1} J_*^{-1/2} < \max_{i=1,...,N} \sigma(g_i) < \gamma_2^{-1} J_*^{-1/2}.$$

Lemma B.8. *Assume Conditions III.6, III.7 and III.8 hold. If $A_\theta = \{g_i : i = 1, ..., N\}$ such that $g_i \in \Omega_N^*$ and $g_i(x) = \theta_i$ for $x \in \Omega_N$. Let $C_N = |A_\theta| = N$ be the cardinality of A_θ. Then for any positive sequence f_N such that $f_N^2 / \log N \to \infty$ and $J_*^{-1} f_N^2 \to 0$ as $N \to \infty$, there exists a sequence $d_N \geq 0$ satisfying the followings.*
(a) If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A_\theta) \leq f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y, v)\|_\sigma(A_\theta) \leq d_N \|y - M_N^*\|_\sigma(A_\theta) \|v\|_\sigma(A_\theta)$.
(b). $d_N f_N \to 0$ as $N \to \infty$.

Lemma B.9. *Assume Conditions III.6, III.7 and III.8 hold and $\sum_{i=1}^N \theta_i = 0$, the asymptotic variance of an individual row parameter, var$(\hat{\theta}_i)$, asymptotically attains*

oracle variance $(\sigma_{i+}^2)^{-1}$ *in the sense that*

$$var(\hat{\theta}_i) = (\sigma_{i+}^2)^{-1} + O(N_*^{-1}J_*^{-1}) \qquad as \quad N \to \infty.$$

Note that by taking sequences f_N and d_N satisfying the conditions in Lemma B.8 and setting $A = A_\theta$, it can be implied easily that Lemmas B.2 and B.3 still hold. Similarly, from $\mathrm{P}(\|R_N\|_\sigma(A_\theta) < \frac{1}{2}f_N) \to 1$ and the results of Lemma B.3, it can be implied as $N \to \infty$, we have with probability tending to 1 that,

$$\|\hat{M}_N - M_N^*\|_\sigma(A_\theta) = O(f_N).$$

It follows, as $N \to \infty$, we have with probability approaching 1 that,

$$\begin{aligned}
\max_{i=1,\ldots,N} |\hat{\theta}_i - \theta_i| &= \max_{i=1,\ldots,N} |g_i(\hat{M}_N) - g_i(M_N^*)| \\
&= \max_{i=1,\ldots,N} |g_i(\hat{M}_N - M_N^*)| \\
&\leq \|\hat{M}_N - M_N^*\|_\sigma(A_\theta) \max_{i=1,\ldots,N} \sigma(g_i) \\
&< \gamma_2^{-1} J_*^{-\frac{1}{2}} \|\hat{M}_N - M_N^*\|_\sigma(A_\theta) \\
&= O\Big\{(\log N)^{\frac{1}{2}} J_*^{-\frac{1}{2}}\Big\} \quad \text{as} \quad N \to \infty,
\end{aligned}$$

where the last step can be implied from the results in Lemma B.7 applied to the fact that with probability tending to 1, $\|\hat{M}_N - M_N^*\|_\sigma(A_\theta) = O\big(f_N\big)$, and the rate requirement of f_N in Lemma B.8, where the minimum order of f_N is determined by

$f_N^2/\log N \to \infty$. It follows that,

$$\|\hat{\theta} - \theta^*\|_\infty = O_p\Big\{(\log N)^{\frac{1}{2}} J_*^{-\frac{1}{2}}\Big\}. \tag{B.5}$$

Combining (B.4) and (B.5), we have $\max_{i,j} |\hat{m}_{ij} - m_{ij}^*| = O_p\big\{(\log J)^{\frac{1}{2}} N_*^{-\frac{1}{2}} + (\log N)^{\frac{1}{2}} J_*^{-\frac{1}{2}}\big\}$. Therefore, the second part of the theorem follows. □

Proof of Theorem III.11. We first seek to show $|\sigma^2(g)/\tilde{\sigma}^2(g) - 1| \to 0$ as $N \to \infty$, where $\sigma^2(g) = \sigma\{g(\hat{M})\}$. Since Conditions III.6, III.7 and III.8 hold and $\|w_g\|_1, \|\tilde{w}_g\|_1 < C$, by Lemma B.10 below,

$$|\sigma^2(g) - \tilde{\sigma}^2(g)| = O(N_*^{-1} J_*^{-1}) \quad \text{as} \quad N \to \infty. \tag{B.6}$$

Hence, it follows

$$\Big|\frac{\sigma^2(g)}{\tilde{\sigma}^2(g)} - 1\Big| = \frac{|\sigma^2(g) - \tilde{\sigma}^2(g)|}{\tilde{\sigma}^2(g)} \to 0 \quad \text{as} \quad N \to \infty,$$

where the last step follows from (B.6) and the definition of $\tilde{\sigma}^2(g)$.

Lemma B.10. *Assume Conditions III.6, III.7 and III.8 hold and* $\sum_{i=1}^N \theta_i = 0$. *Consider a linear function* $g : \Omega_N \mapsto \mathbb{R}$ *with* $g(x) = \sum_{i=1}^N h_i\theta_i + \sum_{j=1}^J h_j'\beta_j$. *If there exists a positive* $C < \infty$ *such that* $\sum_{i=1}^N |h_i| < C$ *and* $\sum_{j=1}^J |h_j'| < C$, *then*

$$\sigma^2(g) = \sum_{i=1}^N h_i^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^J h_j'^2(\sigma_{+j}^2)^{-1} + O(N_*^{-1}J_*^{-1}) \quad \textit{as} \quad N \to \infty.$$

Then if we can show $\sigma(g)^{-1}\{g(\hat{M}) - g(M^*)\} \to \mathrm{N}(0,1)$ in distribution, the first

part of the theorem would follow directly. As a direct application of Proposition III.9, we can re-write function g on Ω_N using $[\cdot,\cdot]_\sigma$ as follows. Let $c_N \in \Omega_N$ be defined by the equation

$$g(x) = [c_N, x]_\sigma = \sum_{i=1}^{N} \sum_{j \in S_J(i)} c_{ij} x_{ij} \sigma_{ij}^2, \qquad x \in \Omega_N.$$

Then we can express,

$$\begin{aligned} g(\hat{M}_N) - g(M_N^*) &= g\big(\hat{M}_N - M_N^*\big) = \big[c_N, \hat{M}_N - M_N^*\big]_\sigma \\ &= \big[c_N, \hat{M}_N - M_N^* - R_N\big]_\sigma + \big[c_N, R_N\big]_\sigma. \end{aligned} \tag{B.7}$$

Recall that $\sigma(g) = \sup_{x \in \Omega_N} \big\{ |[c_N, x]_\sigma| : \|x\|_\sigma \leq 1 \big\}$, the supremum is attained at $x = c_N / \|c_N\|_\sigma$, so $\sigma(g) = \|c_N\|_\sigma$. We consider two possible cases, $w_g = 0$ in case 1 and $w_g \neq 0$ in case 2, and we seek to prove the result of the theorem hold under both cases separately.

We first consider case 1. Similar as in the proof of Theorem III.10, we consider a set A_β consisting of linear functions $g_j \in \Omega_N^*$ on Ω_N such that $g_j(x) = \beta_j$ with $A_\beta = \{g_j : j = 1, ..., J\}$. We now pick a positive sequence f_N satisfying $f_N^2 / \log J \to \infty$ and $f_N^2 N_*^{-1/2} \to 0$ as $N \to \infty$. Then by Lemma B.5, we can pick a sequence $d_N \geq 0$ satisfying Lemma B.5(a) and Lemma B.5(b). Furthermore, it can be implied that Lemmas B.2 and B.3 still hold by taking $A = A_\beta$. Moreover, Lemma B.6 implies there exist $0 < \gamma_1, \gamma_2 < \infty$ and some n such that for all $N > n$,

$$\gamma_1^{-1} N_*^{-\frac{1}{2}} < \sigma(g_j) < \gamma_2^{-1} N_*^{-\frac{1}{2}}, \quad g_j \in A_\beta. \tag{B.8}$$

Now for any $x \in \Omega_N$,

$$\begin{aligned}
|g(x)| &= |\tilde{w}_g^T \beta| \\
&\leq \|\tilde{w}_g\|_1 \max_{j=1,\ldots,J} \{|\beta_j|\} \\
&\leq C \max_{g_j \in A_\beta} \{|g_j(x)|\} \\
&= C \max_{g_j \in A_\beta} \left\{ \frac{|g_j(x)|}{\sigma(g_j)} \sigma(g_j) \right\} \\
&\leq C \left\{ \max_{g_j \in A_\beta} \frac{|g_j(x)|}{\sigma(g_j)} \right\} \max_{g_j \in A_\beta} \sigma(g_j) \\
&= C \|x\|_\sigma(A_\beta) \max_{g_j \in A_\beta} \sigma(g_j) \\
&\leq C \gamma_2^{-1} N_*^{-\frac{1}{2}} \|x\|_\sigma(A_\beta), \qquad \text{(B.9)}
\end{aligned}$$

where the second last step follows from the definition of $\|\cdot\|_\sigma(A_\beta)$ and the last step follows from (B.8). Since case 1 assumes $w_g = 0$, so $g(M) \neq 0$ implies $\tilde{w}_g \neq 0$. Then as a direct consequence of Lemma B.10, there exists some $0 < \gamma_3 < \infty$ such that for all $N > n$,

$$\sigma(g) \geq \gamma_3 N_*^{-\frac{1}{2}}. \qquad \text{(B.10)}$$

As a result of (B.9), we have

$$\left| \left[c_N, \hat{M}_N - M_N^* - R_N \right]_\sigma \right| \leq C \gamma_2^{-1} N_*^{-\frac{1}{2}} \|\hat{M}_N - M_N^* - R_N\|_\sigma(A_\beta). \qquad \text{(B.11)}$$

Note that from (B.7),

$$\frac{g(\hat{M}_N) - g(M_N^*)}{\sigma(g)} = \frac{\left[c_N, \hat{M}_N - M_N^* - R_N\right]_\sigma + \left[c_N, R_N\right]_\sigma}{\sigma(g)}$$

Rearrange gives as $N \to \infty$, with probability tending to 1 that,

$$\begin{aligned}
\left| \frac{g(\hat{M}_N) - g(M_N^*)}{\sigma(g)} - \frac{\left[c_N, R_N\right]_\sigma}{\sigma(g)} \right| &= \frac{\left| [c_N, \hat{M}_N - M_N^* - R_N]_\sigma \right|}{\sigma(g)} \\
&\leq \frac{C\gamma_2^{-1} N_*^{-\frac{1}{2}}}{\sigma(g)} \|\hat{M}_N - M_N^* - R_N\|_\sigma(A_\beta) \\
&\leq C\gamma_2^{-1}\gamma_3^{-1} d_N \left[\|R_N\|_\sigma(A_\beta)\right]^2 \\
&\leq \frac{1}{4} C\gamma_2^{-1}\gamma_3^{-1} d_N f_N^2 \\
&\to 0, \qquad \text{(B.12)}
\end{aligned}$$

where the second line follows from (B.11), the third line can be obtained from (B.10) and Lemma B.3, the second last line can be implied by Lemma B.2 and the last line follows from Lemma B.5. Hence, it turns out that it suffices to show $\left[c_N, R_N\right]_\sigma / \sigma(g) \to \mathrm{N}(0,1)$. Write $Z_N = [c_N, R_N]_\sigma / \sigma(g) = \sum_{i=1}^N \sum_{j \in S_J(i)} \left\{ c_{ij}(Y_{ij} - E_{ij}) \right\} / \|c_N\|_\sigma$ for simplicity. The strategy is to show the moment generating function of Z_N, denoted as $G_{Z_N}(t)$, converges to $\exp\{t^2/2\}$, the moment generating function of the standard Gaussian. Write $c'_{ij} = c_{ij}/\|c_N\|_\sigma = c_{ij}/\sigma(g)$ for simplicity. We consider

the log moment generating function of Z_N,

$$\begin{aligned}
\log G_{Z_N}(t) &= \log \mathbb{E}\big[e^{tZ_N}\big] = \log \mathbb{E}\Big[\exp\Big\{\frac{t}{\sigma(g)}\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}(Y_{ij}-E_{ij})\Big\}\Big] \\
&= -t\sum_{i=1}^{N}\sum_{j\in S_J(i)} c'_{ij}E_{ij} + \log\prod_{i=1}^{N}\prod_{j\in S_J(i)} \mathbb{E}\big\{\exp(tc'_{ij}Y_{ij})\big\} \\
&= -t\sum_{i=1}^{N}\sum_{j\in S_J(i)} c'_{ij}E_{ij} + \sum_{i=1}^{N}\sum_{j\in S_J(i)} \log \mathbb{E}\big\{\exp(tc'_{ij}Y_{ij})\big\} \\
&= \sum_{i=1}^{N}\sum_{j\in S_J(i)} \Big[\log\big\{1+\exp(m^*_{ij})\big\}^{-1} \\
&\quad - \log\big\{1+\exp(tc'_{ij}+m^*_{ij})\big\}^{-1} - tc'_{ij}E_{ij}\Big] \\
&= \sum_{i=1}^{N}\sum_{j\in S_J(i)} \Big[\log\big\{h(m^*_{ij})\big\} - \log\big\{h(tc'_{ij}+m^*_{ij})\big\} - tc'_{ij}E_{ij}\Big], \qquad \text{(B.13)}
\end{aligned}$$

where $h(m_{ij}) = \{1+\exp(m_{ij})\}^{-1}$. We can then apply Taylor expansion to $\log\{h(tc'_{ij}+m^*_{ij})\}$ about m^*_{ij}. For some $t' = \alpha t$ with $0 < \alpha < 1$,

$$\log\{h(tc'_{ij}+m^*_{ij})\} = \log\{h(m^*_{ij})\} - E_{ij}tc'_{ij} - \frac{t^2}{2}c'^2_{ij}\sigma^2(m^*_{ij}+t'c'_{ij}).$$

Substitute into Equation (B.13),

$$\log G_{Z_N}(t) = \frac{t^2}{2}\sum_{i=1}^{N}\sum_{j\in S_J(i)} c'^2_{ij}\sigma^2(m^*_{ij}+t'c'_{ij}), \qquad \|t'c'_N\|_\sigma(A_\beta) \le f_N. \qquad \text{(B.14)}$$

With $\|c'_N\|_\sigma = \|c_N\|_\sigma/\|c_N\|_\sigma = 1$, the summation term in (B.14) can be re-expressed

as follows,

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}'^2\sigma^2(m_{ij}^*+t'c_{ij}') &= \sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}'^2\big\{\sigma^2(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2+\sigma_{ij}^2\big\}\\ &= \|c_N'\|_\sigma^2+\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}'^2\big\{\sigma^2(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\}\\ &= 1+\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}'^2\big\{\sigma^2(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\}.\end{aligned}$$

Note that

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}'^2\big\{\sigma^2(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\} &= \frac{1}{\sigma(g)}\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}\big\{\sigma^2(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\}c_{ij}'\\ &= \frac{1}{\sigma(g)}g\big\{U_N(M_N^*+t'c_N',c_N')\big\}\\ &\le \frac{C\gamma_2^{-1}N_*^{-\frac{1}{2}}}{\sigma(g)}\|U_N(M_N^*+t'c_N',c_N')\|_\sigma(A_\beta)\\ &\le \frac{C\gamma_2^{-1}N_*^{-\frac{1}{2}}}{\sigma(g)}d_N\|t'c_N'\|_\sigma(A_\beta)\|c_N'\|_\sigma(A_\beta)\\ &\le \frac{C\gamma_2^{-1}N_*^{-\frac{1}{2}}}{\sigma(g)}d_Nf_N\\ &\le C\gamma_2^{-1}\gamma_3^{-1}d_Nf_N\\ &\to 0 \quad \text{as} \quad N\to\infty.\end{aligned}$$

The second line follows from $U_{ij}(M_N^*+t'c_N',c_N') = (\sigma_{ij}^2)^{-1}\big\{\sigma(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\}c_{ij}'$. The third last step follows from $\|c_N'\|_\sigma(A_\beta)\le\|c_N'\|_\sigma=1$ and the last step can be implied from Lemma B.5(b). Therefore, $\log G_{Z_N}(t)\to t^2/2$ as $N\to\infty$.

Now consider case 2. We adopt a similar strategy to derive asymptotic normality as in case 1. Define $A_{\theta,\beta}$ to be a set consisting of linear functions $g_i, g_j' \in \Omega_N^*$ on Ω_N such that $g_i(x) = \theta_i$ and $g_j'(x) = \beta_j$, with $A_{\theta,\beta} = \{g_i, g_j' : i = 1, ..., N, j = 1, ..., J\}$. The explicit forms of g_i and g_j' can be found in the proof of Theorem III.10.

From now onwards, we take sequences f_N and d_N as satisfying the conditions in Lemma B.11 below. Note it can be implied that with such f_N and d_N, Lemmas B.2 and B.3 still hold by taking $A = A_{\theta,\beta}$. From Lemmas B.6 and B.9, we know that for any $f \in A_{\theta,\beta}$, there exist $0 < c_1, c_2 < \infty$ and some n such that for all $N > n$,

$$c_1^{-1} N_*^{-\frac{1}{2}} < \sigma(f) < c_2^{-1} J_*^{-\frac{1}{2}}. \tag{B.15}$$

Lemma B.11. *Assume Conditions III.6, III.7 and III.8 hold and $J_*^{-2} N_* (\log N)^2 \to 0$ as $N \to \infty$. If $A_{\theta,\beta} = \{g_i, g_j' : i = 1, ..., N, j = 1, ..., J\}$ such that $g_i, g_j' \in \Omega_N^*$, and $g_i(x) = \theta_i$ and $g_j'(x) = \beta_j$ for $x \in \Omega_N$. Let $C_N = |A_{\theta,\beta}|$, the cardinality of $A_{\theta,\beta}$. Then there exist sequences $f_N > 0$ and $d_N \geq 0$ satisfying the followings.*

(a). As $N \to \infty$, $f_N^2 / \log C_N \to \infty$.

(b). If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A_{\theta,\beta}) \leq f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y, v)\|_\sigma(A_{\theta,\beta}) \leq d_N \|y - M_N^*\|_\sigma(A_{\theta,\beta}) \|v\|_\sigma(A_{\theta,\beta})$. Furthermore, $d_N f_N^2 \to 0$ as $N \to \infty$.

Now for any $x \in \Omega_N$,

$$\begin{aligned}
|g(x)| &= |w_g^T\theta + \tilde{w}_g^T\beta| \\
&\leq \Big(\|w_g\|_1 + \|\tilde{w}_g\|_1\Big) \max_{i=1,\ldots,N, j=1,\ldots,J}\{|\theta_i|, |\beta_j|\} \\
&= \Big(\|w_g\|_1 + \|\tilde{w}_g\|_1\Big) \max_{f\in A_{\theta,\beta}}\{|f(x)|\} \\
&= \Big(\|w_g\|_1 + \|\tilde{w}_g\|_1\Big) \max_{f\in A_{\theta,\beta}}\left\{\frac{|f(x)|}{\sigma(f)}\sigma(f)\right\} \\
&\leq \Big(\|w_g\|_1 + \|\tilde{w}_g\|_1\Big)\left\{\max_{f\in A_{\theta,\beta}}\frac{|f(x)|}{\sigma(f)}\right\}\left\{\max_{f\in A_{\theta,\beta}}\sigma(f)\right\} \\
&< 2Cc_2^{-1}J_*^{-\frac{1}{2}}\|x\|_\sigma(A_{\theta,\beta}), \qquad \text{(B.16)}
\end{aligned}$$

where the last step follows from the definition of $\|\cdot\|_\sigma(A_{\theta,\beta})$, (B.15) and the assumption that $\|w_g\|_1, \|\tilde{w}_g\|_1 < C$. Further note that since $w_g \neq 0$, as a direct consequence of Lemma B.10, there exists some $0 < c_3 < \infty$ such that for all $N > n$,

$$\sigma(g) \geq c_3 J_*^{-\frac{1}{2}}. \qquad \text{(B.17)}$$

As a result of (B.16),

$$\left|\left[c_N, \hat{M}_N - M_N^* - R_N\right]_\sigma\right| \leq 2Cc_2^{-1}J_*^{-\frac{1}{2}}\|\hat{M}_N - M_N^* - R_N\|_\sigma(A_{\theta,\beta}).$$

Again, we have

$$\frac{g(\hat{M}_N) - g(M_N^*)}{\sigma(g)} = \frac{\left[c_N, \hat{M}_N - M_N^* - R_N\right]_\sigma + \left[c_N, R_N\right]_\sigma}{\sigma(g)}$$

As $N \to \infty$, re-arrange gives with probability tending to 1 that,

$$\begin{aligned}
\Big| \frac{g(\hat{M}_N) - g(M_N^*)}{\sigma(g)} - \frac{\big[c_N, R_N\big]_\sigma}{\sigma(g)} \Big| &= \frac{\Big| [c_N, \hat{M}_N - M_N^* - R_N]_\sigma \Big|}{\sigma(g)} \\
&\leq \frac{2Cc_2^{-1} J_*^{-\frac{1}{2}}}{\sigma(g)} \|\hat{M}_N - M_N^* - R_N\|_\sigma(A_{\theta,\beta}) \\
&\leq \frac{2Cc_2^{-1} J_*^{-\frac{1}{2}}}{\sigma(g)} d_N \big[\|R_N\|_\sigma(A_{\theta,\beta})\big]^2 \\
&\leq \frac{1}{2} Cc_2^{-1} c_3^{-1} d_N f_N^2 \\
&\to 0, \qquad \text{(B.18)}
\end{aligned}$$

Again, we can denote $Z_N = \big[c_N, R_N\big]_\sigma / \sigma(g)$ for notation simplicity. Similar as in case 1, we just need to show $Z_N \to \mathrm{N}(0,1)$. We consider the log moment generating function of Z_N, denoted as $\log G_{Z_N}(t)$. Write $c'_{ij} := c_{ij}/\sigma(g)$. Then similarly as in the proof for case 1, we obtain

$$\log G_{Z_N}(t) = \frac{t^2}{2} \sum_{i=1}^{N} \sum_{j \in S_J(i)} c'^2_{ij} \sigma^2(m^*_{ij} + t' c'_{ij}), \qquad \|t' c'_N\|_\sigma(A_{\theta,\beta}) \leq f_N,$$

where,

$$\sum_{i=1}^{N} \sum_{j \in S_J(i)} c'^2_{ij} \sigma^2(m^*_{ij} + t' c'_{ij}) = 1 + \sum_{i=1}^{N} \sum_{j \in S_J(i)} c'^2_{ij} \big\{ \sigma^2(m^*_{ij} + t' c'_{ij}) - \sigma^2_{ij} \big\}.$$

Note that

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}'^2\big\{\sigma^2(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\} &= \frac{1}{\sigma(g)}\sum_{i=1}^{N}\sum_{j\in S_J(i)} c_{ij}\big\{\sigma^2(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\}c_{ij}' \\
&= \frac{1}{\sigma(g)} g\big\{U_N(M_N^*+t'c_N', c_N')\big\} \\
&\leq \frac{2Cc_2^{-1}J_*^{-\frac{1}{2}}}{\sigma(g)}\|U_N(M_N^*+t'c_N', c_N')\|_\sigma(A_{\theta,\beta}) \\
&\leq \frac{2Cc_2^{-1}J_*^{-\frac{1}{2}}}{\sigma(g)} d_N\|t'c_N'\|_\sigma(A_{\theta,\beta})\|c_N'\|_\sigma(A_{\theta,\beta}) \\
&\leq \frac{2Cc_2^{-1}J_*^{-\frac{1}{2}}}{\sigma(g)} d_N f_N \\
&\leq 2Cc_2^{-1}c_3^{-1}d_N f_N \\
&\to 0 \quad \text{as} \quad N\to\infty.
\end{aligned}$$

The second line follows from $U_{ij}(M_N^*+t'c_N', c_N') = (\sigma_{ij}^2)^{-1}\big\{\sigma(m_{ij}^*+t'c_{ij}')-\sigma_{ij}^2\big\}c_{ij}'$. The third last step follows from $\|c_N'\|_\sigma(A_{\theta,\beta}) \leq \|c_N'\|_\sigma = 1$ and the last step can be implied from Lemma B.11(b). Therefore, $\log G_{Z_N}(t) \to \frac{t^2}{2}$ as $N\to\infty$. Hence, the first part of the theorem follows.

Now we seek to prove the second part of the theorem. The strategy is to show

$|\hat{\sigma}^2(g) - \tilde{\sigma}^2(g)|/\tilde{\sigma}^2(g) \to 0$ in probability as $N \to \infty$. Consider

$$\begin{aligned}
\frac{|\hat{\sigma}^2(g) - \tilde{\sigma}^2(g)|}{\tilde{\sigma}^2(g)} &= \frac{\Big| \sum_{i=1}^N w_{gi}^2 \big\{ (\hat{\sigma}_{i+}^2)^{-1} - (\sigma_{i+}^2)^{-1} \big\} + \sum_{j=1}^J \tilde{w}_{gj}^2 \big\{ (\hat{\sigma}_{+j}^2)^{-1} - (\sigma_{+j}^2)^{-1} \big\} \Big|}{\sum_{i=1}^N w_{gi}^2 (\sigma_{i+}^2)^{-1} + \sum_{j=1}^J \tilde{w}_{gj}^2 (\sigma_{+j}^2)^{-1}} \\
&= \frac{\Big| \sum_{i=1}^N w_{gi}^2 \Big\{ \frac{\sigma_{i+}^2 - \hat{\sigma}_{i+}^2}{(\hat{\sigma}_{i+}^2)(\sigma_{i+}^2)} \Big\} + \sum_{j=1}^J \tilde{w}_{gj}^2 \Big\{ \frac{\sigma_{+j}^2 - \hat{\sigma}_{+j}^2}{(\hat{\sigma}_{+j}^2)(\sigma_{+j}^2)} \Big\} \Big|}{\sum_{i=1}^N w_{gi}^2 (\sigma_{i+}^2)^{-1} + \sum_{j=1}^J \tilde{w}_{gj}^2 (\sigma_{+j}^2)^{-1}} \\
&\le \frac{\Big| \sum_{i=1}^N w_{gi}^2 \Big\{ \frac{\sum_{j \in S_J(i)} |\sigma_{ij}^2 - \hat{\sigma}_{ij}^2|}{(\hat{\sigma}_{i+}^2)(\sigma_{i+}^2)} \Big\} + \sum_{j=1}^J \tilde{w}_{gj}^2 \Big\{ \frac{\sum_{i \in S_N(j)} |\sigma_{ij}^2 - \hat{\sigma}_{ij}^2|}{(\hat{\sigma}_{+j}^2)(\sigma_{+j}^2)} \Big\} \Big|}{\sum_{i=1}^N w_{gi}^2 (\sigma_{i+}^2)^{-1} + \sum_{j=1}^J \tilde{w}_{gj}^2 (\sigma_{+j}^2)^{-1}}.
\end{aligned} \tag{B.19}$$

Since $m_{ij}^*, \hat{m}_{ij} \in \mathbb{R}$, $0 < \sigma_{ij}^2, \hat{\sigma}_{ij}^2 < 1$. Note that there exist $0 < c_4, c_5 < \infty$ that

$$\sigma_{i+}^2, \hat{\sigma}_{i+}^2 > c_4 J_*, \qquad \sigma_{+j}^2, \hat{\sigma}_{+j}^2 > c_5 N_*.$$

Further note that there exists a positive $c_6 < \infty$ such that

$$\begin{aligned}
\max_{i,j,z_{ij}=1} |\sigma_{ij}^2 - \hat{\sigma}_{ij}^2| &\le c_6 \max_{i,j,z_{ij}=1} |m_{ij}^* - \hat{m}_{ij}| \\
&= o_p(1), \qquad \text{as } N \to \infty.
\end{aligned}$$

where the last line follows from (B.3). It follows

$$\frac{\sum_{j \in S_J(i)} |\sigma_{ij}^2 - \hat{\sigma}_{ij}^2|}{(\hat{\sigma}_{i+}^2)(\sigma_{i+}^2)} = o_p\big(J_*^{-1}\big), \tag{B.20}$$

$$\frac{\sum_{i \in S_N(j)} |\sigma_{ij}^2 - \hat{\sigma}_{ij}^2|}{(\hat{\sigma}_{+j}^2)(\sigma_{+j}^2)} = o_p\big(N_*^{-1}\big). \tag{B.21}$$

Moreover, we note that $\|w_g\|_1, \|\tilde{w}_g\|_1 < C$ implies that $\sum_{i=1}^{N} w_{gi}^2 < c_7$ and $\sum_{j=1}^{J} \tilde{w}_{gj}^2 < c_7$ for some $c_7 < \infty$. From (B.19), it can be implied that

$$\frac{|\hat{\sigma}^2(g) - \tilde{\sigma}^2(g)|}{\tilde{\sigma}^2(g)} = o_p(1), \qquad \text{as } N \to \infty,$$

where the above result follows from (B.20), (B.21) and the assumption that $g(x) \neq 0$ for any $x \in \Omega_N$. Since we have shown $\tilde{\sigma}(g)^{-1}\{g(\hat{M}) - g(M^*)\} \to \mathrm{N}(0,1)$ in distribution in the first part of the proof, it follows that $\hat{\sigma}(g)^{-1}\{g(\hat{M}) - g(M^*)\} \to \mathrm{N}(0,1)$ in distribution as $N \to \infty$. □

Proof of Proposition III.9. We prove the first part of the proposition by direct construction; in particular, we find the solutions for θ and β, respectively, given equations $\sum_{i=1}^{N} \theta_i = 0$ and $\theta_i - \beta_j = m_{ij}$, $i = 1, ..., N, j = 1, ..., J$, for which $z_{ij} = 1$. We first construct the solution for β_j, $j = 1, ..., J$. The idea is to include all the row parameters θ_i so that we can apply the constraint $\sum_{i=1}^{N} \theta_i = 0$. Denote $S_J(i) = \{j = 1, ...J : z_{ij} = 1\}$, $S_N(j) = \{i = 1, ..., N : z_{ij} = 1\}$, and $S_{N_\phi}(j) = \{1, 2, ..., N\} \setminus S_N(j)$. Then for any $i \in S_N(j)$, we use $m_{ij} = \theta_i - \beta_j$ in the construction. While for each $i \in S_{N_\phi}(j)$, applying Condition III.8, there must exist $1 \leq i_{i1}, i_{i2}, ..., i_{ik} \leq N$ and $1 \leq j_{i1}, j_{i2}, ..., j_{ik} \leq J$ such that

$$z_{i,j_{i1}} = z_{i_{i1},j_{i1}} = z_{i_{i1},j_{i2}} = z_{i_{i2},j_{i2}} = ... = z_{i_{ik},j_{ik}} = z_{i_{ik},j} = 1,$$

with

$$\begin{aligned}
&m_{i,j_{i1}} - m_{i_{i1},j_{i1}} + m_{i_{i1},j_{i2}} - m_{i_{i2},j_{i2}} + ... - m_{i_{ik},j_{ik}} + m_{i_{ik},j} \\
=&(\theta_i - \beta_{j_{i1}}) - (\theta_{i_{i1}} - \beta_{j_{i1}}) + (\theta_{i_{i1}} - \beta_{j_{i2}}) - (\theta_{i_{i2}} - \beta_{j_{i2}}) + \\
&... - (\theta_{i_{ik}} - \beta_{j_{ik}}) + (\theta_{i_{ik}} - \beta_j) \\
=&\theta_i - \beta_j.
\end{aligned}$$

Therefore, the solution for β_j is simply

$$\begin{aligned}
\beta_j = &- \frac{1}{N}\Big\{ \sum_{i \in S_N(j)} m_{ij} \\
&+ \sum_{i \in S_{N_\phi}(j)} \Big(m_{i,j_{i1}} - m_{i_{i1},j_{i1}} + m_{i_{i1},j_{i2}} - m_{i_{i2},j_{i2}} + ... - m_{i_{ik},j_{ik}} + m_{i_{ik},j} \Big) \Big\}.
\end{aligned}$$

To find solution for θ_i,

$$\begin{aligned}
\theta_i = &\frac{1}{|S_J(i)|} \sum_{j \in S_J(i)} \Big[m_{ij} - \frac{1}{N}\Big\{ \sum_{i' \in S_N(j)} m_{i'j} \\
&+ \sum_{i' \in S_{N_\phi}(j)} \Big(m_{i',j_{i'1}} - m_{i'_{i'1},j_{i'1}} + m_{i'_{i'1},j_{i'2}} - m_{i'_{i'2},j_{i'2}} + ... - m_{i'_{i'k},j_{i'k}} + m_{i'_{i'k},j} \Big) \Big\} \Big],
\end{aligned}$$

where $|S_J(i)|$ denotes the cardinality of $S_j(i)$. This concludes the proof for the first part of the proposition.

We can view the row parameters and column parameters as a bipartite graph $\mathcal{G}$, with one part consisting of row parameters as nodes (denoted as $\{i = 1, ..., N\}$ for simplicity) and the other consisting of column parameters as nodes (denoted as

$\{j = 1, ..., J\}$ for simplicity). If $z_{ij} = 1$, then there is an edge connecting i and j in $\mathcal{G}$. For the second part of the proposition, note if Condition III.8 is not satisfied, then there exists at least one pair of (i, j) such that there does not exist a path connecting them in graph $\mathcal{G}$. This means that claim: $\mathcal{G}$ can be separated into at least two sub-graphs. Denote the two sub-graphs by $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively. The above claim can be proved by a contradiction argument as follows. Suppose not, then there exist either $i'_1 \in \mathcal{G}_1$ and $j'_2 \in \mathcal{G}_2$ with $z_{i'_1 j'_2} = 1$, or $j'_1 \in \mathcal{G}_1$ and $i'_2 \in \mathcal{G}_2$ with $z_{i'_2 j'_1} = 1$. By assumption there must exist a path connecting any two nodes within each of the two sub-graphs, otherwise we could split $\mathcal{G}$ into two sub-graphs. Therefore, there must exist a path connecting the pair (i, j). A contradiction.

Now, denote $\{\theta_{i_1}, \beta_{j_1} : 1 \leq i_1 \leq N, 1 \leq j_1 \leq J\}$ and $\{\theta_{i_2}, \beta_{j_2} : 1 \leq i_2 \leq N, 1 \leq j_2 \leq J\}$ as the values associated with the nodes in $\mathcal{G}_1$ and in $\mathcal{G}_2$ respectively and together also serving as a solution set satisfying $\sum_{i=1}^{N} \theta_i = 0$ and $\theta_i - \beta_j = m_{ij},\ i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$. Let n_{i_1} and n_{i_2} denote the number of row parameters in $\mathcal{G}_1$ and in $\mathcal{G}_2$ respectively. Let $\tau = n_{i_1}/n_{i_2}$. For any constant a, let $\tilde{\theta}_{i_1} = \theta_{i_1} + a, \tilde{\beta}_{j_1} = \beta_{j_1} + a$ and $\tilde{\theta}_{i_2} = \theta_{i_2} - \tau a, \tilde{\beta}_{j_2} = \beta_{j_2} - \tau a$. We can check easily that $(\tilde{\theta}, \tilde{\beta})$ is also a solution to the system but $(\tilde{\theta}, \tilde{\beta}) \neq (\theta, \beta)$. This concludes the proof for the second part of the proposition. □

B.2 Proofs of Supporting Lemmas

We first give some intuition on how to obtain the approximation formula for $\sigma^2(g)$, as summarized in Lemmas B.12, B.13 and B.14 below. Lemmas B.12, B.13 and B.14 will be used in the proofs of other supporting lemmas, which will be given

later in this section.

First note that it is a property of the exponential family that $\sigma(g) = \sup_{x\in\Omega_N}\{|g(x)| : \|x\|_\sigma^2 \leq 1\}$. $\sigma^2(g)$ can be viewed as the solution to a constrained quadratic programming problem, i.e.

$$\max_{\theta,\beta}\Big\{\sum_{i=1}^{N}\sum_{j\in S_J(i)} w_{ij}(\theta_i-\beta_j)\Big\}^2 \quad \text{such that} \quad \sum_{i=1}^{N}\sum_{j\in S_J(i)}\sigma_{ij}^2(\theta_i-\beta_j)^2 \leq 1, \sum_{i=1}^{N}\theta_i = 0. \tag{B.22}$$

An explicit form is often difficult to derive, so an approximation is desired for both implementation and inference purposes. We consider a three-way decomposition of the coefficients of g that lies in the constrained solution space, and convert this quadratic programming to a linear system from which $\sigma^2(g)$ can be solved. The results are summarized in Lemma B.12 below.

Lemma B.12. *Define a vector* $d(g) = \big\{d_{ij}(g) : i = 1,...,N, j = 1,...,J, z_{ij} = 1, d_{ij}(g) \in \mathbb{R}\big\}$ *with a three-way decomposition* $d_{ij}(g) = b(g) + f_i(g) + m_j(g)$, *such that* $[d(g), x]_\sigma = g(x)$ *for* $x \in \Omega_N$ *and* $f_i(g), m_j(g)$ *satisfying*

$$\sum_{i=1}^{N}\sigma_{i+}^2 f_i(g) = 0, \tag{B.23}$$

$$\sum_{j=1}^{J}\sigma_{+j}^2 m_j(g) = 0. \tag{B.24}$$

Then, we have

$$\sigma^2(g) = b^2(g)\sigma^2_{++} + \sum_{i=1}^{N} \sigma^2_{i+} f_i^2(g) + \sum_{j=1}^{J} \sigma^2_{+j} m_j^2(g) + 2\sum_{i=1}^{N} \sum_{j \in S_J(i)} \sigma^2_{ij} f_i(g) m_j(g). \tag{B.25}$$

Proof. Note $\sigma^2(g)$ is a solution to the quadratically constrained quadratic programming problem (B.22). From *Haberman* (1977), the construction of $d(g)$ in the lemma lies in the required solution space of (B.22). As a result, $\sigma^2(g)$ can be expressed directly as $\sigma^2(g) = \|d(g)\|^2_\sigma$. We just need to find an explicit expression of $\|d(g)\|^2_\sigma$ in terms of $b(g), f_i(g), m_j(g)$.

First consider $x \in \Omega_N$ such that $x_{ij} = y$ are identical for all $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$. Then in such cases,

$$\begin{aligned}
g(x) &= [d(g), x]_\sigma \\
&= \sum_{i=1}^{N} \sum_{j \in S_J(i)} \big\{ b(g) + f_i(g) + m_j(g) \big\} \sigma^2_{ij} y \\
&= b(g)\sigma^2_{++} y + \sum_{i=1}^{N} \Big(\sum_{j \in S_J(i)} \sigma^2_{ij} \Big) f_i(g) y + \sum_{j=1}^{J} \Big(\sum_{i \in S_N(j)} \sigma^2_{ij} \Big) m_j(g) y \\
&= b(g)\sigma^2_{++} y + \sum_{i=1}^{N} \sigma^2_{i+} f_i(g) y + \sum_{j=1}^{J} \sigma^2_{+j} m_j(g) y \\
&= b(g)\sigma^2_{++} y,
\end{aligned} \tag{B.26}$$

where the last step follows from (B.23) and (B.24). Also by the original definition of

g, we have

$$g(x) = \sum_{i=1}^{N} \sum_{j \in S_J(i)} w_{ij} y = w_{++} y. \tag{B.27}$$

Since (B.26) and (B.27) hold for any y, we must have

$$b(g) = (\sigma^2_{++})^{-1} w_{++}. \tag{B.28}$$

Next consider $x \in \Omega_N$ such that $x_{ij} = y_i$, $y_i \in \mathbb{R}$, for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$, then

$$g(x) = [d(g), x]_\sigma = \sum_{i=1}^{N} \sum_{j \in S_J(i)} d_{ij}(g) \sigma^2_{ij} y_i = \sum_{i=1}^{N} y_i \Big(\sum_{j \in S_J(i)} d_{ij}(g) \sigma^2_{ij} \Big). \tag{B.29}$$

From the original definition of g,

$$g(x) = \sum_{i=1}^{N} \sum_{j \in S_J(i)} w_{ij} y_i = \sum_{i=1}^{N} y_i \Big(\sum_{j \in S_J(i)} w_{ij} \Big). \tag{B.30}$$

Since (B.29) = (B.30) for any y_i, it follows that

$$\sum_{j \in S_J(i)} d_{ij}(g) \sigma^2_{ij} = \sum_{j \in S_J(i)} w_{ij} = w_{i+}, \quad i = 1, ..., N. \tag{B.31}$$

Consider

$$
\begin{aligned}
f_i(g) + m_j(g) &= d_{ij}(g) - b(g) \\
\sum_{j \in S_J(i)} \big\{ f_i(g) + m_j(g) \big\} \sigma_{ij}^2 &= \sum_{j \in S_J(i)} \big\{ d_{ij}(g) - b(g) \big\} \sigma_{ij}^2 \\
\sigma_{i+}^2 f_i(g) + \sum_{j \in S_J(i)} \sigma_{ij}^2 m_j(g) &= \sum_{j \in S_J(i)} d_{ij}(g) \sigma_{ij}^2 - \sigma_{i+}^2 b(g) \\
\sigma_{i+}^2 f_i(g) + \sum_{j \in S_J(i)} \sigma_{ij}^2 m_j(g) &= w_{i+} - (\sigma_{++}^2)^{-1} w_{++} \sigma_{i+}^2, \quad i = 1, ..., N, \qquad \text{(B.32)}
\end{aligned}
$$

where the last line follows from (B.28) and (B.31). Similarly, we consider $x \in \Omega_N$ such that $x_{ij} = y_j$, $y_j \in \mathbb{R}$ for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$, then

$$
\begin{aligned}
g(x) = [d(g), x]_\sigma &= \sum_{j=1}^{J} \sum_{i \in S_N(j)} d_{ij}(g) \sigma_{ij}^2 y_j \\
&= \sum_{j=1}^{J} y_j \Big(\sum_{i \in S_N(j)} d_{ij}(g) \sigma_{ij}^2 \Big). \qquad \text{(B.33)}
\end{aligned}
$$

Again by the original definition of g,

$$
g(x) = \sum_{j=1}^{J} \sum_{i \in S_N(j)} w_{ij} y_j = \sum_{j=1}^{J} y_j \Big(\sum_{i \in S_N(j)} w_{ij} \Big). \qquad \text{(B.34)}
$$

Since (B.33) = (B.34) for any $y_j \in \mathbb{R}$, it follows

$$
\sum_{i \in S_N(j)} d_{ij}(g) \sigma_{ij}^2 = \sum_{i \in S_N(j)} w_{ij} = w_{+j}. \qquad \text{(B.35)}
$$

Similarly,

$$
\begin{aligned}
f_i(g) + m_j(g) &= d_{ij}(g) - b(g) \\
\sum_{i \in S_N(j)} \big\{ f_i(g) + m_j(g) \big\} \sigma_{ij}^2 &= \sum_{i \in S_N(j)} \big\{ d_{ij}(g) - b(g) \big\} \sigma_{ij}^2 \\
\sigma_{+j}^2 m_j(g) + \sum_{i \in S_N(j)} \sigma_{ij}^2 f_i(g) &= \sum_{i \in S_N(j)} d_{ij}(g) \sigma_{ij}^2 - \sigma_{+j}^2 b(g) \\
\sigma_{+j}^2 m_j(g) + \sum_{i \in S_N(j)} \sigma_{ij}^2 f_i(g) &= w_{+j} - (\sigma_{++}^2)^{-1} w_{++} \sigma_{+j}^2, \quad j = 1, ..., J, \qquad \text{(B.36)}
\end{aligned}
$$

where the last line follows from (B.28) and (B.35). Note that all $b(g), f_i(g), m_j(g)$ can be obtained by solving a system of $N+J+1$ linear equations from (B.28), (B.32) and (B.36). Now we seek to derive a simplified expression for $\|d(g)\|_\sigma^2$ in terms of $b(g), f_i(g), m_j(g)$.

$$
\begin{aligned}
\sigma^2(g) =& \|d(g)\|_\sigma^2 \\
=& \sum_{i=1}^{N} \sum_{j \in S_J(i)} \sigma_{ij}^2 \big\{ b(g) + f_i(g) + m_j(g) \big\}^2 \\
=& b(g) \sum_{i=1}^{N} \sum_{j \in S_J(i)} \sigma_{ij}^2 \big\{ b(g) + f_i(g) + m_j(g) \big\} \qquad \text{(B.37)} \\
&+ \sum_{i=1}^{N} \sum_{j \in S_J(i)} f_i(g) \sigma_{ij}^2 \big\{ b(g) + f_i(g) + m_j(g) \big\} \qquad \text{(B.38)} \\
&+ \sum_{i=1}^{N} \sum_{j \in S_J(i)} m_j(g) \sigma_{ij}^2 \big\{ b(g) + f_i(g) + m_j(g) \big\}. \qquad \text{(B.39)}
\end{aligned}
$$

Let us consider each of these three terms separately,

$$\begin{aligned}(\text{B.37}) &= b^2(g)\sigma^2_{++} + b(g)\sum_{i=1}^{N} f_i(g)\Big(\sum_{j\in S_J(i)}\sigma^2_{ij}\Big) + b(g)\sum_{j=1}^{J} m_j(g)\Big(\sum_{i\in S_N(j)}\sigma^2_{ij}\Big)\\ &= b^2(g)\sigma^2_{++} + b(g)\sum_{i=1}^{N}\sigma^2_{i+}f_i(g) + b(g)\sum_{j=1}^{J}\sigma^2_{+j}m_j(g)\\ &= b^2(g)\sigma^2_{++}.\end{aligned}$$

$$\begin{aligned}(\text{B.38}) &= b(g)\sum_{i=1}^{N} f_i(g)\Big(\sum_{j\in S_J(i)}\sigma^2_{ij}\Big) + \sum_{i=1}^{N} f_i^2(g)\Big(\sum_{j\in S_J(i)}\sigma^2_{ij}\Big) + \sum_{i=1}^{N}\sum_{j\in S_J(i)}\sigma^2_{ij}f_i(g)m_j(g)\\ &= b(g)\sum_{i=1}^{N} f_i(g)\sigma^2_{i+} + \sum_{i=1}^{N}\sigma^2_{i+}f_i^2(g) + \sum_{i=1}^{N}\sum_{j\in S_J(i)}\sigma^2_{ij}f_i(g)m_j(g)\\ &= \sum_{i=1}^{N}\sigma^2_{i+}f_i^2(g) + \sum_{i=1}^{N}\sum_{j\in S_J(i)}\sigma^2_{ij}f_i(g)m_j(g).\end{aligned}$$

$$\begin{aligned}(\text{B.39}) =& b(g)\sum_{j=1}^{J} m_j(g)\Big(\sum_{i\in S_N(j)}\sigma^2_{ij}\Big) + \sum_{j=1}^{J} m_j^2(g)\Big(\sum_{i\in S_N(j)}\sigma^2_{ij}\Big)\\ &+ \sum_{i=1}^{N}\sum_{j\in S_J(i)}\sigma^2_{ij}f_i(g)m_j(g)\\ =& b(g)\sum_{j=1}^{J}\sigma^2_{+j}m_j(g) + \sum_{j=1}^{J}\sigma^2_{+j}m_j^2(g) + \sum_{i=1}^{N}\sum_{j\in S_J(i)}\sigma^2_{ij}f_i(g)m_j(g)\\ =& \sum_{j=1}^{J}\sigma^2_{+j}m_j^2(g) + \sum_{i=1}^{N}\sum_{j\in S_J(i)}\sigma^2_{ij}f_i(g)m_j(g).\end{aligned}$$

Combining three terms together, the result of the lemma follows with

$$\sigma^2(g) = \|d(g)\|_\sigma^2 = b^2(g)\sigma_{++}^2 + \sum_{i=1}^{N} \sigma_{i+}^2 f_i^2(g) + \sum_{j=1}^{J} \sigma_{+j}^2 m_j^2(g) + 2\sum_{i=1}^{N} \sum_{j \in S_J(i)} \sigma_{ij}^2 f_i(g) m_j(g).$$

□

As in the proof of Lemma B.12, we can solve a system of $N+J+1$ linear equations from (B.28), (B.32) and (B.36) for $f_i(g), i = 1, ..., N$, $m_j(g), j = 1, ..., J$ and $b(g)$. Then an exact expression for $\sigma^2(g)$ can be obtained by substituting these values into (B.25). However, when N and J are large, it is difficult to solve this large system of linear equations. Furthermore, to study the order of $\sigma^2(g)$, we need an analytical form for analysis. The following set-ups are used to find an approximation for $\sigma^2(g)$. Define $\gamma_N > 0$ to be the largest number such that for all $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$x^2\sigma_{ij}^2 \geq \gamma_N \Big(\frac{1}{|S_J(i)|} x^2 \sigma_{i+}^2 + \frac{1}{|S_N(j)|} x^2 \sigma_{+j}^2 \Big), \quad x \in \mathbb{R}, \tag{B.40}$$

where $|S_J(i)|$ and $|S_N(j)|$ are the cardinalities of $S_J(i)$ and $S_N(j)$ respectively. Note that there exist some $\gamma > 0$ such that $\gamma_N > \gamma$ for all N. For $i = 1, ..., N$ and

$j = 1, ..., J$, further define

$$f_i'(g) = (\sigma_{i+}^2)^{-1} w_{i+} - (\sigma_{++}^2)^{-1} w_{++}, \tag{B.41}$$

$$m_j'(g) = (\sigma_{+j}^2)^{-1} w_{+j} - (\sigma_{++}^2)^{-1} w_{++}, \tag{B.42}$$

$$f_i''(g) = f_i(g) - f_i'(g), \tag{B.43}$$

$$m_j''(g) = m_j(g) - m_j'(g), . \tag{B.44}$$

Then for $i = 1, ..., N, j = 1, ..., J$ with $z_{ij} = 1$, define

$$\breve{\sigma}_{ij}^2 = \sigma_{ij}^2 - \gamma_N \Big(\frac{1}{|S_J(i)|} \sigma_{i+}^2 + \frac{1}{|S_N(j)|} \sigma_{+j}^2 \Big), \tag{B.45}$$

$$d_{ij}'(g) = b(g) + f_i'(g) + m_j'(g), \tag{B.46}$$

$$d_{ij}''(g) = f_i''(g) + m_j''(g) = d_{ij}(g) - d_{ij}'(g). \tag{B.47}$$

By triangle inequality, (B.47) then implies

$$\|d'(g)\|_\sigma - \|d''(g)\|_\sigma \leq \|d(g)\|_\sigma \leq \|d'(g)\|_\sigma + \|d''(g)\|_\sigma.$$

We seek to use $\|d'(g)\|_\sigma$ as an approximation to $\sigma(g) = \|d(g)\|_\sigma$ while showing $\|d''(g)\|_\sigma$ is a negligible term asymptotically under certain conditions. The analytical expression for $\|d'(g)\|_\sigma$ is given in Lemma B.13 below.

Lemma B.13. *If $d'(g)$ is defined as in* (B.46)*, then*

$$\|d'(g)\|_\sigma^2 = \sum_{i=1}^{N} w_{i+}^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^{J} w_{+j}^2(\sigma_{+j}^2)^{-1} + 2\sum_{i=1}^{N}\sum_{j\in S_J(i)} w_{i+}w_{+j}\sigma_{ij}^2(\sigma_{i+}^2)^{-1}(\sigma_{+j}^2)^{-1} - 3w_{++}^2(\sigma_{++}^2)^{-1}.$$

Proof. Following from the definition of $d'(g)$, we can write

$$\begin{aligned}
&\|d'(g)\|_\sigma^2 \\
&= \sum_{i=1}^{N}\sum_{j\in S_J(i)} \sigma_{ij}^2\big\{b(g) + (\sigma_{i+}^2)^{-1}w_{i+} + (\sigma_{+j}^2)^{-1}w_{+j} - 2(\sigma_{++}^2)^{-1}w_{++}\big\}^2 \\
&= b(g)\sum_{i=1}^{N}\sum_{j\in S_J(i)} \sigma_{ij}^2\big\{b(g) + (\sigma_{i+}^2)^{-1}w_{i+} + (\sigma_{+j}^2)^{-1}w_{+j} - 2(\sigma_{++}^2)^{-1}w_{++}\big\} && \text{(B.48)} \\
&\quad + \sum_{i=1}^{N}\sum_{j\in S_J(i)} \sigma_{ij}^2(\sigma_{i+}^2)^{-1}w_{i+}\big\{b(g) + (\sigma_{i+}^2)^{-1}w_{i+} + (\sigma_{+j}^2)^{-1}w_{+j} - 2(\sigma_{++}^2)^{-1}w_{++}\big\} && \text{(B.49)} \\
&\quad + \sum_{i=1}^{N}\sum_{j\in S_J(i)} \sigma_{ij}^2(\sigma_{+j}^2)^{-1}w_{+j}\big\{b(g) + (\sigma_{i+}^2)^{-1}w_{i+} + (\sigma_{+j}^2)^{-1}w_{+j} - 2(\sigma_{++}^2)^{-1}w_{++}\big\} && \text{(B.50)} \\
&\quad - 2\sum_{i=1}^{N}\sum_{j\in S_J(i)} \sigma_{ij}^2(\sigma_{++}^2)^{-1}w_{++}\big\{b(g) + (\sigma_{i+}^2)^{-1}w_{i+} + (\sigma_{+j}^2)^{-1}w_{+j} - 2(\sigma_{++}^2)^{-1}w_{++}\big\}. && \text{(B.51)}
\end{aligned}$$

We evaluate each of these four terms separately. For the first term,

$$\begin{aligned}
\text{(B.48)} =& b^2(g)\sigma_{++}^2 + b(g)\sum_{i=1}^{N}\Big(\sum_{j\in S_J(i)}\sigma_{ij}^2\Big)(\sigma_{i+}^2)^{-1}w_{i+} - b(g)w_{++} \\
&+ b(g)\sum_{j=1}^{J} w_{+j} - b(g)w_{++} \\
=& b^2(g)\sigma_{++}^2 \\
=& (\sigma_{++}^2)^{-1}w_{++}^2,
\end{aligned}$$

where the last line follows from (B.28). Now consider the second term,

$$\begin{aligned}
\text{(B.49)} =& b(g)w_{++} + \sum_{i=1}^{N} w_{i+}^2(\sigma_{i+}^2)^{-1} \\
&+ \sum_{i=1}^{N}\sum_{j\in S_J(i)} \sigma_{ij}^2(\sigma_{i+}^2)^{-1}w_{i+}(\sigma_{+j}^2)^{-1}w_{+j} - 2(\sigma_{++}^2)^{-1}w_{++}^2 \\
=& -(\sigma_{++}^2)^{-1}w_{++}^2 + \sum_{i=1}^{N} w_{i+}^2(\sigma_{i+}^2)^{-1} + \sum_{i=1}^{N}\sum_{j\in S_J(i)} w_{i+}w_{+j}\sigma_{ij}^2(\sigma_{i+}^2)^{-1}(\sigma_{+j}^2)^{-1}.
\end{aligned}$$

Now consider the third term,

$$\begin{aligned}
\text{(B.50)} =& b(g)w_{++} + \sum_{i=1}^{N}\sum_{j\in S_J(i)} \sigma_{ij}^2(\sigma_{+j}^2)^{-1}w_{+j}(\sigma_{i+}^2)^{-1}w_{i+} \\
&+ \sum_{j=1}^{J} w_{+j}^2(\sigma_{+j}^2)^{-1} - 2(\sigma_{++}^2)^{-1}w_{++}^2 \\
=& -(\sigma_{++}^2)^{-1}w_{++}^2 + \sum_{j=1}^{J} w_{+j}^2(\sigma_{+j}^2)^{-1} + \sum_{i=1}^{N}\sum_{j\in S_J(i)} w_{i+}w_{+j}\sigma_{ij}^2(\sigma_{i+}^2)^{-1}(\sigma_{+j}^2)^{-1}.
\end{aligned}$$

Now consider the last term,

$$\begin{aligned}(\text{B.51}) &= -2b(g)w_{++} - 2(\sigma^2_{++})^{-1}w^2_{++} - 2(\sigma^2_{++})^{-1}w^2_{++} + 4b(g)w_{++} \\ &= -2(\sigma^2_{++})^{-1}w^2_{++} - 2(\sigma^2_{++})^{-1}w^2_{++} - 2(\sigma^2_{++})^{-1}w^2_{++} + 4(\sigma^2_{++})^{-1}w^2_{++} \\ &= -2w^2_{++}(\sigma^2_{++})^{-1}.\end{aligned}$$

Combining all these four terms together, we obtain

$$\begin{aligned}\|d'(g)\|^2_\sigma =& \sum_{i=1}^{N} w^2_{i+}(\sigma^2_{i+})^{-1} + \sum_{j=1}^{J} w^2_{+j}(\sigma^2_{+j})^{-1} \\ &+ 2\sum_{i=1}^{N}\sum_{j\in S_J(i)} w_{i+}w_{+j}\sigma^2_{ij}(\sigma^2_{i+})^{-1}(\sigma^2_{+j})^{-1} - 3w^2_{++}(\sigma^2_{++})^{-1}.\end{aligned}$$

Hence the result of the lemma follows. □

Lemma B.14 below gives an analytical upper bound for $\|d''(g)\|_\sigma$ so that we can show it is a negligible term under certain conditions. Define

$$l_i = -\sum_{j\in S_J(i)} w_{+j}\sigma^2_{ij}(\sigma^2_{+j})^{-1} + w_{++}\sigma^2_{i+}(\sigma^2_{++})^{-1}, \quad i = 1, ..., N, \tag{B.52}$$

$$v_j = -\sum_{i\in S_N(j)} w_{i+}\sigma^2_{ij}(\sigma^2_{i+})^{-1} + w_{++}\sigma^2_{+j}(\sigma^2_{++})^{-1}, \quad j = 1, ..., J. \tag{B.53}$$

Lemma B.14. *If l_i and v_j are defined as in* (B.52) *and* (B.53)*, respectively, then*

$$\|d''(g)\|_\sigma \leq \gamma_N^{-1}\Big[\sum_{i=1}^{N} l_i^2(\sigma^2_{i+})^{-1} + \sum_{j=1}^{J} v_j^2(\sigma^2_{+j})^{-1}\Big].$$

Proof. From the definitions of f_i'', m_j'', l_i and v_j as in (B.43), (B.44), (B.52) and (B.53), respectively, it can be easily verified that

$$\sigma_{i+}^2 f_i'' + \sum_{j \in S_J(i)} \sigma_{ij}^2 m_j'' = l_i, \quad i = 1, ..., N,$$
$$\sigma_{+j}^2 m_j'' + \sum_{i \in S_N(j)} \sigma_{ij}^2 f_i'' = v_j, \quad j = 1, ..., J.$$

It can be shown $\|d''(g)\|_\sigma^2 = \sum_{i=1}^N f_i'' l_i + \sum_{j=1}^J m_j'' v_j$, which can be seen as follows,

$$\begin{aligned}
\sum_{i=1}^N f_i'' l_i + \sum_{j=1}^J m_j'' v_j &= \sum_{i=1}^N f_i'' \big(\sigma_{i+}^2 f_i'' + \sum_{j \in S_J(i)} \sigma_{ij}^2 m_j''\big) + \sum_{j=1}^J m_j'' \big(\sigma_{+j}^2 m_j'' + \sum_{i \in S_N(j)} \sigma_{ij}^2 f_i''\big) \\
&= \sum_{i=1}^N \sigma_{i+}^2 f_i''^2 + \sum_{j=1}^J \sigma_{+j}^2 m_j''^2 + 2 \sum_{i=1}^N \sum_{j \in S_J(i)} f_i'' m_j'' \sigma_{ij}^2 \\
&= \sum_{i=1}^N \sum_{j \in S_J(i)} (f_i'' + m_j'')^2 \sigma_{ij}^2 \\
&= \|d''(g)\|_\sigma^2.
\end{aligned}$$

Furthermore, by *Rao* (1973), $\sum_{i=1}^N f_i'' l_i + \sum_{j=1}^J m_j'' v_j$ is the largest value of $\big(\sum_{i=1}^N x_i l_i + \sum_{j=1}^J y_j v_j\big)^2$, for $x_i \in \mathbb{R}$, $i = 1, ..., N$ and $y_j \in \mathbb{R}$, $j = 1, ..., J$ such that

$$\sum_{i \in S_N(j)} \frac{1}{|S_J(i)|} \sigma_{i+}^2 x_i = 0, \quad j = 1, ..., J,$$
$$\sum_{j \in S_J(i)} \frac{1}{|S_N(j)|} \sigma_{+j}^2 y_j = 0, \quad i = 1, ..., N,$$
$$D(x, y) = \sum_{i=1}^N \sum_{j \in S_J(i)} \sigma_{ij}^2 (x_i + y_j)^2 \leq 1.$$

Note

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{j\in S_J(i)}(x_i+y_j)^2\check{\sigma}_{ij}^2\\
=&\sum_{i=1}^{N}\sum_{j\in S_J(i)}(x_i+y_j)^2\Big\{\sigma_{ij}^2-\gamma_N\big(\frac{1}{|S_J(i)|}\sigma_{i+}^2+\frac{1}{|S_N(j)|}\sigma_{+j}^2\big)\Big\}\\
=&D(x,y)-\gamma_N\sum_{i=1}^{N}\sum_{j\in S_J(i)}(x_i+y_j)^2\Big\{\frac{1}{|S_J(i)|}\sigma_{i+}^2+\frac{1}{|S_N(j)|}\sigma_{+j}^2\Big\}\\
=&D(x,y)\\
&-\gamma_N\Big\{\sum_{i=1}^{N}\big(x_i^2\sigma_{i+}^2+\sum_{j\in S_j(i)}\frac{1}{|S_N(j)|}x_i^2\sigma_{+j}^2\big)+\sum_{j=1}^{J}\big(y_j^2\sigma_{+j}^2+\sum_{i\in S_N(j)}\frac{1}{|S_J(i)|}y_j^2\sigma_{i+}^2\big)\Big\}\\
&-2\gamma_N\sum_{j=1}^{J}y_j\Big\{\sum_{i\in S_N(j)}\frac{1}{|S_J(i)|}\sigma_{i+}^2x_i\Big\}-2\gamma_N\sum_{i=1}^{N}x_i\Big\{\sum_{j\in S_J(i)}\frac{1}{|S_N(j)|}\sigma_{+j}^2y_j\Big\}\\
=&D(x,y)\\
&-\gamma_N\Big\{\sum_{i=1}^{N}\big(x_i^2\sigma_{i+}^2+\sum_{j\in S_j(i)}\frac{1}{|S_N(j)|}x_i^2\sigma_{+j}^2\big)+\sum_{j=1}^{J}\big(y_j^2\sigma_{+j}^2+\sum_{i\in S_N(j)}\frac{1}{|S_J(i)|}y_j^2\sigma_{i+}^2\big)\Big\}.
\end{aligned}
$$

Re-arrange gives,

$$
\begin{aligned}
D(x,y)=&\gamma_N\Big\{\sum_{i=1}^{N}\big(x_i^2\sigma_{i+}^2+\sum_{j\in S_j(i)}\frac{1}{|S_N(j)|}x_i^2\sigma_{+j}^2\big)+\sum_{j=1}^{J}\big(y_j^2\sigma_{+j}^2+\sum_{i\in S_N(j)}\frac{1}{|S_J(i)|}y_j^2\sigma_{i+}^2\big)\Big\}\\
&+\sum_{i=1}^{N}\sum_{j\in S_J(i)}(x_i+y_j)^2\check{\sigma}_{ij}^2\\
&\geq\gamma_N\Big\{\sum_{i=1}^{N}x_i^2\sigma_{i+}^2+\sum_{j=1}^{J}y_j^2\sigma_{+j}^2\Big\}.
\end{aligned}
$$

It follows that

$$\|d''(g)\|_\sigma \leq \gamma_N^{-1}\Big[\sum_{i=1}^{N} l_i^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^{J} v_j^2(\sigma_{+j}^2)^{-1}\Big].$$

□

Next, we give proofs for the supporting lemmas used in the proofs of Proposition III.9 and the proofs of Theorems III.10 and III.11.

LEMMA B.1. *Assume Conditions III.6, III.7 and III.8 hold. If* $A_p = \{f_{ij} : i = 1, ..., N, j = 1, ..., J, z_{ij} = 1\}$ *such that* $f_{ij}(x) = x_{ij}$ *for* $x \in \Omega_N$. *Let* $C_N = |A_p|$, *the cardinality of* A_p. *There exist sequences* $f_N > 0$ *and* $d_N \geq 0$ *satisfying the followings.*
(a). As $N \to \infty$, $f_N^2 / \log C_N \to \infty$.
(b). As $N \to \infty$, $f_N^2(N_*^{-1} + J_*^{-1}) \to 0$.
(c). If $y, v \in \Omega_N$ *and* $\|y - M_N^*\|_\sigma(A_p) \leq f_N$, *then there exists* $n < \infty$ *such that for all* $N > n$, $\|U_N(y, v)\|_\sigma(A_p) \leq d_N \|y - M_N^*\|_\sigma(A_p)\|v\|_\sigma(A_p)$. *Furthermore,* $d_N f_N \to 0$ *as* $N \to \infty$.

Proof. Condition III.6(b) assumes $J_*^{-1} \log N \to 0$. If $x_N \ll y_N$ means that $y_N^{-1} x_N \to 0$ as $N \to \infty$, then Condition III.6(b) implies that $\log N^* \ll J_*$. Then there must

exist a sequence $f_N > 0$ such that $\log N^* \ll f_N^2 \ll J_*$.

$$\begin{aligned} f_N^2 / \log C_N &\geq \frac{f_N^2}{\log(J^* N^*)} \\ &= \frac{f_N^2}{\log J^* + \log N^*} \\ &\geq \frac{f_N^2}{2 \log N^*} \to \infty \quad \text{as} \quad N \to \infty. \end{aligned}$$

The first inequality follows from the fact that $J_* N_* \leq C_N \leq J^* N^*$. The last line follows from $\log N^* \ll f_N^2$. Therefore, the result of part (a) is satisfied. We further note

$$f_N^2 (N_*^{-1} + J_*^{-1}) \leq \frac{2 f_N^2}{J_*} \to 0 \quad \text{as} \quad N \to \infty. \tag{B.54}$$

The last line follows from $f_N^2 \ll J_*$. Therefore, part (b) of the lemma follows. To verify part (c), first note by Lemma B.4, for any point maps $f_{ij} \in A_p$, there exist $0 < \tau_1, \tau_2 < \infty$ such that for all $N > n$,

$$\tau_1^{-1} \left(N_*^{-1} + J_*^{-1} \right)^{\frac{1}{2}} < \sigma(f_{ij}) < \tau_2^{-1} \left(N_*^{-1} + J_*^{-1} \right)^{\frac{1}{2}}. \tag{B.55}$$

By the definition of $\| \cdot \|_\sigma(A_p)$, we have for any $y \in \Omega_N, f_{ij} \in A_p$,

$$|f_{ij}(y)| \leq \|y\|_\sigma(A_p) \sigma(f_{ij}). \tag{B.56}$$

It follows from (B.55) and (B.56) that for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|y_{ij}| \leq \tau_2^{-1} \|y\|_\sigma(A_p) \left(N_*^{-1} + J_*^{-1}\right)^{\frac{1}{2}}. \tag{B.57}$$

Since $|\sigma^2(y_{ij}) - \sigma_{ij}^2| \leq 1$, note that there exists a positive $\tau_3 < \infty$ such that for any $y \in \Omega_N$, one has for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|\sigma^2(y_{ij}) - \sigma_{ij}^2| \leq \tau_3 |y_{ij} - m_{ij}^*|. \tag{B.58}$$

Since A_p consists of point maps only, by the definition of $\|\cdot\|_\sigma(A_p)$, we have $\|U_N(y,v)\|_\sigma(A_p)$ is the maximum value of $|f_{ij}\{U_N(y,v)\}|/\sigma(f_{ij})$ over $f_{ij} \in A_p$. Therefore, upper bounding $\|U_N(y,v)\|_\sigma(A_p)$ is equivalent to upper bounding all $|U_{ij}(y,v)|/\sigma(f_{ij})$. Note that for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$\begin{aligned}
|U_{ij}(y,v)| &= \Big| \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} \Big[d_{i'j'}(f_{ij}) \{\sigma^2(y_{i'j'}) - \sigma_{i'j'}^2\} v_{i'j'} \Big] \Big| \\
&\leq \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} \{|d_{i'j'}(f_{ij})|\} \{|\sigma^2(y_{i'j'}) - \sigma_{i'j'}^2|\} \{|v_{i'j'}|\} \\
&\leq \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} \{|d_{i'j'}(f_{ij})|\} \{\tau_3 |y_{i'j'} - m_{i'j'}^*|\} \{|v_{i'j'}|\} \\
&\leq \tau_2^{-2} \tau_3 \left(N_*^{-1} + J_*^{-1}\right) \Big\{ \|y - M_N^*\|_\sigma(A_p) \|v\|_\sigma(A_p) \Big\} \Big\{ \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} |d_{i'j'}(f_{ij})| \Big\},
\end{aligned}$$

where the second last line follows from (B.58) and the last line follows from (B.57).

Further note that

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d_{i'j'}(f_{ij})| \le \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(f_{ij})| + \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(f_{ij})|.$$

By definition, $d'_{i'j'}(g) = (\sigma^2_{i'+})^{-1}w_{i'+} + (\sigma^2_{+j'})^{-1}w_{+j'} - (\sigma_{++})^{-1}w_{++}$ for any $g \in \Omega^*_N$. When $g = f_{ij}$, $w_{i'+} = 0$ if $i' \neq i$, and $w_{i'+} = 1$ if $i' = i$, $w_{+j'} = 0$ if $j' \neq j$, and $w_{+j'} = 1$ if $j' = j$, and $w_{++} = 1$. Therefore, we can rewrite

$$\begin{aligned}
\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(f_{ij})| =& \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}\left|(\sigma^2_{i'+})^{-1}w_{i'+} + (\sigma^2_{+j'})^{-1}w_{+j'} - (\sigma_{++})^{-1}w_{++}\right| \\
\le& \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}(\sigma^2_{i'+})^{-1}|w_{i'+}| + \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}(\sigma^2_{+j'})^{-1}|w_{+j'}| \\
&+ \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}(\sigma_{++})^{-1}|w_{++}| \\
=& \sum_{j'\in S_J(i)}(\sigma^2_{i+})^{-1} + \sum_{i'\in S_N(j)}(\sigma^2_{+j})^{-1} + \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}(\sigma_{++})^{-1} \le \tau_4,
\end{aligned}$$

where τ_4 is some positive constant such that $\tau_4 < \infty$. Note also that there exists $\tau_5 < \infty$ such that

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(f_{ij})| \le (J^*N^*)^{\frac{1}{2}}\|d''(f_{ij})\|_\sigma \le \tau_5.$$

As a result,

$$\|U_N(y,v)\|_\sigma(A_p) \le \tau_1\tau_2^{-2}\tau_3(\tau_4+\tau_5)\left(N_*^{-1}+J_*^{-1}\right)^{\frac{1}{2}}\left\{\|y - M_N^*\|_\sigma(A_p)\|v\|_\sigma(A_p)\right\}.$$

Therefore, we can set $d_N = \tau_1\tau_2^{-2}\tau_3(\tau_4+\tau_5)\,(N_*^{-1}+J_*^{-1})^{\frac{1}{2}}$. By (B.54), we have $f_N(N_*^{-1}+J_*^{-1})^{1/2} \to 0$ as $N \to \infty$. Therefore, it follows

$$d_N f_N = \tau_1\tau_2^{-2}\tau_3(\tau_4+\tau_5)\left(N_*^{-1}+J_*^{-1}\right)^{\frac{1}{2}} f_N \to 0, \quad \text{as} \quad N \to \infty.$$

Hence, the result of part (c) is also satisfied. □

LEMMA B.2. *Let $A \subset \Omega_N^*$. Let C_N denote the cardinality of A. If there exist sequences $f_N > 0$ and $d_N \geq 0$ satisfying (a). $0 < C_N < \infty$ and $f_N^2/\log C_n \to \infty$ as $N \to \infty$, (b). If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A) \leq f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y,v)\|_\sigma(A) \leq d_N\|y - M_N^*\|_\sigma(A)\|v\|_\sigma(A)$, (c). $d_N f_N \to 0$ as $N \to \infty$. Then $P\big(\|R_N\|_\sigma(A) < \frac{1}{2}f_N\big) \to 1$ as $N \to \infty$.*

Proof. Denote $A = \{g_k : k = 1, ..., C_N\}$ and let $w_k \in \Omega_N$ be defined for $k = 1, ..., C_N$ by $g_k(x) = [w_k, x]_\sigma, x \in \Omega_N$. Let $W_k = \|w_k\|_\sigma^{-1}\sum_{i=1}^N \sum_{j\in S_J(i)} w_{ijk}(Y_{ij} - E_{ij})$ for $k = 1, ..., C_N$ so that $\|R_N\|_\sigma(A) = \max_{k=1,...,C_N}|W_k|$. We consider the log moment generating function of W_k, denoted as $\log G_k(t)$. Write $w_k' = w_k/\|w_k\|_\sigma$,

$k = 1, ..., C_N$, for simplicity, and we have

$$\begin{aligned}
\log G_k(t) &= \log \mathbb{E}[e^{tW_k}] \\
&= \log \mathbb{E}\Big[\exp\Big\{ \frac{t}{\|w_k\|_\sigma} \sum_{i=1}^{N} \sum_{j \in S_J(i)} w_{ijk}(Y_{ij} - E_{ij}) \Big\} \Big] \\
&= -t \sum_{i=1}^{N} \sum_{j \in S_J(i)} w'_{ijk} E_{ij} + \log \prod_{i=1}^{N} \prod_{j \in S_J(i)} \mathbb{E}\big\{ \exp(t w'_{ijk} Y_{ij}) \big\}, \quad \text{by independence} \\
&= -t \sum_{i=1}^{N} \sum_{j \in S_J(i)} w'_{ijk} E_{ij} + \sum_{i=1}^{N} \sum_{j \in S_J(i)} \log \mathbb{E}\big\{ \exp(t w'_{ijk} Y_{ij}) \big\} \\
&= \sum_{i=1}^{N} \sum_{j \in S_J(i)} \Big[\log\{1 + \exp(m^*_{ij})\}^{-1} - \log\{1 + \exp(t w'_{ijk} + m^*_{ij})\}^{-1} - t w'_{ijk} E_{ij} \Big] \\
&= \sum_{i=1}^{N} \sum_{j \in S_J(i)} \Big[\log\{h(m^*_{ij})\} - \log\{h(t w'_{ijk} + m^*_{ij})\} - t w'_{ijk} E_{ij} \Big], \qquad \text{(B.59)}
\end{aligned}$$

where we have denoted $h(x) = \{1 + \exp(x)\}^{-1}$. We apply Taylor expansion to $\log\{h(t w'_{ijk} + m^*_{ij})\}$ with respect to m^*_{ij}. For some $t' = \alpha t$ with $0 < \alpha < 1$, we have,

$$\log\{h(t w'_{ijk} + m^*_{ij})\} = \log h(m^*_{ij}) - E_{ij} t w'_{ijk} - \frac{t^2}{2} w'^2_{ijk} \sigma^2\big(m^*_{ij} + t' w'_{ijk}\big).$$

Substitute into (B.59),

$$\log G_k(t) = \frac{t^2}{2} \sum_{i=1}^{N} \sum_{j \in S_J(i)} w'^2_{ijk} \sigma^2\big(m^*_{ij} + t' w'_{ijk}\big), \qquad |t| \le f_N.$$

By *Bahadur* (1971),

$$P\Big(W_k \geq \frac{1}{2}f_N\Big) \leq \exp\Big(-\frac{1}{4}f_N^2\Big)G_k\Big(\frac{1}{2}f_N\Big), \quad k = 1, ..., C_N,$$

and

$$P\Big(-W_k \geq \frac{1}{2}f_N\Big) \leq \exp\Big(-\frac{1}{4}f_N^2\Big)G_k\Big(-\frac{1}{2}f_N\Big), \quad k = 1, ..., C_N.$$

Furthermore note that,

$$\log G_k\Big(\frac{1}{2}f_N\Big), \log G_k\Big(-\frac{1}{2}f_N\Big) \leq \frac{1}{8}f_N^2\Big(1 + \frac{d_N f_N}{2}\Big) \quad k = 1, ..., C_N.$$

Applying the Bonferroni inequality,

$$\begin{aligned} P\Big\{\|R_N\|_\sigma(A) \geq \frac{1}{2}f_N\Big\} &\leq 2C_N \exp\Big\{-\frac{1}{8}f_N^2\big(1 - \frac{d_N f_N}{2}\big)\Big\} \\ &= 2\exp\Big\{\log C_N - \frac{1}{8}f_N^2\big(1 - \frac{d_N f_N}{2}\big)\Big\} \\ &\to 0 \quad \text{as} \quad N \to \infty, \end{aligned}$$

where the last step follows from the assumption $f_N^2/\log C_N \to \infty$ as $N \to \infty$. Hence the result of the lemma follows. □

LEMMA B.3. *Assume Conditions III.6, III.7 and III.8 hold. Let $A \subset \Omega_N^*$. If there exist sequences $f_N > 0$ and $d_N \geq 0$ satisfying (a). $P\big(\|R_N\|_\sigma(A) < \frac{1}{2}f_N\big) \to 1$ as $N \to \infty$, (b). If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A) \leq f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y, v)\|_\sigma(A) \leq d_N\|y - M_N^*\|_\sigma(A)\|v\|_\sigma(A)$, (c). $d_N f_N \to 0$ as*

$N \to \infty$. *Then, as* $N \to \infty$, *with probability approaching 1 that,*

$$\Big|\frac{\|\hat{M}_N - M_N^*\|_\sigma(A)}{\|R_N\|_\sigma(A)} - 1\Big| \le d_N^{\frac{1}{2}} \to 0 \quad \text{and} \quad \|\hat{M}_N - M_N^* - R_N\|_\sigma(A) \le d_N \|R_N\|_\sigma^2(A).$$

Proof. Write $z_N = \|R_N\|_\sigma(A)$ for simplicity. Consider a sequence $\{h_{Nk} : k = 0, 1, ...\}$, with $h_{N0} = 0$ and $h_{N(k+1)} = z_N + d_N h_{Nk}^2/2$ for $k = 0, 1, 2, \ldots$. Define another sequence

$$l_N = \frac{2z_N}{1 + (1 - 2z_N d_N)^{\frac{1}{2}}}.$$

By Kantorovich and Akilov (1964, pages 695-711), if $z_N < \frac{1}{2} f_N$ and $z_N d_N < \frac{1}{2}$ (which hold with probability tending to 1 by (a), (b) and (c)), it follows

$$\|t_{Nk} - \hat{M}_N\|_\sigma(A) \le l_N - h_{Nk}, \quad k = 0, 1, 2, ..., \tag{B.60}$$

where $\{t_{Nk} : k = 0, 1, ...\}$ is the sequence constructed in the proof of Theorem III.10. When $k = 0$, (B.60) implies $\|M_N^* - \hat{M}_N\|_\sigma(A) \le l_N$. When $k = 1$, (B.60) implies

$$\|M_N^* + R_N - \hat{M}_N\|_\sigma(A) \le l_N - z_N. \tag{B.61}$$

It follows that $\big|\|M_N^* - \hat{M}_N\|_\sigma(A) - \|R_N\|_\sigma(A)\big| \le l_N - z_N$, where

$$l_N - z_N = \frac{z_N\{1 - (1 - 2z_N d_N)^{\frac{1}{2}}\}}{1 + (1 - 2z_N d_N)^{\frac{1}{2}}}.$$

If we view $x = z_N d_N$ and $f(x) = \{1-(1-2x)^{1/2}\}/\{1+(1-2x)^{1/2}\}$. We note $f(0) = 0$, $f(1/2) = 1$ and $f'(0) = 1/4 < 1$ and $f''(x) > 0$ for all $x < 1/2$. Therefore, $f(x) < x$

for all $x < 1/2$. Hence, whenever $d_N z_N < 1/2$, we must have $l_N - z_N \leq d_N z_N^2$. We know that with probability tending to 1 that $d_N z_N < 1/2$. Hence the second part of the lemma follows from (B.61). Also as $N \to \infty$, with probability approaching 1 that,

$$\left| \|\hat{M}_N - M_N^*\|_\sigma(A) - \|R_N\|_\sigma(A) \right|^2 \leq d_N \|R_N\|_\sigma^2(A). \tag{B.62}$$

Re-write (B.62), the result of the first part of the lemma then follows. □

LEMMA B.4. *Assume Conditions III.6, III.7 and III.8 hold and* $\sum_{i=1}^N \theta_i = 0$*, the asymptotic variance of the maximum likelihood estimator of* m_{ij}^**,* $var(\hat{m}_{ij})$*, for any* $i = 1, ..., N$ *and* $j = 1, ..., J$*, takes the form,*

$$var(\hat{m}_{ij}) = (\sigma_{i+}^2)^{-1} + (\sigma_{+j}^2)^{-1} + O(N_*^{-1} J_*^{-1}) \quad \text{as} \quad N \to \infty.$$

Proof. If $z_{ij} = 1$, then we can simply use a linear function f_{ij} with $f_{ij}(x) = x_{ij}$. We apply $\|d'(f_{ij})\|_\sigma^2$ to approximate $\sigma^2(f_{ij})$. With $w_{i+} = 1, w_{k+} = 0$, for all $k = 1, ..., i-1, i+1, ..., N, w_{+j} = 1, w_{+l} = 0$ for all $l = 1, ..., j-1, j+1, ..., J$ and $w_{++} = 1$. We obtain

$$\|d'(f_{ij})\|_\sigma^2 = (\sigma_{i+}^2)^{-1} + (\sigma_{+j}^2)^{-1} + O\left(N_*^{-1} J_*^{-1}\right) \qquad \text{as} \quad N \to \infty.$$

If $z_{ij} = 0$, then we can apply Condition III.8, there must exist $1 \leq i_1, i_2, ..., i_k \leq N$ and $1 \leq j_1, j_2, ..., j_k \leq J$ such that $z_{ij_1} = z_{i_1 j_1} = z_{i_1 j_2} = z_{i_2 j_2} = ... = z_{i_k j_k} = z_{i_k j} = 1$.

Consider a linear function g_2 defined as

$$\begin{aligned} g_{ij}(x) &= x_{ij_1} - x_{i_1j_1} + x_{i_1j_2} - x_{i_2j_2} + ... + x_{i_{k-1}j_k} - x_{i_kj_k} + x_{i_kj} \\ &= \theta_i - \beta_j. \end{aligned}$$

In this case, similarly we have $w_{i+} = 1, w_{k+} = 0$, for all $k = 1, ..., i-1, i+1, ..., N, w_{+j} = 1, w_{+l} = 0$ for all $l = 1, ..., j-1, j+1, ..., J$ and $w_{++} = 1$. Note these values are exactly the same as those of g_1. Therefore,

$$\|d'(g_{ij})\|_\sigma^2 = (\sigma_{i+}^2)^{-1} + (\sigma_{+j}^2)^{-1} + O\left(N_*^{-1}J_*^{-1}\right) \qquad \text{as} \quad N \to \infty.$$

In both cases, $\|d''\|_\sigma^2$ has a small order. To see this, note that in both cases above,

$$\begin{aligned} l_p &= \begin{cases} -\sigma_{pj}^2(\sigma_{+j}^2)^{-1} + \sigma_{p+}^2(\sigma_{++}^2)^{-1} & \text{if} \quad z_{pj} = 1 \\ \sigma_{p+}^2(\sigma_{++}^2)^{-1} & \text{if} \quad z_{pj} = 0 \end{cases} \\ &= O\left(N_*^{-1}\right) \qquad \text{as} \quad N \to \infty, \qquad p = 1, ..., N. \end{aligned}$$

$$\begin{aligned} v_q &= \begin{cases} -\sigma_{iq}^2(\sigma_{i+}^2)^{-1} + \sigma_{+q}^2(\sigma_{++}^2)^{-1} & \text{if} \quad z_{iq} = 1 \\ \sigma_{+q}^2(\sigma_{++}^2)^{-1} & \text{if} \quad z_{iq} = 0 \end{cases} \\ &= O\left(J_*^{-1}\right) \qquad \text{as} \quad N \to \infty, \qquad q = 1, ..., J. \end{aligned}$$

It follows that

$$\|d''(f_{ij})\|_\sigma^2 = \|d''(g_{ij})\|_\sigma^2 \le \gamma_N^{-2}\Big\{\sum_{p=1}^{N} l_p^2(\sigma_{p+}^2)^{-1} + \sum_{q=1}^{J} v_q^2(\sigma_{+q}^2)^{-1}\Big\}^2$$
$$= O\left(N_*^{-1}J_*^{-2}\right) \qquad \text{as} \quad N \to \infty,$$

where the last line follows from Condition III.6(a). Since for any $g \in \Omega_N^*$,

$$\left(\|d'(g)\|_\sigma - \|d''(g)\|_\sigma\right)^2 \le \sigma^2(g) \le \left(\|d'(g)\|_\sigma + \|d''(g)\|_\sigma\right)^2,$$

it follows $\mathrm{var}(\hat{m}_{ij}) = (\sigma_{i+}^2)^{-1} + (\sigma_{+j}^2)^{-1} + O(N_*^{-1}J_*^{-1})$ as $N \to \infty$. □

LEMMA B.5. *Assume Conditions III.6, III.7 and III.8 hold. If $A_\beta = \{g_j : j = 1, ..., J\}$ such that $g_j \in \Omega_N^*$ and $g_j(x) = \beta_j$ for $x \in \Omega_N$. Let $C_N = |A_\beta| = J$ be the cardinality of A_β. For any positive sequence f_N such that $f_N^2/\log J \to \infty$ and $f_N^2 N_*^{-1/2} \to 0$ as $N \to \infty$, there exists a sequence $d_N \ge 0$ satisfying the followings.*
(a). If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A_\beta) \le f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y, v)\|_\sigma(A_\beta) \le d_N\|y - M_N^*\|_\sigma(A_\beta)\|v\|_\sigma(A_\beta)$.
(b). $d_N f_N^2 \to 0$ as $N \to \infty$.

Proof. First we note since we have assumed $N > J$, we must have $\log J \ll N_*^{1/2}$ by Condition III.6(a), so the rate requirements for f_N is valid. To find a valid d_N, we seek to upper bound $\|U_N(y, v)\|_\sigma(A_\beta)$ and then show that $d_N f_N \to 0$ as $N \to \infty$ for all f_N satisfying the rate requirements $f_N^2/\log J \to \infty$ and $f_N^2 N_*^{-1/2} \to 0$ as $N \to \infty$.

For any $y, v \in \Omega_N$, by the definition of $\|\cdot\|_\sigma(A_\beta)$, we have

$$\|U_N(y,v)\|_\sigma(A_\beta) = \max_{g_j \in A_\beta} |g_j\{U_N(y,v)\}|/\sigma(g_j).$$

First note that by Lemma B.6, $\sigma^2(g_j) = (\sigma^2_{+j})^{-1} + O\big\{(N_* J_*)^{-1}\big\}$ for any $g_j \in A_\beta$. Therefore, there exist positive $0 < c_1, c_2 < \infty$ such that for all $N > n$, $c_1^{-1} N_*^{-1/2} < \sigma(g_j) < c_2^{-1} N_*^{-1/2}$, for all $g_j \in A_\beta$. So we just need to find an upper bound for $|g_j\{U_N(y,v)\}|$ that holds for all $g_j \in A_\beta$. Consider

$$\begin{aligned} |g_j\{U_N(y,v)\}| &= \Big| \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} d_{i'j'}(g_j)\{\sigma^2(y_{i'j'}) - \sigma^2_{i'j'}\} v_{i'j'} \Big| \\ &\le \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} |d_{i'j'}(g_j)| \cdot |\sigma^2(y_{i'j'}) - \sigma^2_{i'j'}| \cdot |v_{i'j'}|. \end{aligned}$$

Note $0 \le \sigma^2(y_{ij}), \sigma^2_{ij} \le 1$, so $|\sigma^2(y_{ij}) - \sigma^2_{ij}| \le 1$. It can be implied that there exists some positive $c_3 < \infty$ such that $|\sigma^2(y_{ij}) - \sigma^2_{ij}| \le c_3 |g_j(y - M_N^*)|$. Again, by the definition of $\|\cdot\|_\sigma(A_\beta)$, we have $|g_j(y - M_N^*)| \le \|y - M_N^*\|_\sigma(A_\beta)\sigma(g_j)$. Therefore, for all $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|\sigma^2(y_{ij}) - \sigma^2_{ij}| \le c_2^{-1} c_3 N_*^{-1/2} \|y - M_N^*\|_\sigma(A_\beta).$$

On the other hand, using a similar strategy, we can show that there exists a positive $c_4 < \infty$ such that for all $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|v_{ij}| \le c_2^{-1} c_4 N_*^{-1/2} \|v\|_\sigma(A_\beta).$$

Further note that

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d_{i'j'}(g_j)| \le \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(g_j)| + \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(g_j)|.$$

By definition, we know $d'_{i'j'} = (\sigma^2_{i'+})^{-1}w_{i'+} + (\sigma^2_{+j'})^{-1}w_{+j'} - (\sigma^2_{++})^{-1}w_{++}$. For any $g_j \in A_\beta$, $w_{i'+} = -1/N$, for $i' = 1, ..., N$, $w_{+j'} = -1$ if $j' = j$ and $w_{+j'} = 0$ if $j' \neq j$, $w_{++} = -1$. Hence,

$$d'_{i'j'}(g_j) = \begin{cases} -\frac{1}{N}(\sigma^2_{i'+})^{-1} - (\sigma^2_{+j'})^{-1} + (\sigma^2_{++})^{-1} & \text{if} \quad j' = j \\ -\frac{1}{N}(\sigma^2_{i'+})^{-1} + (\sigma^2_{++})^{-1} & \text{if} \quad j' \neq j. \end{cases}$$

It follows

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(g_j)| \le \frac{J^*}{N}\sum_{i'=1}^{N}(\sigma^2_{i'+})^{-1} + N^*(\sigma^2_{+j})^{-1} + \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}(\sigma^2_{++})^{-1} \le c_5,$$

for some positive $c_5 < \infty$. On the other hand,

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(g_j)| \le (N^*J^*)^{\frac{1}{2}}\|d''(g_j)\|_\sigma \le c_6,$$

for some positive $c_6 < \infty$. The last step follows from Condition III.6(c) and Lemma

B.6 which implies that $\|d''(g_j)\|_\sigma^2 = O\left(N_*^{-1}J_*^{-1}\right)$. Overall,

$$\begin{aligned}\|U_N(y,v)\|_\sigma(A_\beta) &= \max_{g_j\in A_\beta} |g_j\{U_N(y,v)\}|/\sigma(g_j) \\ &\le \max_{g_j\in A_\beta} |g_j\{U_N(y,v)\}| \cdot \max_{g_j\in A_\beta}\{\sigma^{-1}(g_j)\}. \\ &\le c_1 c_2^{-2} c_3 c_4 (c_5+c_6) N_*^{-\frac{1}{2}} \|y - M_N^*\|_\sigma(A_\beta)\|v\|_\sigma(A_\beta).\end{aligned}$$

Note that by taking $d_N = c_1 c_2^{-2} c_3 c_4 (c_5+c_6) N_*^{-1/2}$, part (a) of the lemma follows. Furthermore, by the rate requirement of f_N, for any positive sequence f_N such that $\log J \ll f_N^2 \ll N_*^{1/2}$, it can be seen easily that $d_N f_N^2 \to 0$ as $N \to \infty$. Therefore, part (b) of the lemma follows. □

LEMMA B.6. *Assume Conditions III.6, III.7 and III.8 hold and* $\sum_{i=1}^N \theta_i = 0$. *The asymptotic variance of the maximum likelihood estimator of an individual column parameter, var*$(\hat{\beta}_j)$*, asymptotically attains the oracle variance* $(\sigma_{+j}^2)^{-1}$ *in the sense that*

$$var(\hat{\beta}_j) = (\sigma_{+j}^2)^{-1} + O(N_*^{-1}J_*^{-1}) \qquad \text{as} \quad N \to \infty. \tag{B.63}$$

Proof. We seek to construct a linear function $g_j \in \Omega_N^*$ such that $g_j(x) = \beta_j$ so that we can use $\|d'(g_j)\|_\sigma^2$ defined in Lemma B.13 to approximate $\text{var}(\hat{\beta}_j)$. To construct such a g_j, we may want to include all x_{ij}, $i = 1, ..., N$, in g_j so that we can apply the constraint $\sum_{i=1}^N \theta_i = 0$ to solve for β_j. For $i \in S_N(j)$, we use $x_{ij} = \theta_i - \beta_j$ directly. For each $i \in S_{N_\phi}(j)$, by Condition III.8, there must exist $1 \le i_{i1}, i_{i2}, ..., i_{ik} \le N$ and $1 \le j_{i1}, j_{i2}, ..., j_{ik} \le J$ such that $z_{i,j_{i1}} = z_{i_{i1},j_{i1}} = z_{i_{i1},j_{i2}} = z_{i_{i2},j_{i2}} = ... = z_{i_{ik},j_{ik}} =$

$z_{i_{ik},j} = 1$, with

$$\begin{aligned}
& x_{i,j_{i1}} - x_{i_{i1},j_{i1}} + x_{i_{i1},j_{i2}} - x_{i_{i2},j_{i2}} + ... - x_{i_{ik},j_{ik}} + x_{i_{ik},j} \\
=&(\theta_i - \beta_{j_{i1}}) - (\theta_{i_{i1}} - \beta_{j_{i1}}) + (\theta_{i_{i1}} - \beta_{j_{i2}}) - (\theta_{i_{i2}} - \beta_{j_{i2}}) + \\
& ... - (\theta_{i_{ik}} - \beta_{j_{ik}}) + (\theta_{i_{ik}} - \beta_j) \\
=&\theta_i - \beta_j.
\end{aligned}$$

Therefore, we can construct g to be

$$\begin{aligned}
g_j(x) =& -\frac{1}{N}\Big\{ \sum_{i \in S_N(j)} x_{ij} \\
& + \sum_{i \in S_{N_\phi}(j)} \Big(x_{i,j_{i1}} - x_{i_{i1},j_{i1}} + x_{i_{i1},j_{i2}} - x_{i_{i2},j_{i2}} + ... - x_{i_{ik},j_{ik}} + x_{i_{ik},j} \Big) \Big\} \\
=& -\frac{1}{N}\Big\{ \Big(\sum_{i=1}^{N} \theta_i \Big) - N\beta_j \Big\} \\
=& \beta_j.
\end{aligned}$$

Use $\|d'(g_j)\|_\sigma^2$ from Lemma B.13 to approximate $\sigma^2(g_j)$, with $w_{i+} = -1/N$, for all $i = 1, ..., N$, $w_{+j} = -1, w_{+l} = 0$ for all $l = 1, ...j-1, j+1, ..., J$ and $w_{++} = -1$. It follows

$$\begin{aligned}
\|d'(g_j)\|_\sigma^2 &= (\sigma_{+j}^2)^{-1} + \frac{1}{N^2}\sum_{i=1}^{N}(\sigma_{i+}^2)^{-1} + \frac{2}{N}\sum_{i \in S_N(j)} \sigma_{ij}^2(\sigma_{i+}^2)^{-1}(\sigma_{+j}^2)^{-1} - 3(\sigma_{++}^2)^{-1} \\
&= \left(\sigma_{+j}^2\right)^{-1} + O\left(N_*^{-1}J_*^{-1}\right) \qquad \text{as} \quad N \to \infty.
\end{aligned}$$

To see whether $\|d'(g_j)\|_\sigma^2$ is a good approximation for $\sigma^2(g_j)$, we need to evaluate the order of $\|d''(g_j)\|_\sigma^2$ from Lemma B.14. Note

$$\begin{aligned} l_i &= \begin{cases} \sigma_{ij}^2(\sigma_{+j}^2)^{-1} - \sigma_{i+}^2(\sigma_{++}^2)^{-1} & \text{if} \quad z_{ij} = 1 \\ -\sigma_{i+}^2(\sigma_{++}^2)^{-1} & \text{if} \quad z_{ij} = 0 \end{cases} \\ &= O\left(N_*^{-1}\right) \qquad \text{as} \quad N \to \infty, \qquad i = 1, ..., N. \end{aligned}$$

$$\begin{aligned} v_q &= \frac{1}{N} \sum_{i \in S_N(q)} \sigma_{iq}^2(\sigma_{i+}^2)^{-1} - \sigma_{+q}^2(\sigma_{++}^2)^{-1} \\ &= O\left(J_*^{-1}\right) \qquad \text{as} \quad N \to \infty, \qquad q = 1, ..., J. \end{aligned}$$

Applying Lemma B.14, we have

$$\begin{aligned} \|d''(g_j)\|_\sigma^2 &\le \gamma_N^{-2} \Big\{ \sum_{i=1}^N l_i^2(\sigma_{i+}^2)^{-1} + \sum_{q=1}^J v_q^2(\sigma_{+q}^2)^{-1} \Big\}^2 \\ &= O\left(N_*^{-1} J_*^{-2}\right) \qquad \text{as} \quad N \to \infty, \end{aligned}$$

where the last line follows from Condition III.6(a). Since

$$\left(\|d'(g_j)\|_\sigma - \|d''(g_j)\|_\sigma\right)^2 \le \sigma^2(g_j) \le \left(\|d'(g_j)\|_\sigma + \|d''(g_j)\|_\sigma\right)^2,$$

It follows that $\text{var}(\hat{\beta}_j) = (\sigma_{+j}^2)^{-1} + O\left(N_*^{-1} J_*^{-1}\right)$ as $N \to \infty$. □

LEMMA B.7. Let a_N and c_N be positive sequences. As $N \to \infty$, suppose that $a_N = O(b_N)$, for any sequence b_N satisfying $b_N/c_N \to \infty$. Then $a_N = O(c_N)$ as

$N \to \infty$.

Proof. We prove the lemma result by a contradiction argument. Suppose the lemma result does not hold. Then there exists a subsequence N_n, such that

$$L_{N_n} = \frac{a_{N_n}}{c_{N_n}} \to \infty.$$

We define a sequence

$$b_N = \begin{cases} Nc_N & \text{if } N \notin \{N_1, N_2, ...\}, \\ L_{N_n}^{\frac{1}{2}} c_{N_n} & \text{if } N = N_n. \end{cases}$$

Then b_N satisfies $b_N/c_N \to \infty$, but "$a_N = O(b_N)$" does not hold. Contradiction. Hence, the lemma result holds.

□

LEMMA B.8. *Assume Conditions III.6, III.7 and III.8 hold. If $A_\theta = \{g_i : i = 1, ..., N\}$ such that $g_i \in \Omega_N^*$ and $g_i(x) = \theta_i$ for $x \in \Omega_N$. Let $C_N = |A_\theta| = N$ be the cardinality of A_θ. Then for any positive sequence f_N such that $f_N^2/\log N \to \infty$ and $J_*^{-1} f_N^2 \to 0$ as $N \to \infty$, there exists a sequence $d_N \geq 0$ satisfying the followings.*
(a) If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A_\theta) \leq f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y, v)\|_\sigma(A_\theta) \leq d_N \|y - M_N^*\|_\sigma(A_\theta) \|v\|_\sigma(A_\theta)$.
(b). $d_N f_N \to 0$ as $N \to \infty$.

Proof. We first note that from Condition III.6(b), $\log N \ll J_*$ as $N \to \infty$. Therefore, the rate requirements for the sequence f_N, $f_N^2/\log N \to \infty$ and $J_*^{-1} f_N^2 \to 0$ as $N \to$

∞, are valid. Now we seek to upper bound $\|U_N(y,v)\|_\sigma(A_\theta)$ to find a sequence d_N and then show that $d_N f_N \to 0$ for any f_N satisfying $f_N^2/\log N \to \infty$ and $J_*^{-1} f_N^2 \to 0$ as $N \to \infty$. For any $y, v \in \Omega_N$, by the definition of $\|\cdot\|_\sigma(A_\theta)$,

$$\|U_N(y,v)\|_\sigma(A_\theta) = \max_{g_i \in A_\theta} |g_i\{U_N(y,v)\}|/\sigma(g_i).$$

Note that by Lemma B.9, we know that $\sigma^2(g_i) = (\sigma_{i+}^2)^{-1} + O\{N_*^{-1}J_*^{-1}\}$ for any $g_i \in A_\theta$. Hence, there exist positive $0 < \gamma_1, \gamma_2 < \infty$ such that for any $i = 1, ..., N$,

$$\gamma_1^{-1} J_*^{-1/2} < \sigma(g_i) < \gamma_2^{-1} J_*^{-1/2}.$$

So we just need to find an upper bound for $|g_i\{U_N(y,v)\}|$ that holds for all $g_i \in A_\theta$. For any $g_i \in A_\theta$, we have

$$\begin{aligned} |g_i\{U_N(y,v)\}| &= \Big| \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} d_{i'j'}(g_i)\{\sigma^2(y_{i'j'}) - \sigma_{i'j'}^2\} v_{i'j'} \Big| \\ &\le \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} |d_{i'j'}(g_i)| \cdot |\sigma^2(y_{i'j'}) - \sigma_{i'j'}^2| \cdot |v_{i'j'}|. \end{aligned}$$

Since $\sigma^2(y_{ij}), \sigma_{ij}^2 < 1$, so $|\sigma^2(y_{ij}) - \sigma_{ij}^2| \le 1$. It can be implied that there exists a positive $\gamma_3 < \infty$ such that $|\sigma^2(y_{ij}) - \sigma_{ij}^2| \le \gamma_3 |g_i(y - M_N^*)|$. From the definition of $\|\cdot\|_\sigma(A_\theta)$, $|g_i(y - M_N^*)| \le \|y - M_N^*\|_\sigma(A_\theta)\sigma(g_i)$ for any $g_i \in A_\theta$. Then it follows that for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|\sigma^2(y_{ij}) - \sigma_{ij}^2| \le \gamma_2^{-1}\gamma_3 J_*^{-1/2} \|y - M_N^*\|_\sigma(A_\theta).$$

Using a similar strategy, we can also show that there exists a positive $\gamma_4 < \infty$ such that for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|v_{ij}| \leq \gamma_2^{-1}\gamma_4 J_*^{-1/2}\|v\|_\sigma(A_\theta).$$

Similarly, we have

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d_{i'j'}(g_i)| \leq \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(g_i)| + \sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(g_i)|.$$

By definition, we know $d'_{i'j'} = (\sigma^2_{i'+})^{-1}w_{i'+} + (\sigma^2_{+j'})^{-1}w_{+j'} - (\sigma^2_{++})^{-1}w_{++}$. For any $g_i \in A_\theta$, $w_{i'+} = 1 - 1/N$, if $i' = i$, and $w_{i'+} = -1/N$ for $i' \neq i$, $w_{+j'} = 0$ for all $j' = 1, ..., J$ and $w_{++} = 0$. Hence,

$$d'_{i'j'}(g_i) = \begin{cases} (1 - \frac{1}{N})(\sigma^2_{i'+})^{-1} & \text{if} \quad i' = i \\ -\frac{1}{N}(\sigma^2_{i'+})^{-1} & \text{if} \quad i' \neq i. \end{cases}$$

It follows

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(g_i)| = \sum_{j'\in S_J(i)}\Big(1 - \frac{1}{N}\Big)(\sigma^2_{i+})^{-1} + \sum_{i'=1,i'\neq i}^{N}\sum_{j'\in S_J(i')}\frac{1}{N}(\sigma^2_{i'+})^{-1} \leq \gamma_5,$$

for some positive $\gamma_5 < \infty$. On the other hand,

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(g_j)| \leq (N^*J^*)^{\frac{1}{2}}\|d''(g_i)\|_\sigma \leq \gamma_6,$$

for some positive $\gamma_6 < \infty$. The last step follows from Condition III.6(c) and Lemma B.9 which implies that $\|d''(g_j)\|_\sigma^2 = O\left(N_*^{-1}J_*^{-1}\right)$. Overall,

$$\begin{aligned}\|U_N(y,v)\|_\sigma(A_\theta) &= \max_{g_i \in A_\theta} |g_i\{U_N(y,v)\}|/\sigma(g_i) \\ &\leq \max_{g_i \in A_\theta} |g_i\{U_N(y,v)\}| \cdot \max_{g_i \in A_\theta}\{\sigma^{-1}(g_i)\} \\ &\leq \gamma_1\gamma_2^{-2}\gamma_3\gamma_4(\gamma_5+\gamma_6)J_*^{-\frac{1}{2}}\|y - M_N^*\|_\sigma(A_\theta)\|v\|_\sigma(A_\theta).\end{aligned}$$

So we can set $d_N = \gamma_1\gamma_2^{-2}\gamma_3\gamma_4(\gamma_5+\gamma_6)J_*^{-\frac{1}{2}}$. Furthermore, by the rate requirement of f_N, for any positive sequence f_N such that $(\log N)^{1/2} \ll f_N \ll J_*^{1/2}$, we must have $d_N f_N \to 0$ as $N \to \infty$. Therefore, both part (a) and part (b) of the lemma are satisfied. □

LEMMA B.9. *Assume Conditions III.6, III.7 and III.8 hold and* $\sum_{i=1}^N \theta_i = 0$*, the asymptotic variance of an individual row parameter,* $var(\hat{\theta}_i)$*, asymptotically attains oracle variance* $(\sigma_{i+}^2)^{-1}$ *in the sense that*

$$var(\hat{\theta}_i) = (\sigma_{i+}^2)^{-1} + O\left(N_*^{-1}J_*^{-1}\right) \qquad as \quad N \to \infty. \tag{B.64}$$

Proof. We seek to construct a linear function $g_i \in \Omega_N^*$ such that $g_i(x) = \theta_i$ so that we can use $\|d'(g_i)\|_\sigma^2$ in Lemma B.13 to approximate $\text{var}(\hat{\theta}_i)$. Fix some $j \in S_J(i)$, i.e. $z_{ij} = 1$, since Condition III.8 holds, we can use the linear function g_j constructed in the proof of Theorem III.10 to represent β_j, i.e. $g_j(x) = \beta_j$. Hence, g_i can easily be

constructed with

$$
\begin{aligned}
g_i(x) &= \frac{1}{|S_J(i)|}\sum_{j\in S_J(i)}\{x_{ij}+g_j(x)\} \\
&= \frac{1}{|S_J(i)|}\sum_{j\in S_J(i)}\Big[x_{ij}-\frac{1}{N}\Big\{\sum_{i'\in S_N(j)}x_{i'j} \\
&+ \sum_{i'\in S_{N_\phi}(j)}\Big(x_{i',j_{i'1}}-x_{i'_{i'1},j_{i'1}}+x_{i'_{i'1},j_{i'2}}-x_{i'_{i'2},j_{i'2}}+\ldots-x_{i'_{i'k},j_{i'k}}+x_{i'_{i'k},j}\Big)\Big\}\Big] \\
&= \theta_i.
\end{aligned}
$$

We use $\|d'(g_i)\|^2_\sigma$ from Lemma B.13 to approximate $\sigma^2(g_i)$, with $w_{i+}=1-N^{-1}, w_{k+}=-N^{-1}$, for all $k=1,...,i-1,i+1,...,N, w_{+j}=0$, for all $j=1,...,J, w_{++}=0$, we obtain

$$
\begin{aligned}
\|d'(g_i)\|^2_\sigma &= \big(1-\frac{1}{N}\big)^2(\sigma^2_{i+})^{-1}+\frac{1}{N^2}\sum_{k=1,k\neq i}^{N}(\sigma^2_{k+})^{-1} \\
&= (\sigma^2_{i+})^{-1}+O\left(N_*^{-1}J_*^{-1}\right) \qquad \text{as} \quad N\to\infty.
\end{aligned}
$$

To see whether $\|d'(g_i)\|^2_\sigma$ is a good approximation for $\sigma^2(g_i)$, we evaluate the order of $\|d''(g_i)\|^2_\sigma$. Note that in this case

$$
\begin{aligned}
&l_p = 0, \qquad p=1,...,N. \\
&v_q = \begin{cases} \frac{1}{N}\sum_{k\in S_N(q),k\neq i}\sigma^2_{kq}(\sigma^2_{k+})^{-1}-(1-\frac{1}{N})\sigma^2_{iq}(\sigma_{i+})^{-1} & \text{if} \quad z_{iq}=1 \\ \frac{1}{N}\sum_{k\in S_N(q)}\sigma^2_{kq}(\sigma^2_{k+})^{-1} & \text{if} \quad z_{iq}=0 \end{cases} \\
&\quad = O\left(J_*^{-1}\right) \qquad \text{as} \quad N\to\infty, \qquad q=1,...,J.
\end{aligned}
$$

It follows that

$$\|d''(g_i)\|_\sigma^2 \leq \gamma_N^{-2}\Big\{\sum_{i=1}^N l_i^2(\sigma_{i+}^2)^{-1} + \sum_{q=1}^J v_q^2(\sigma_{+q}^2)^{-1}\Big\}^2$$
$$= O\left(N_*^{-2}J_*^{-1}\right) \qquad \text{as} \quad N \to \infty,$$

where the last line follows from Condition III.6(a). Since

$$\left(\|d'(g_i)\|_\sigma - \|d''(g_i)\|_\sigma\right)^2 \leq \sigma^2(g_i) \leq \left(\|d'(g_i)\|_\sigma + \|d''(g_i)\|_\sigma\right)^2,$$

it follows that $\text{var}(\hat{\theta}_i) = (\sigma_{i+}^2)^{-1} + O(N_*^{-1}J_*^{-1})$ as $N \to \infty$. □

LEMMA B.10. *Assume Conditions III.6, III.7 and III.8 hold and* $\sum_{i=1}^N \theta_i = 0$. *Consider a linear function* $g : \Omega_N \mapsto \mathbb{R}$ *with* $g(M) = \sum_{i=1}^N h_i\theta_i + \sum_{j=1}^J h'_j\beta_j$. *If there exists a positive* $C < \infty$ *such that* $\sum_{i=1}^N |h_i| < C$ *and* $\sum_{j=1}^J |h'_j| < C$, *then*

$$\sigma^2(g) = \sum_{i=1}^N h_i^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^J h_j'^2(\sigma_{+j}^2)^{-1} + O\left(N_*^{-1}J_*^{-1}\right) \quad \text{as} \quad N \to \infty.$$

Proof. By Proposition III.9, we can reexpress function g in terms of m_{ij} for $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$ with $g(M_N) = \sum_{i=1}^N \sum_{j \in S_J(i)} w_{ij}(g)m_{ij}$. In particular, we

have,

$$w_{i+}(g) = h_i\big(1 - \frac{1}{N}\big) - \frac{1}{N}\sum_{i'=1,i'\neq i}^{N} h_{i'} - \frac{1}{N}\sum_{j=1}^{J} h'_j \qquad i = 1, ..., N,$$

$$w_{+j}(g) = -h'_j, \qquad j = 1, ..., J,$$

$$w_{++}(g) = -\sum_{j=1}^{J} h'_j.$$

We apply $\|d'(g)\|_\sigma^2$ from Lemma B.13 to approximate $\sigma^2(g)$. Note that

$$\begin{aligned}\|d'(g)\|_\sigma^2 =& \sum_{i=1}^{N} w_{i+}^2(g)(\sigma_{i+}^2)^{-1} + \sum_{j=1}^{J} w_{+j}^2(g)(\sigma_{+j}^2)^{-1} \\ &+ 2\sum_{i=1}^{N}\sum_{j\in S_j(i)} \sigma_{ij}^2(\sigma_{i+}^2)^{-1} w_{i+}(g)(\sigma_{+j}^2)^{-1} w_{+j}(g) - 3(\sigma_{++}^2)^{-1} w_{++}^2(g) \\ &= \sum_{i=1}^{N} h_i^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^{J} h_j'^2(\sigma_{+j}^2)^{-1} + O\left(N_*^{-1}J_*^{-1}\right) \quad \text{as} \quad N \to \infty,\end{aligned}$$

where the last step follows from the assumption that $\sum_{i=1}^{N} |h_i| < C$ and $\sum_{j=1}^{J} |h'_j| < C$. To see whether $\|d'(g)\|_\sigma^2$ is a good approximation for $\sigma^2(g)$, we need to evaluate the order of $\|d''(g)\|_\sigma^2$. Note that for $i = 1, ..., N$,

$$\begin{aligned}l_i =& -\sum_{j\in S_J(i)} \sigma_{ij}^2(\sigma_{+j}^2)^{-1} w_{+j}(g) + \sigma_{i+}^2(\sigma_{++}^2)^{-1} w_{++}(g) \\ =& \sum_{j\in S_J(i)} \sigma_{ij}^2(\sigma_{+j}^2)^{-1} h'_j - \sigma_{i+}^2(\sigma_{++}^2)^{-1}\sum_{j=1}^{J} h'_j = O\left(N_*^{-1}\right) \quad \text{as} \quad N \to \infty, \qquad \text{(B.65)}\end{aligned}$$

where the last step follows from $\sum_{j=1}^{J}|h'_j| < C$. Similarly for $j = 1, ..., J$,

$$\begin{aligned} v_j =& -\sum_{i\in S_N(j)} \sigma_{ij}^2(\sigma_{i+}^2)^{-1}w_{i+}(g) + \sigma_{+j}^2(\sigma_{++}^2)^{-1}w_{++}(g) \\ =& -\sum_{i\in S_N(j)} \sigma_{ij}^2(\sigma_{i+}^2)^{-1}\Big\{h_i\big(1-\frac{1}{N}\big) - \frac{1}{N}\sum_{i'=1,i'\neq i}^{N} h_{i'} - \frac{1}{N}\sum_{j=1}^{J} h'_j\Big\} \\ & - \sigma_{+j}^2(\sigma_{++}^2)^{-1}\sum_{j=1}^{J} h'_j \\ =& O\left(J_*^{-1}\right) \quad \text{as} \quad N\to\infty, \end{aligned} \tag{B.66}$$

where the last step follows from $\sum_{j=1}^{J}|h'_j| < C$ and $\sum_{i=1}^{N}|h_i| < C$. Hence, we have

$$\begin{aligned} \|d''(g)\|_\sigma^2 &\le \gamma_N^{-2}\Big\{\sum_{i=1}^{N} l_i^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^{J} v_j^2(\sigma_{+j}^2)^{-1}\Big\}^2 \\ &= O\left(N_*^{-1}J_*^{-2}\right) \quad \text{as} \quad N\to\infty, \end{aligned}$$

where the last step can be seen easily from (B.65), (B.66) and Condition III.6(a) that there exists a constant $k > 0$, such that $N_* \ge kN^{2/3}$ and $J_* \ge kJ^{2/3}$. It follows that

$$\sigma^2(g) = \sum_{i=1}^{N} h_i^2(\sigma_{i+}^2)^{-1} + \sum_{j=1}^{J} h_j'^2(\sigma_{+j}^2)^{-1} + O\left(N_*^{-1}J_*^{-1}\right) \qquad \text{as} \quad N\to\infty.$$

Hence, the result of the lemma follows. □

LEMMA B.11. *Assume Conditions III.6, III.7 and III.8 hold and* $J_*^{-2}N_*(\log N)^2 \to 0$ *as* $N\to\infty$. *If* $A_{\theta,\beta} = \{g_i, g'_j : i = 1, ..., N, j = 1, ..., J\}$ *such that* $g_i, g'_j \in \Omega_N^*$, *and*

$g_i(x) = \theta_i$ and $g_j'(x) = \beta_j$ for $x \in \Omega_N$. Let $C_N = |A_{\theta,\beta}|$, the cardinality of $A_{\theta,\beta}$. Then there exist sequences $f_N > 0$ and $d_N \geq 0$ satisfying the followings.

(a). As $N \to \infty$, $f_N^2 / \log C_N \to \infty$.

(b). If $y, v \in \Omega_N$ and $\|y - M_N^*\|_\sigma(A_{\theta,\beta}) \leq f_N$, then there exists $n < \infty$ such that for all $N > n$, $\|U_N(y,v)\|_\sigma(A_{\theta,\beta}) \leq d_N \|y - M_N^*\|_\sigma(A_{\theta,\beta}) \|v\|_\sigma(A_{\theta,\beta})$. Furthermore, $d_N f_N^2 \to 0$ as $N \to \infty$.

Proof. Since $J_*^{-2} N_* (\log N)^2 \to 0$ as $N \to \infty$, there must exist a positive sequence L_N such that $L_N \to \infty$ but $J_*^{-1} N_*^{1/2} (\log N) L_N \to 0$ as $N \to \infty$. Furthermore, note that

$$\log(C_N) = \log(N + J) \leq \log(2N) = \log(2) + \log(N) = O(\log(N)) \quad \text{as} \quad N \to \infty.$$

Let $f_N^2 = \{\log(N)\} L_N$. It is easy to see that the constructed f_N satisfies part (a) of the lemma.

Now we consider part (b). We seek to find an upper bound for $\|U_N(y,z)\|_\sigma(A_{\theta,\beta})$ in order to find d_N and then show that $d_N f_N^2 \to 0$ as $N \to \infty$. For any $y, v \in \Omega_N$, by the definition of $\|\cdot\|_\sigma(A_{\theta,\beta})$,

$$\|U_N(y,v)\|_\sigma(A_{\theta,\beta}) = \max_{f \in A_{\theta,\beta}} |f\{U_N(y,v)\}| / \sigma(f).$$

First note from (B.63) and (B.64), we know that for any $f \in A_{\theta,\beta}$, there exist

$0 < c_1, c_2 < \infty$ such that for all $N > n$,

$$c_1^{-1} N_*^{-\frac{1}{2}} < \sigma(f) < c_2^{-1} J_*^{-\frac{1}{2}}.$$

So we just need to find an upper bound for $|f\{U_N(y,v)\}|$ that holds for all $f \in A_{\theta,\beta}$. Note that

$$\begin{aligned} |f\{U_N(y,v)\}| &= \Big| \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} d_{i'j'}(f)\{\sigma^2(y_{i'j'}) - \sigma^2_{i'j'}\} v_{i'j'} \Big| \\ &\leq \sum_{i'=1}^{N} \sum_{j' \in S_J(i')} |d_{i'j'}(f)| \cdot |\sigma^2(y_{i'j'}) - \sigma^2_{i'j'}| \cdot |v_{i'j'}|. \end{aligned} \tag{B.67}$$

Note $0 \leq \sigma^2(y_{ij}), \sigma^2_{ij} \leq 1$, so $|\sigma^2(y_{ij}) - \sigma^2_{ij}| \leq 1$. It can be implied that $|\sigma^2(y_{ij}) - \sigma^2_{ij}| \leq c_3 |f(y - M_N^*)|$ for some positive $c_3 < \infty$. By the definition of $\|\cdot\|_\sigma(A_{\theta,\beta})$, we have $|f(y - M_N^*)| \leq \|y - M_N^*\|_\sigma(A_{\theta,\beta})\sigma(f)$. Hence, it follows that for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|\sigma^2(y_{ij}) - \sigma^2_{ij}| \leq c_2^{-1} c_3 J_*^{-1/2} \|y - M_N^*\|_\sigma(A_{\theta,\beta}).$$

Using a similar strategy, we can show that there exists a positive $c_4 < \infty$ such that for any $i = 1, ..., N, j = 1, ..., J, z_{ij} = 1$,

$$|v_{ij}| \leq c_2^{-1} c_4 J_*^{-1/2} \|v\|_\sigma(A_{\theta,\beta}).$$

Further, note also that

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d_{i'j'}(f)|\leq\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(f)|+\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(f)|.$$

By definition, $d'_{i'j'}=(\sigma^2_{i'+})^{-1}w_{i'+}+(\sigma^2_{+j'})^{-1}w_{+j'}-(\sigma^2_{++})^{-1}w_{++}$. For any $f\in A_{\theta,\beta}$, either $f=g'_j$ or $f=g_i$. When $f=g'_j$, $w_{i'+}=-1/N$, for $i'=1,...,N$, $w_{+j'}=-1$ if $j'=j$ and $w_{+j'}=0$ if $j'\neq j$, $w_{++}=-1$. Hence,

$$d'_{i'j'}(g'_j)=\begin{cases}-\frac{1}{N}(\sigma^2_{i'+})^{-1}-(\sigma^2_{+j'})^{-1}+(\sigma^2_{++})^{-1} & \text{if}\quad j'=j\\ -\frac{1}{N}(\sigma^2_{i'+})^{-1}+(\sigma^2_{++})^{-1} & \text{if}\quad j'\neq j\end{cases}$$

It follows

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d'_{i'j'}(g'_j)|\leq\frac{J^*}{N}\sum_{i'=1}^{N}(\sigma^2_{i'+})^{-1}+N^*(\sigma^2_{+j})^{-1}+\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}(\sigma^2_{++})^{-1}\leq c_5,$$

for some positive $c_5<\infty$. Furthermore,

$$\sum_{i'=1}^{N}\sum_{j'\in S_J(i')}|d''_{i'j'}(g'_j)|\leq(N^*J^*)^{\frac{1}{2}}\|d''(g'_j)\|_\sigma\leq c_6,$$

for some positive $c_6<\infty$. The last step follows from Condition III.6(c) and (B.63) which implies that $\|d''(g'_j)\|^2_\sigma=O\left(N_*^{-1}J_*^{-1}\right)$. On the other hand, when $f=g_i$, we have $w_{i'+}=1-1/N$, if $i'=i$, and $w_{i'+}=-1/N$ for $i'\neq i$, $w_{+j'}=0$ for all $j'=1,...,J$

and $w_{++} = 0$. Hence,

$$d'_{i'j'}(g_i) = \begin{cases} (1 - \frac{1}{N})(\sigma^2_{i'+})^{-1} & \text{if} \quad i' = i \\ -\frac{1}{N}(\sigma^2_{i'+})^{-1} & \text{if} \quad i' \neq i. \end{cases}$$

It follows

$$\sum_{i'=1}^{N} \sum_{j' \in S_J(i')} |d'_{i'j'}(g_i)| = \sum_{j' \in S_J(i)} \Big(1 - \frac{1}{N}\Big)(\sigma^2_{i+})^{-1} - \sum_{i'=1, i' \neq i}^{N} \sum_{j' \in S_J(i')} \frac{1}{N}(\sigma^2_{i'+})^{-1} \leq c_7,$$

for some positive $c_7 < \infty$. Furthermore,

$$\sum_{i'=1}^{N} \sum_{j' \in S_J(i')} |d''_{i'j'}(g_i)| \leq (N^* J^*)^{\frac{1}{2}} \|d''(g_i)\|_\sigma \leq c_8,$$

for some positive $c_8 < \infty$. The last step follows from Condition III.6(c) and (B.64) which implies that $\|d''(g_i)\|^2_\sigma = O\left(N_*^{-1} J_*^{-1}\right)$. Overall,

$$\begin{aligned} \|U_N(y,v)\|_\sigma(A_{\theta,\beta}) &= \max_{f \in A_{\theta,\beta}} |f\{U_N(y,v)\}|/\sigma(f) \\ &\leq \max_{f \in A_{\theta,\beta}} |f\{U_N(y,v)\}| \max_{f \in A_{\theta,\beta}} \{\sigma(f)^{-1}\} \\ &\leq c_1 c_2^{-2} c_3 c_4 \max\{c_5 + c_6, c_7 + c_8\} J_*^{-1} N_*^{\frac{1}{2}} \|y - M_N^*\|_\sigma(A_{\theta,\beta}) \|v\|_\sigma(A_{\theta,\beta}). \end{aligned}$$

Note that in this case we can take $d_N = c_1 c_2^{-2} c_3 c_4 \max\{c_5 + c_6, c_7 + c_8\} J_*^{-1} N_*^{1/2}$. We have

$$d_N f_N^2 = c_1 c_2^{-2} c_3 c_4 \max\{c_5 + c_6, c_7 + c_8\} J_*^{-1} N_*^{1/2} \log(N) L_N \to 0 \quad \text{as} \quad N \to \infty.$$

Hence both parts (a) and (b) of the lemma are satisfied. □

B.3 Senator Rankings

With the same set-up as in Section 3.5.2 of Chapter III, we present a full list of rankings for senators serving the 111th, the 112th and the 113th United States senate according to their estimated conservativeness scores. The results are summarized in Tables B.1 and B.2 below. We observe from Table B.1 that all the top 62 most conservative senators predicted by the model are Republicans. While the Democrats and the independent politicians are predicted to have much lower conservativeness scores as presented in Table B.2. This aligns well with the public perceptions about the Republican party and the Democratic party. Standard errors of the estimated row parameters (i.e. senators' conservativeness scores) are also included to facilitate inferences.

Table B.1: Ranking of the top 62 most conservative senators predicted by the model. Rep represents the Republican party and states are listed in their standard abbreviations. $\hat{\theta}$ represents the conservativeness score of senators and s.e.($\hat{\theta}$) is the standard error of the estimated conservativeness score.

Rank	Senator	State	Party	$\hat{\theta}$	s.e.($\hat{\theta}$)	Rank	Senator	State	Party	$\hat{\theta}$	s.e.($\hat{\theta}$)
1	Demint	SC	Rep	5.87	0.157	2	Lee	UT	Rep	5.73	0.138
3	Cruz	TX	Rep	5.65	0.195	4	Coburn	OK	Rep	5.25	0.114
5	Paul	KY	Rep	5.24	0.129	6	Scott	SC	Rep	5.17	0.176
7	Bunning	KY	Rep	4.92	0.204	8	Johnson	WI	Rep	4.84	0.119
9	Risch	ID	Rep	4.81	0.102	10	Inhofe	OK	Rep	4.69	0.103
11	Crapo	ID	Rep	4.56	0.097	12	Sessions	AL	Rep	4.48	0.096
13	Enzi	WY	Rep	4.36	0.094	14	Barasso	WY	Rep	4.35	0.094
15	Cornyn	TX	Rep	4.33	0.095	16	Rubio	FL	Rep	4.25	0.112
17	Ensign	NV	Rep	4.24	0.166	18	Vitter	LA	Rep	4.20	0.094
19	Fischer	NE	Rep	4.14	0.145	20	Toomey	PA	Rep	4.12	0.109
21	Kyl	AZ	Rep	4.10	0.115	22	Roberts	KS	Rep	4.06	0.091
23	Mcconnell	KY	Rep	4.02	0.089	24	Thune	SD	Rep	3.95	0.088
25	Burr	NC	Rep	3.95	0.090	26	Moran	KS	Rep	3.89	0.109
27	Grassley	IA	Rep	3.80	0.086	28	Shelby	AL	Rep	3.78	0.086
29	Boozman	AR	Rep	3.68	0.105	30	Chambliss	GA	Rep	3.65	0.087
31	Mccain	AZ	Rep	3.65	0.086	32	Brownback	KS	Rep	3.61	0.153
33	Coats	IN	Rep	3.51	0.101	34	Johanns	NE	Rep	3.39	0.082
35	Isakson	GA	Rep	3.38	0.082	36	Hatch	UT	Rep	3.38	0.083
37	Lemieux	FL	Rep	3.34	0.188	38	Blunt	MO	Rep	3.31	0.099
39	Wicker	MS	Rep	3.29	0.080	40	Portman	OH	Rep	3.28	0.098
41	Corker	TN	Rep	3.27	0.080	42	Heller	NV	Rep	3.26	0.100
43	Hutchison	TX	Rep	3.25	0.105	44	Graham	SC	Rep	3.18	0.080
45	Flake	AZ	Rep	3.03	0.125	46	Ayotte	NH	Rep	3.02	0.095
47	Hoeven	ND	Rep	2.97	0.094	48	Bennett	UT	Rep	2.74	0.127
49	Alexander	TN	Rep	2.71	0.075	50	Kirk	IL	Rep	2.67	0.105
51	Cochran	MS	Rep	2.63	0.075	52	Chiesa	NJ	Rep	2.61	0.343
53	Gregg	NH	Rep	2.59	0.127	54	Martinez	FL	Rep	2.47	0.186
55	Lugar	IN	Rep	2.29	0.088	56	Bond	MO	Rep	2.25	0.118
57	Murkowski	AK	Rep	1.47	0.066	58	Brown	MA	Rep	1.29	0.103
59	Voinovich	OH	Rep	1.22	0.102	60	Snowe	ME	Rep	1.06	0.080
61	Specter	PA	Rep	1.03	0.192	62	Collins	ME	Rep	0.82	0.064

Table B.2: Ranking of the top 63-139 most conservative senators predicted by the model. Dem and Ind represent the Democratic party and independent politician, respectively. States are presented in their standard abbreviations. $\hat{\theta}$ represents the conservativeness score of senators and s.e.($\hat{\theta}$) is the standard error of the estimated conservativeness score.

Rank	Senator	State	Party	$\hat{\theta}$	s.e.($\hat{\theta}$)	Rank	Senator	State	Party	$\hat{\theta}$	s.e.($\hat{\theta}$)
63	Nelson	NE	Dem	-0.05	0.084	64	Bayh	IN	Dem	-0.13	0.104
65	Manchin	WV	Dem	-0.66	0.099	66	Feingold	WI	Dem	-0.92	0.115
67	Lincoln	AR	Dem	-0.96	0.119	68	Mccaskill	MO	Dem	-1.15	0.083
69	Webb	VA	Dem	-1.49	0.108	70	Pryor	AR	Dem	-1.63	0.094
71	Lieberman	CT	Dem	-1.68	0.113	72	Heitkamp	ND	Dem	-1.87	0.183
73	Donnelly	IN	Dem	-1.87	0.182	74	Hagan	NC	Dem	-1.90	0.100
75	Byrd	WV	Dem	-2.00	0.217	76	Warner	VA	Dem	-2.06	0.105
77	Landrieu	LA	Dem	-2.07	0.106	78	Tester	MT	Dem	-2.11	0.105
79	Baucus	MT	Dem	-2.11	0.112	80	Bennet	CO	Dem	-2.16	0.107
81	Klobuchar	MN	Dem	-2.26	0.109	82	Conrad	ND	Dem	-2.29	0.131
83	King	ME	Ind	-2.30	0.208	84	Nelson	FL	Dem	-2.32	0.112
85	Kohl	WI	Dem	-2.34	0.131	86	Carper	DE	Dem	-2.36	0.112
87	Udall	CO	Dem	-2.39	0.113	88	Begich	AK	Dem	-2.43	0.116
89	Dorgan	ND	Dem	-2.44	0.167	90	Reid	NV	Dem	-2.68	0.122
91	Shaheen	NH	Dem	-2.76	0.125	92	Kaine	VA	Dem	-2.80	0.246
93	Casey	PA	Dem	-2.83	0.127	94	Cantwell	WA	Dem	-2.84	0.127
95	Coons	DE	Dem	-2.84	0.170	96	Specter	PA	Dem	-2.84	0.222
97	Walsh	MT	Dem	-2.85	0.395	98	Wyden	OR	Dem	-2.97	0.132
99	Bingaman	NM	Dem	-3.03	0.155	100	Johnson	SD	Dem	-3.09	0.137
101	Stabenow	MI	Dem	-3.11	0.137	102	Cowan	MA	Dem	-3.19	0.439
103	Merkley	OR	Dem	-3.19	0.140	104	Sanders	VT	Ind	-3.23	0.143
105	Feinstein	CA	Dem	-3.24	0.143	106	Kerry	MA	Dem	-3.25	0.165
107	Kaufman	DE	Dem	-3.28	0.219	108	Murray	WA	Dem	-3.29	0.143
109	Heinrich	NM	Dem	-3.30	0.290	110	Menendez	NJ	Dem	-3.32	0.144
111	Inouye	HI	Dem	-3.33	0.169	112	Boxer	CA	Dem	-3.35	0.148
113	Dodd	CT	Dem	-3.38	0.218	114	Warren	MA	Dem	-3.45	0.307
115	Levin	MI	Dem	-3.52	0.152	116	Blumenthal	CT	Dem	-3.52	0.214
117	Kirk	MA	Dem	-3.54	0.716	118	Akaka	HI	Dem	-3.54	0.174
119	Franken	MN	Dem	-3.55	0.166	120	Rockefeller	WV	Dem	-3.56	0.161
121	Mikulski	MD	Dem	-3.60	0.158	122	Leahy	VT	Dem	-3.63	0.158
123	Harkin	IA	Dem	-3.64	0.158	124	Lautenberg	NJ	Dem	-3.65	0.179
125	Schumer	NY	Dem	-3.65	0.159	126	Reed	RI	Dem	-3.67	0.157
127	Gillibrand	NY	Dem	-3.67	0.158	128	Murphy	CT	Dem	-3.68	0.327
129	Markey	MA	Dem	-3.73	0.465	130	Whitehouse	RI	Dem	-3.74	0.163
131	Cardin	MD	Dem	-3.82	0.163	132	Durbin	IL	Dem	-3.83	0.164
133	Udall	NM	Dem	-3.85	0.165	134	Brown	OH	Dem	-3.89	0.168
135	Baldwin	WI	Dem	-3.90	0.352	136	Booker	NJ	Dem	-4.14	0.572
137	Hirono	HI	Dem	-4.17	0.383	138	Burris	IL	Dem	-4.43	0.297
139	Schatz	HI	Dem	-4.74	0.468						

APPENDIX C

Appendix of Chapter 4

Appendix C contains proofs of all the propositions and theorems developed in Chapter IV, additional results for simulation studies and real application to PISA 2018 data, and discussion of the asymptotic distribution of $\tilde{\Xi}$ and the implementation details of Algorithms 3 and 4 in Chapter IV.

C.1 Proofs of Propositions and Theorems

Proof of Proposition IV.1. Note h is differentiable for all $c \neq 0$ with,

$$\nabla h(c) = \sum_{j=1}^{J} |a_j| \cdot sign(a_j^* c - \gamma_j^*), \qquad c \neq 0.$$

Further note that $sign(a_j^* c - \gamma_j^*) = 0$ when $c = \gamma_j^*/a_j^*$. Hence we have

$$sign(a_j^* c - \gamma_j^*) > 0 \quad \text{whenever} \quad c > \frac{\gamma_j^*}{a_j^*}, \tag{C.1}$$

$$sign(a_j^* c - \gamma_j^*) < 0 \quad \text{whenever} \quad c < \frac{\gamma_j^*}{a_j^*}. \tag{C.2}$$

Consider the right derivative (positive directional derivative) of h at 0 from $+1$ direction,

$$\partial h^+(0) := \lim_{c \downarrow 0} \frac{h(c) - h(0)}{c}.$$

By the definition of right derivative of h at 0, (C.1) and (C.2), we can rewrite $\partial h^+(0)$ equivalently as follows,

$$\partial h^+(0) = \sum_{j=1}^{J} |a_j^*| \left(-I\big(\frac{\gamma_j^*}{a_j^*} > 0\big) + I\big(\frac{\gamma_j^*}{a_j^*} \leq 0\big) \right). \tag{C.3}$$

Similarly, define the left derivative (negative directional derivative) of h at 0 from -1 direction,

$$\partial h^-(0) := \lim_{c \uparrow 0} \frac{h(c) - h(0)}{c}.$$

By the definition of left derivative $\partial h^-(0)$, (C.1) and (C.2), we can rewrite $\partial h^-(0)$ equivalently as follows,

$$\partial h^-(0) = \sum_{j=1}^{J} |a_j^*| \left(-I\big(\frac{\gamma_j^*}{a_j^*} \geq 0\big) + I\big(\frac{\gamma_j^*}{a_j^*} < 0\big) \right). \tag{C.4}$$

Since h is convex, we must have $\arg\min_c h(c) = 0$ if and only if $\partial h^+(0) > 0$ and $\partial h^-(0) < 0$ (*Boyd and Vandenberghe*, 2004; *Shor*, 2012). From (C.3), (C.4) and the fact that ML1 Condition (4.5) is equivalent to $\arg\min_c h(c) = 0$, the result of the proposition follows directly. □

Proof of Theorem IV.3. Since the model with constraint $\gamma_1^\dagger = 0$ is identifiable, by classical asymptotic theory for MLE (*van der Vaart*, 2000), we have $\tilde{\Xi}$ converges in probability to $\Xi^\dagger$. That is, as $N \to \infty$, for any $\epsilon > 0$, we must have with probability tending to 1 that $|\tilde{\mu} - \mu^\dagger| \leq \epsilon$, $|\tilde{\gamma}_j - \gamma_j^\dagger| \leq \epsilon, |\tilde{a}_j - a_j^\dagger| \leq \epsilon$ and $|\tilde{d}_j - d_j^\dagger| \leq \epsilon$, for any $j = 1, ..., J$. Denote $f(c) = \sum_{j=1}^J |\gamma_j^\dagger - ca_j^\dagger|$ as a function of c. Similarly, denote $f_N(c) = \sum_{j=1}^J |\tilde{\gamma}_j - c\tilde{a}_j|$. Let $c^\dagger = \arg\min_c f(c)$ and $\hat{c} = \arg\min_c f_N(c)$, respectively. We seek to establish that $\hat{c}$ will converge in probability to $c^\dagger$ as $N \to \infty$. First note that by regularity conditions, there exists $C_1 < \infty$ such that $J, |\gamma_j^\dagger|, |a_j^\dagger| \leq C_1$. Then, there must exist $C_2 < \infty$ such that $|c^\dagger| \leq C_2$. Furthermore, note f_N is clearly continuous and convex in c, so consistency will follow if f_N can be shown to converge point-wise to f that is uniquely minimized at the true value $c^\dagger$ (typically uniform convergence is needed, but point-wise convergence of convex functions implies their uniform convergence on compact subsets). Following the model identifiability and the ML1 condition (4.5), $c^\dagger$ is unique. To see this, suppose for contradiction that there exist c_1 and c_2 such that $c_1 \neq c_2$ and $c_1 = \arg\min_c f(c)$ and $c_2 = \arg\min_c f(c)$. First note that $a_j^\dagger = a_j^*$ for all $j = 1, ..., J$. Then by model identifiability, there exists

c_3 such that $\gamma_j^\dagger = \gamma_j^* + c_3 a_j^*$. So we have

$$c_1 = \arg\min_c \sum_{j=1}^{J} |\gamma_j^* + (c_3 - c)a_j^*|$$

and

$$c_2 = \arg\min_c \sum_{j=1}^{J} |\gamma_j^* + (c_3 - c)a_j^*|.$$

Hence, $\gamma^* = \gamma^\dagger + (c_3 - c_1)a_j^*$ and $\gamma^* = \gamma^\dagger + (c_3 - c_2)a_j^*$. If ML1 condition (4.5) holds, then $c_3 = c_1$ and $c_3 = c_2$. This contradicts the assumption $c_1 \neq c_2$. Hence, $c^\dagger$ must be unique. For any $|c| \leq C_2$,

$$\begin{aligned}
&|f_N(c) - f(c)| \\
&= \Big| \sum_{j=1}^{J} \Big(|\tilde{\gamma}_j - c\tilde{a}_j| - |\gamma_j^\dagger - ca_j^\dagger| \Big) \Big| \\
&\leq \Big| \sum_{j=1}^{J} \Big(|(\tilde{\gamma}_j - c\tilde{a}_j) - (\gamma_j^\dagger - ca_j^\dagger)| \Big) \Big| \\
&= \Big| \sum_{j=1}^{J} \Big(|(\tilde{\gamma}_j - \gamma_j^\dagger) + c(a_j^\dagger - \tilde{a}_j)| \Big) \Big| \\
&\leq \sum_{j=1}^{J} \Big(|\tilde{\gamma}_j - \gamma_j^\dagger| + |c| \cdot |a_j^\dagger - \tilde{a}_j| \Big) \\
&\leq J\epsilon + |c|\epsilon. \\
&\leq (C_1 + C_2)\epsilon.
\end{aligned}$$

Take $\epsilon_1 = (C_1+C_2)\epsilon$, it follows that for any fixed $|c| \leq C_2$, $P\big(|f_N(c)-f(c)| \leq \epsilon_1\big) \to 1$ as $N \to \infty$. Moreover, following from the uniqueness of $c^\dagger$ and the continuity and the convexity of $f_N(\cdot)$ in c, we must have $|\hat{c} - c^\dagger| = o_P(1)$ as $N \to \infty$.

Note that $\hat{\mu} = \tilde{\mu} + \hat{c}$, $\hat{\gamma}_j = \tilde{\gamma}_j - \hat{c}\tilde{a}_j$, $\hat{a}_j = \tilde{a}_j$, $\hat{d}_j = \tilde{d}_j$ for all $j = 1, ..., J$. From the model identifiability and the ML1 condition (4.5), we know that $\mu^* = \mu^\dagger + c^\dagger$, $\gamma_j^* = \gamma_j^\dagger - c^\dagger a_j^\dagger$, $a_j^* = a_j^\dagger$, $d_j^* = d_j^\dagger$ for all $j = 1, ..., J$. Since $|\hat{c} - c^\dagger| = o_P(1), |\tilde{\mu} - \mu^\dagger| = o_P(1)$, $|\tilde{\gamma}_j - \gamma_j^\dagger| = o_P(1), |\tilde{a}_j - a_j^\dagger| = o_P(1), |\tilde{d}_j - d_j^\dagger| = o_P(1)$ as $N \to \infty$, it follows directly from the Slutsky's Theorem that $|\hat{\mu} - \mu^*| = o_P(1), |\hat{\gamma}_j - \gamma_j^*| = o_P(1), |\hat{a}_j - a_j^*| = o_P(1)$, $|\hat{d}_j - d_j^*| = o_P(1)$ as $N \to \infty$. □

Proof of Theorem IV.4. Following from the results of Theorem IV.3, we have $|\hat{\gamma}_j - \gamma_j^*| = o_P(1)$ as $N \to \infty$ for all $j = 1, ..., J$, i.e., for any $\epsilon > 0$, $\lim_{N\to\infty} P(|\hat{\gamma}_j - \gamma_j^*| \leq \epsilon) = 1$. First note that $|\hat{\gamma}_j - \gamma_j^*| = o_P(1)$ implies that $\big||\hat{\gamma}_j| - |\gamma_j^*|\big| = o_P(1)$ as $N \to \infty$ for all $j = 1, ..., J$. That is, for any $\epsilon_1 > 0$, $\lim_{N\to\infty} P(\big||\hat{\gamma}_j| - |\gamma_j^*|\big| \leq \epsilon_1) = 1$. When $\gamma_j^* \neq 0$, let $\epsilon_{2j} := |\gamma_j^*| - \delta$. By the assumption that $0 < \delta < \min\{|\gamma_j^*| : \gamma_j^* \neq 0\}$, we know that $\epsilon_{2j} > 0$. Take $\epsilon_1 = 0.5\epsilon_{2j}$. Then,

$$\begin{aligned} |\hat{\gamma}_j| - \delta &\geq |\gamma_j^*| - \epsilon_1 - \delta \\ &= \epsilon_{2j} - \epsilon_1 \\ &= \frac{1}{2}\epsilon_{2j}. \end{aligned}$$

Hence, we have

$$\lim_{N\to\infty} P(|\hat{\gamma}_j| - \delta \geq \frac{1}{2}\epsilon_{2j}) = 1,$$

which then implies that

$$\lim_{N\to\infty} P(|\hat{\gamma}_j| - \delta > 0) = 1.$$

Therefore, we have,

$$\lim_{N\to\infty} P(1_{\{|\hat{\gamma}_j|\le\delta\}} = 1_{\{\gamma_j^*=0\}}) = 1.$$

□

C.2 Asymptotic Distribution of $\tilde{\Xi}$

Since the model is identifiable with constraint $\gamma_1^\dagger = 0$ and all the regularity conditions in Theorem 5.39 of *van der Vaart* (2000) are satisfied, hence, by Theorem 5.39 in *van der Vaart* (2000), $\tilde{\Xi} \to \mathrm{N}(\Xi^\dagger, \Sigma^*)$ in distribution as $N \to \infty$. In practice, we use the inverse of the observed Fisher information matrix, denoted by $\hat{\Sigma}_N$, which is a consistent estimator of Σ^*, to draw Monte Carlo samples. Below, we give procedures to evaluate $\hat{\Sigma}_N$ from the marginal log-likelihood.

Following the notations in Chapter IV, we first provide the complete data log-likelihood function,

$$\begin{aligned} l(\Xi; Y) = \sum_{i=1}^{N} \Big[& \log\Big\{ \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(\theta_i - \mu x_i)^2}{2} \right) \Big\} \\ & + \sum_{j=1}^{J} \big\{ y_{ij}(a_j\theta_i + d_j + \gamma_j x_i) - \log(1 + \exp\{a_j\theta_i + d_j + \gamma_j x_i\}) \big\} \Big]. \end{aligned}$$

Since the latent θ_i is considered as a random variable such that $\theta_i \mid x_i \sim \mathrm{N}(\mu x_i, 1)$,

we will work with the marginal log-likelihood function,

$$mll(\Xi;Y)=\sum_{i=1}^{N}\log\Big\{\int\left(\prod_{j=1}^{J}\frac{\exp(y_{ij}(a_j\theta_i+d_j+\gamma_j x_i))}{1+\exp(a_j\theta_i+d_j+\gamma_j x_i)}\right)\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(\theta_i-\mu x_i)^2}{2}\right)d\theta_i\Big\}.$$

Note that the observed Fisher information matrix $I(\Xi)$ cannot be directly obtained from the $mll(\Xi;Y)$ due to the intractable integral. Instead, we apply the Louis Identity (*Louis*, 1982) to evaluate the observed Fisher information matrix. Let $S(\Xi;Y)$ and $B(\Xi;Y)$ denote the gradient vector and the negative of the hessian matrix of the complete data log-likelihood function, respectively. Then by the Louis Identity, $I(\Xi)$ can be expressed as

$$I(\Xi)=\mathbf{E}_\theta[B(\Xi;Y)\mid Y]-\mathbf{E}_\theta[S(\Xi;Y)S(\Xi;Y)^T\mid Y]+\mathbf{E}_\theta[S(\Xi;Y)\mid Y]\mathbf{E}_\theta[S(\Xi;Y)\mid Y]^T.$$

Denote $p_{ij}=\exp\{y_{ij}(a_j\theta_i+d_j+\gamma_j x_i)\}/[1+\exp\{y_{ij}(a_j\theta_i+d_j+\gamma_j x_i)\}]$. Then, in particular,

$$\begin{aligned}S(\Xi;Y)=&\frac{\partial l(\Xi;Y)}{\partial\Xi}\\=&\Big\{\frac{\partial l(\Xi;Y)}{\partial\mu},...,\frac{\partial l(\Xi;Y)}{\partial a_j},...,\frac{\partial l(\Xi;Y)}{\partial d_j},...,\frac{\partial l(\Xi;Y)}{\partial\gamma_j},...\Big\}\\=&\Big\{\sum_{i=1}^{N}x_i(\theta_i-\mu),...,\sum_{i=1}^{N}\theta_i(y_{ij}-p_{ij}),...,\sum_{i=1}^{N}(p_{ij}-y_{ij}),...,\sum_{i=1}^{N}x_i(y_{ij}-p_{ij})\Big\}.\end{aligned}$$

Furthermore, note that $B(\Xi;Y)=-\partial^2 l(\Xi;Y)/\partial\Xi\partial\Xi^T$ is a (3J+1) by (3J +1) matrix

with the only non-zero entries,

$$\begin{aligned}
\frac{\partial^2 l(\Xi;Y)}{\partial \mu^2} &= -\sum_{i=1}^{N} x_i, \\
\frac{\partial^2 l(\Xi;Y)}{\partial a_j^2} &= -\sum_{i=1}^{N} \theta_i^2 p_{ij}(1-p_{ij}), \\
\frac{\partial^2 l(\Xi;Y)}{\partial d_j^2} &= -\sum_{i=1}^{N} p_{ij}(1-p_{ij}), \\
\frac{\partial^2 l(\Xi;Y)}{\partial \gamma_j^2} &= -\sum_{i=1}^{N} x_i^2 p_{ij}(1-p_{ij}), \\
\frac{\partial^2 l(\Xi;Y)}{\partial a_j \partial d_j} &= \sum_{i=1}^{N} \theta_i p_{ij}(1-p_{ij}), \\
\frac{\partial^2 l(\Xi;Y)}{\partial a_j \partial \gamma_j} &= -\sum_{i=1}^{N} \theta_i x_i p_{ij}(1-p_{ij}), \\
\frac{\partial^2 l(\Xi;Y)}{\partial d_j \partial \gamma_j} &= \sum_{i=1}^{N} x_i p_{ij}(1-p_{ij}).
\end{aligned}$$

To implement, we can use Gauss Hermite quadrature to approximate the expectation of these terms so as to obtain $\hat{I}(\tilde{\Xi})$. Then $\hat{\Sigma}_N$ can be evaluated with $\hat{\Sigma}_N = \hat{I}^{-1}(\tilde{\Xi})$. This then enables Step 1 of Algorithm 1, where Monte Carlo samples of $\Xi^\dagger$ can be simulated from $\mathrm{N}(\tilde{\Xi}, \hat{\Sigma}_N)$.

Proof of Proposition IV.5. By considering for each of $k = 1, ..., p$, separately, the proofs are essentially same as that for Proposition IV.1. □

Proof of Proposition IV.7. Let $\Psi^\dagger = \{\boldsymbol{\mu}^\dagger, \boldsymbol{\sigma}^\dagger, \boldsymbol{a}^\dagger, \boldsymbol{d}^\dagger, \boldsymbol{\gamma}^\dagger\}$ be the model parameters such that $\boldsymbol{\gamma}_1^\dagger = \mathbf{0}$. Since model (4.9) with constraint $\boldsymbol{\gamma}_1^\dagger = \mathbf{0}$ is identifiable, by classical asymptotic theory for MLE (*van der Vaart*, 2000), $\tilde{\Psi} \xrightarrow{p} \Psi^\dagger$ as $N \to \infty$.

Equivalently, as $N \to \infty$, for any $\epsilon > 0$ and for all $j = 1, ..., J, k = 1, ..., p$, we have with probability tending to 1 that $|\tilde{\mu}_k - \mu_k^\dagger| \leq \epsilon$, $|\tilde{\sigma}_k - \sigma_k^\dagger| \leq \epsilon$, $|\tilde{a}_j - a_j^\dagger| \leq \epsilon$, $|\tilde{d}_j - d_j^\dagger| \leq \epsilon$, $|\tilde{\gamma}_{jk} - \gamma_{jk}^\dagger| \leq \epsilon$.

We fix a $k \in \{1, ..., p\}$. Note that the argument below would work for any $k = 1, ..., p$. Denote $f_k(c_k) = \sum_{j=1}^J |\gamma_{jk}^\dagger - c_k a_j^\dagger|$ as a function of c_k. Similarly, denote $f_{Nk}(c_k) = \sum_{j=1}^J |\tilde{\gamma}_{jk} - c_k \tilde{a}_j|$. Let $c_k^\dagger = \arg\min_{c_k} f_k(c_k)$ and $\hat{c}_k = \arg\min_{c_k} f_{Nk}(c_k)$, respectively. We aim to establish that $\hat{c}_k$ will converge in probability to $c_k^\dagger$ as $N \to \infty$. By regularity conditions, there exists $C_{k1} < \infty$ such that $J, |\gamma_{jk}^\dagger|, |a_j^\dagger| \leq C_{k1}$. Then, there must exist $C_{k2} < \infty$ such that $|c_k^\dagger| \leq C_{k2}$. Furthermore, note f_{Nk} is clearly continuous and convex in c_k, so consistency will follow if f_{Nk} can be shown to converge point-wise to f_k that is uniquely minimized at the true value $c_k^\dagger$ (point-wise convergence of convex functions implies their uniform convergence on compact subsets).

Following the model identifiability and Condition (4.10), $c_k^\dagger$ is unique. To see this, suppose for contradiction that there exist c_{k1} and c_{k2} such that $c_{k1} \neq c_{k2}$ and $c_{k1} = \arg\min_{c_k} f_k(c_k)$ and $c_{k2} = \arg\min_{c_k} f_k(c_k)$. First note that $a_j^\dagger = a_j^*$ for all $j = 1, ..., J$. Then by model identifiability, there exists c_{k3} such that $\gamma_{jk}^\dagger = \gamma_{jk}^* + c_{k3} a_j^*$. So we have

$$c_{k1} = \arg\min_{c_k} \sum_{j=1}^J |\gamma_{jk}^* + (c_{k3} - c_k) a_j^*|$$

and

$$c_{k2} = \arg\min_{c_k} \sum_{j=1}^J |\gamma_{jk}^* + (c_{k3} - c_k) a_j^*|.$$

Hence, $\gamma_{jk}^* = \gamma_{jk}^\dagger - (c_{k3} - c_{k1}) a_j^*$ and $\gamma_{jk}^* = \gamma_{jk}^\dagger - (c_{k3} - c_{k2}) a_j^*$. If Condition (4.10)

holds, then $c_{k3} = c_{k1}$ and $c_{k3} = c_{k2}$. This contradicts the assumption $c_{k1} \neq c_{k2}$. Hence, $c_k^\dagger$ must be unique.

For any $|c_k| \leq C_{k2}$,

$$
\begin{aligned}
&|f_{Nk}(c_k) - f_k(c_k)| \\
&= \Big| \sum_{j=1}^{J} \Big(|\tilde{\gamma}_{jk} - c_k \tilde{a}_j| - |\gamma_{jk}^\dagger - c_k a_j^\dagger| \Big) \Big| \\
&\leq \Big| \sum_{j=1}^{J} \Big(|(\tilde{\gamma}_{jk} - c_k \tilde{a}_j) - (\gamma_{jk}^\dagger - c_k a_j^\dagger)| \Big) \Big| \\
&= \Big| \sum_{j=1}^{J} \Big(|(\tilde{\gamma}_{jk} - \gamma_{jk}^\dagger) + c_k (a_j^\dagger - \tilde{a}_j)| \Big) \Big| \\
&\leq \sum_{j=1}^{J} \Big(|\tilde{\gamma}_{jk} - \gamma_{jk}^\dagger| + |c_k| \cdot |a_j^\dagger - \tilde{a}_j| \Big) \\
&\leq J\epsilon + |c_k|\epsilon. \\
&\leq (C_{k1} + C_{k2})\epsilon.
\end{aligned}
$$

Take $\epsilon_{k1} = (C_{k1}+C_{k2})\epsilon$, it follows that for any fixed $|c_k| \leq C_{k2}$, $P\big(|f_{Nk}(c_k) - f_k(c_k)| \leq \epsilon_{k1}\big) \to 1$ as $N \to \infty$. Moreover, following from the uniqueness of $c_k^\dagger$ and the continuity and the convexity of $f_{kN}(\cdot)$ in c_k, we must have $|\hat{c}_k - c_k^\dagger| = o_P(1)$ as $N \to \infty$.

Note that $\hat{\mu}_k = \tilde{\mu}_k + \hat{c}_k$, $\hat{\sigma}_k = \tilde{\sigma}_k$, $\hat{\gamma}_{jk} = \tilde{\gamma}_{jk} - \hat{c}_k \tilde{a}_j$, $\hat{a}_j = \tilde{a}_j$, $\hat{d}_j = \tilde{d}_j$ for $j = 1, ..., J$. From the model identifiability and Condition (4.10), we know that

$\mu_k^* = \mu_k^\dagger + c_k^\dagger$, $\sigma_k^* = \sigma_k^\dagger$, $\gamma_{jk}^* = \gamma_{jk}^\dagger - c_k^\dagger a_j^\dagger$, $a_j^* = a_j^\dagger$, $d_j^* = d_j^\dagger$ for all $j = 1, ..., J$. Since $|\hat{c}_k - c_k^\dagger| = o_P(1), |\tilde{\mu}_k - \mu_k^\dagger| = o_P(1)$, $|\tilde{\sigma}_k - \sigma_k^\dagger| = o_P(1)$, $|\tilde{\gamma}_{jk} - \gamma_{jk}^\dagger| = o_P(1), |\tilde{a}_j - a_j^\dagger| = o_P(1), |\tilde{d}_j - d_j^\dagger| = o_P(1)$ as $N \to \infty$, it follows directly from the Slutsky's Theorem that $|\hat{\mu}_k - \mu_k^*| = o_P(1)$, $|\hat{\sigma}_k - \sigma_k^*| = o_P(1)$, $|\hat{\gamma}_{jk} - \gamma_{jk}^*| = o_P(1), |\hat{a}_j - a_j^*| = o_P(1)$, $|\hat{d}_j - d_j^*| = o_P(1)$ as $N \to \infty$. Hence the first part of the proposition follows.

Now for the second part of the proposition, first note that $|\hat{\gamma}_{jk} - \gamma_{jk}^*| = o_P(1)$ implies that $\big||\hat{\gamma}_{jk}| - |\gamma_{jk}^*|\big| = o_P(1)$ as $N \to \infty$ for all $j = 1, ..., J$. That is, for any $\epsilon_{k2} > 0$, $\lim_{N\to\infty} P(\big||\hat{\gamma}_{jk}| - |\gamma_{jk}^*|\big| \leq \epsilon_{k2}) = 1$. When $\gamma_{jk}^* \neq 0$, let $\epsilon_{k3j} := |\gamma_{jk}^*| - \delta$. By the assumption that $0 < \delta < \min\{|\gamma_{jk}^*| : \gamma_{jk}^* \neq 0\}$, we know that $\epsilon_{k3j} > 0$. Take $\epsilon_{k2} = 0.5\epsilon_{k3j}$. Then,

$$\begin{aligned} |\hat{\gamma}_{jk}| - \delta &\geq |\gamma_{jk}^*| - \epsilon_{k2} - \delta \\ &= \epsilon_{k3j} - \epsilon_{k2} \\ &= \frac{1}{2}\epsilon_{k3j}. \end{aligned}$$

Hence, we have

$$\lim_{N\to\infty} P(|\hat{\gamma}_{jk}| - \delta \geq \frac{1}{2}\epsilon_{k3j}) = 1,$$

which then implies that

$$\lim_{N\to\infty} P(|\hat{\gamma}_{jk}| - \delta > 0) = 1.$$

Therefore, we have,

$$\lim_{N\to\infty} P(1_{\{|\hat{\gamma}_{jk}| \leq \delta\}} = 1_{\{\gamma_{jk}^* = 0\}}) = 1.$$

Therefore, the second part of the proposition follows. □

C.3 Additional Simulation Results for Multiple Group DIF Analysis

We present some additional simulation results for Section 4.3.3 in Chapter IV. It may be of interest of practitioners to perform statistical inference on individual DIF effects γ_{jk}. We examine the validity of using the proposed method and the LRT method to construct Wald intervals for individual γ_{jk} and check whether the empirical coverage rate is close to the nominal values. For both the proposed method and the LRT method, since empirical standard errors for each $\hat{\gamma}_{jk}$ can be obtained from M Monte Carlo samples directly, we first construct 95% Wald intervals for each γ_{jk} and validate its coverage rate. The results are summarized in Figures C.1 and C.2. We observe that for both methods, the empirical coverage rates for 95% Wald intervals are close to the nominal 95%, suggesting valid inference results for individual γ_{jk} can be obtained.

We further explore whether the false discovery rate (FDR) can be controlled in testing multiple items simultaneously. In specific, the Benjamini-Hochberg (B-H) (*Benjamini and Hochberg*, 1995) procedure is applied to the p-values obtained from the two methods. The results are given in Table C.1. As we can see, FDR is controlled well below/around the targeted level of 10% for both the proposed method and the LRT method with 1, 3, and 5 anchor items, under all settings.

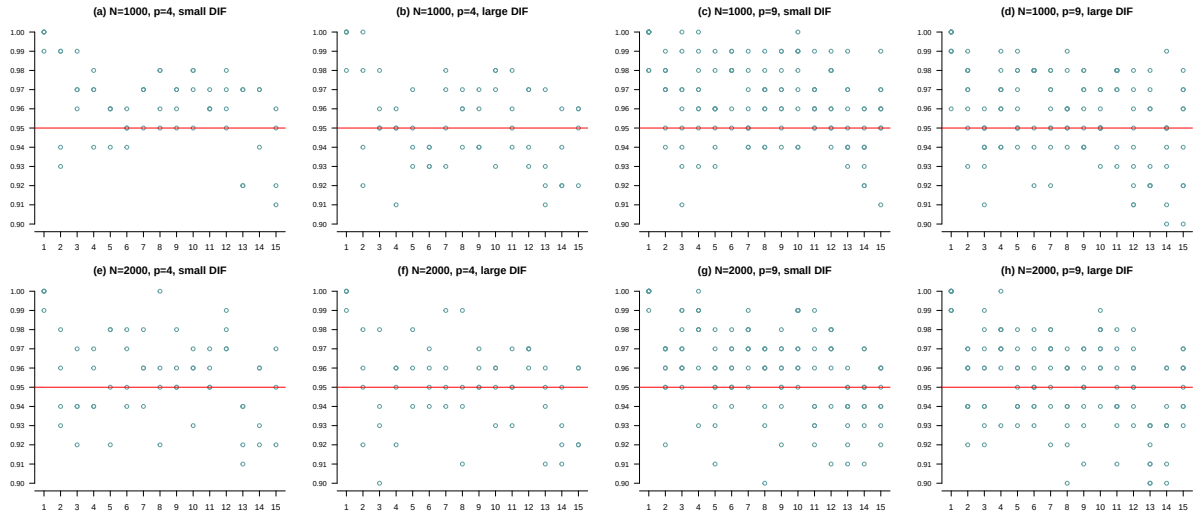

Figure C.1: Scatter-plots of the coverage rates of the 95% Wald intervals for γ_{jk}^*'s using Algorithm 7. X-axis and Y-axis represent the item indices and coverage rates respectively.

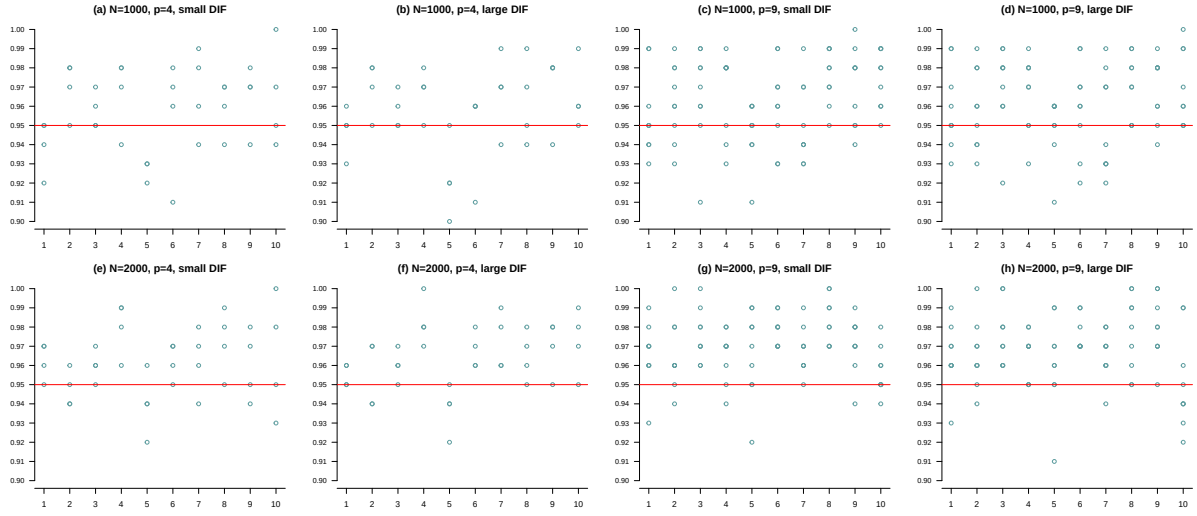

Figure C.2: Scatter-plots of the coverage rates of the 95% Wald intervals for γ_{jk}^*'s using the LRT method with five anchor items. X-axis and Y-axis represent the item indices and coverage rates respectively.

Table C.1: Comparison of average FDR out of 100 Monte Carlo samples of the proposed method and the LRT method with 1, 3 and 5 anchor items respectively when FDR is set to be 10% with B-H procedure. Small and Large denote the small DIF setting and the large DIF setting, respectively.

	$p = 4$				$p = 9$			
	$N = 1000$		$N = 2000$		$N = 1000$		$N = 2000$	
	Small	Large	Small	Large	Small	Large	Small	Large
Proposed method	0.053	0.055	0.071	0.058	0.007	0.020	0.027	0.024
LRT (1 anchor item)	0.072	0.061	0.044	0.042	0.052	0.102	0.048	0.040
LRT (3 anchor items)	0.046	0.050	0.070	0.078	0.043	0.070	0.053	0.056
LRT (5 anchor items)	0.088	0.088	0.056	0.052	0.106	0.096	0.057	0.054

C.4 Additional Results for PISA 2018 Application

Index the items used in the three test domains in Table C.2. We present the estimated parameters for Model (4.9) in the main article on the reading, the math and the science data respectively. The results for $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{d}}$ are summarized in Table C.3. The results for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$ are summarized in Table C.4. The results for $\hat{\boldsymbol{\gamma}}$ are summarized in Tables C.5 and C.6 for the reading data, Tables C.7 and C.8 for the math data, and Tables C.9 and C.10 for the science data.

Table C.2: List of indices for the PISA 2018 reading items, the math items and the science items respectively.

Index	Reading Items	Math Items	Science	Items Index	Reading Items	Math Items	Science Items
1	CR424Q02S	CM033Q01S	CS408Q01S	39	CR104Q02S	CM909Q03S	CS527Q01S
2	CR424Q03S	CM474Q01S	CS408Q04S	40	DR466Q02C	CM949Q01S	CS527Q03S
3	CR424Q07S	CM155Q01S	CS408Q05S	41	CR466Q03S	CM949Q02S	CS527Q04S
4	CR220Q01S	CM155Q04S	CS413Q06S	42	CR466Q06S	CM00GQ01S	CS428Q01S
5	CR220Q02S	CM411Q01S	CS413Q04S	43	CR412Q01S	CM998Q04S	CS428Q03S
6	CR220Q04S	CM411Q02S	CS413Q05S	44	CR412Q05S		CS634Q01S
7	CR220Q05S	CM803Q01S	CS635Q02S	45	DR412Q08C		CS634Q04S
8	CR220Q06S	CM442Q02S	CS604Q02S	46	DR432Q01C		CS629Q02S
9	CR067Q01S	CM034Q01S	CS625Q02S	47	DR432Q05C		CS629Q04S
10	CR456Q01S	CM305Q01S	CS625Q03S	48	CR432Q06S		CS648Q02S
11	DR456Q02C	CM496Q01S	CS626Q01S	49	DR219Q01C		CS648Q03S
12	DR456Q06C	CM496Q02S	CS626Q02S	50	DR219Q01EC		CS498Q02S
13	DR420Q02C	CM423Q01S	CS626Q03S	51	DR219Q02C		CS605Q01S
14	DR420Q06C	CM192Q01S	CS425Q05S	52	DR460Q01C		CS605Q02S
15	DR420Q09C	CM603Q01S	CS425Q02S	53	CR460Q05S		CS605Q03S
16	DR455Q02C	CM571Q01S	CS438Q01S	54	CR460Q06S		CS646Q01S
17	DR455Q03C	CM564Q01S	CS438Q02S	55	DR406Q01C		CS646Q02S
18	CR455Q04S	CM564Q02S	CS608Q01S	56	DR406Q05C		CS646Q03S
19	CR455Q05S	CM447Q01S	CS608Q02S	57	DR406Q02C		CS620Q01S
20	CR055Q01S	CM273Q01S	CS608Q03S	58	CR227Q01S		CS620Q02S
21	DR055Q02C	CM408Q01S	CS643Q01S	59	DR227Q03C		CS645Q03S
22	DR055Q05C	CM420Q01S	CS643Q02S	60	DR227Q06C		CS478Q02S
23	CR111Q01S	CM446Q01S	CS643Q04S	61	DR102Q04C		CS478Q03S
24	CR446Q03S	CM559Q01S	CS610Q02S	62	DR102Q05C		CS415Q07S
25	DR446Q06C	CM828Q03S	CS466Q01S	63	CR102Q07S		CS415Q02S
26	CR437Q01S	CM464Q01S	CS256Q01S	64			CS415Q08S
27	DR437Q07C	CM800Q01S	CS326Q03S	65			CS627Q01S
28	CR437Q06S	CM982Q01S	CS326Q04S	66			CS627Q03S
29	CR404Q03S	CM982Q02S	CS602Q01S	67			CS627Q04S
30	CR404Q06S	CM982Q03S	CS602Q02S	68			CS607Q01S
31	CR404Q07S	CM982Q04S	CS602Q04S	69			CS607Q02S
32	DR404Q10AC	CM992Q01S	CS603Q01S	70			CS638Q01S
33	DR404Q10BC	CM992Q02S	CS603Q03S	71			CS638Q02S
34	CR453Q01S	CM915Q01S	CS603Q04S	72			CS638Q04S
35	DR453Q04C	CM915Q02S	CS603Q05S	73			CS615Q07S
36	CR453Q05S	CM906Q01S	CS657Q01S	74			CS615Q01S
37	DR453Q06C	CM909Q01S	CS657Q02S	75			CS615Q02S
38	CR104Q01S	CM909Q02S	CS657Q03S	76			CS615Q05S

Table C.3: Estimated $\hat{a}_j$ and $\hat{d}_j$ parameters for Model (4.9) on the PISA 2018 reading data, math data and science data respectively.

	Reading		Math		Science			Reading		Math		Science	
Item Index	$\hat{a}_j$	$\hat{d}_j$	$\hat{a}_j$	$\hat{d}_j$	$\hat{a}_j$	$\hat{d}_j$	Item Index	$\hat{a}_j$	$\hat{d}_j$	$\hat{a}_j$	$\hat{d}_j$	$\hat{a}_j$	$\hat{d}_j$
1	0.45	-0.61	0.74	1.89	0.98	0.23	39	0.52	-0.19	2.31	-1.02	1.01	-1.91
2	0.51	0.38	0.76	0.97	0.57	0.35	40	1.10	-0.48	1.28	1.22	0.42	0.70
3	1.33	1.59	1.01	1.34	0.88	-1.13	41	0.65	-2.59	1.19	-0.93	0.78	0.47
4	1.09	-1.40	0.77	0.44	1.49	-0.95	42	1.67	2.56	1.36	-3.22	1.10	0.17
5	1.36	0.72	1.28	0.22	0.96	-0.01	43	0.85	2.58	0.25	-0.57	1.44	1.63
6	1.14	0.32	0.76	-0.00	0.79	0.95	44	0.83	0.06			1.32	-1.62
7	1.92	2.58	1.63	-1.03	1.30	1.35	45	1.08	0.04			1.43	0.12
8	1.25	0.47	1.60	-1.07	0.84	-0.18	46	1.68	2.98			0.73	-0.17
9	1.17	1.81	1.13	-0.39	0.75	0.77	47	1.40	1.94			0.73	0.51
10	2.13	3.99	0.22	-0.31	1.06	0.16	48	1.78	-3.08			0.77	-0.33
11	1.16	1.73	1.34	0.17	0.75	0.64	49	1.31	2.49			0.51	0.58
12	1.28	1.28	1.03	0.99	0.74	0.22	50	1.08	1.42			0.60	-0.30
13	1.01	0.68	0.70	2.08	1.03	1.11	51	1.28	2.13			0.89	0.11
14	0.66	-0.42	1.11	-0.30	0.66	0.95	52	1.39	0.41			1.37	-0.88
15	0.86	0.93	0.83	-0.59	0.94	0.22	53	1.69	1.39			0.72	0.29
16	0.82	0.22	1.11	-0.12	0.88	1.44	54	0.86	0.32			1.73	2.57
17	0.94	1.84	0.58	-0.23	1.31	1.21	55	1.22	-0.27			1.11	0.35
18	1.24	0.56	0.67	-0.19	0.51	-0.53	56	1.31	0.88			0.93	1.10
19	1.38	-1.22	1.14	1.04	1.18	0.75	57	0.85	-0.87			0.99	1.97
20	1.41	1.31	0.80	-0.14	0.63	-0.10	58	0.86	0.02			0.76	-0.56
21	1.25	0.19	1.15	-0.13	2.00	2.07	59	1.31	0.21			0.92	0.28
22	1.83	1.37	0.94	0.56	1.65	0.25	60	1.49	1.69			1.19	0.12
23	1.45	0.78	1.45	1.49	1.59	-1.01	61	1.19	-1.60			0.71	0.45
24	1.78	3.33	0.74	0.49	1.28	2.34	62	1.08	-0.85			0.79	1.54
25	1.15	1.25	0.91	-0.76	0.87	1.03	63	1.23	2.31			1.50	2.08
26	0.68	-0.06	1.83	-1.61	0.66	2.63	64					0.97	0.36
27	0.76	-1.70	0.51	1.75	1.00	0.44	65					0.58	-0.29
28	0.89	0.18	0.70	1.89	0.84	-1.34	66					0.77	1.71
29	1.43	1.64	0.71	-0.54	1.15	2.23	67					1.27	1.36
30	0.85	0.06	0.93	0.56	0.65	-0.89	68					1.36	3.12
31	0.99	-0.76	1.23	-0.37	1.13	1.27	69					1.55	0.91
32	1.94	0.49	1.02	1.70	1.67	1.69	70					1.42	0.47
33	1.69	-0.52	1.59	-2.43	1.26	1.49	71					1.43	2.27
34	1.42	1.82	0.75	0.01	0.75	0.60	72					1.10	0.10
35	1.11	1.17	1.09	1.72	1.09	0.62	73					1.13	-1.47
36	1.35	-0.13	0.99	0.56	0.56	1.14	74					1.04	1.86
37	1.31	2.00	1.30	3.33	0.73	-0.39	75					1.49	-0.29
38	1.21	0.77	1.35	0.84	1.09	0.02	76					0.48	-1.46

Table C.4: Estimated $\hat{\mu}_k$ and $\hat{\sigma}_k$ parameters for Model (4.9) on the PISA 2018 reading data, math data and science data, respectively.

	Reading		Math		Science	
Country	$\hat{\mu}_k$	$\hat{\sigma}_k$	$\hat{\mu}_k$	$\hat{\sigma}_k$	$\hat{\mu}_k$	$\hat{\sigma}_k$
Australia	0.00	1.00	0.00	1.00	0.00	1.00
Austria	-0.02	1.14	-0.02	1.20	-0.12	1.14
Belgium	0.08	1.07	-0.00	1.24	-0.13	1.10
Canada	0.28	1.06	0.05	1.15	0.09	1.07
Chile	-0.40	0.96	-0.98	1.00	-0.69	0.93
Colombia	-0.89	0.96	-1.35	0.99	-1.06	0.92
Czech Republic	0.02	1.08	-0.03	1.20	-0.09	1.09
Denmark	0.06	1.00	0.00	1.09	-0.10	1.05
Estonia	0.43	1.01	0.17	1.10	0.25	1.02
Finland	0.32	1.08	-0.00	1.07	0.18	1.13
France	0.09	1.08	-0.10	1.17	-0.14	1.05
Germany	0.12	1.13	-0.02	1.28	-0.06	1.16
Greece	-0.28	1.06	-0.69	1.12	-0.68	0.91
Hungary	-0.09	1.05	-0.41	1.18	-0.32	1.07
Iceland	-0.03	1.25	-0.13	1.17	-0.29	1.09
Ireland	0.23	0.96	-0.15	1.04	-0.14	1.01
Israel	-0.20	1.27	-0.48	1.34	-0.45	1.18
Italy	-0.06	1.07	-0.14	1.20	-0.37	1.00
Japan	0.18	1.01	0.41	1.14	0.25	1.13
Korea	0.20	1.08	0.17	1.13	0.29	1.17
Latvia	-0.05	1.00	-0.16	1.02	-0.15	0.98
Lithuania	-0.10	1.06	-0.34	1.19	-0.27	1.06
Luxembourg	-0.14	1.21	-0.30	1.26	-0.31	1.13
Mexico	-0.73	0.88	-1.23	0.95	-0.99	0.80
Netherlands	-0.05	1.08	-0.13	1.24	-0.11	1.18
New Zealand	0.17	1.11	-0.08	1.18	-0.01	1.17
Norway	0.17	1.13	-0.06	1.11	-0.20	1.10
Poland	0.27	0.99	0.09	1.17	0.09	1.08
Portugal	-0.14	1.04	-0.16	1.18	-0.18	1.05
Slovak Republic	-0.31	1.09	-0.29	1.23	-0.50	1.10
Slovenia	0.13	1.02	0.07	1.16	-0.03	1.03
Spain	-0.12	0.96	-0.31	1.10	-0.39	0.98
Sweden	0.22	1.15	-0.12	1.15	-0.12	1.11
Switzerland	-0.05	1.16	0.13	1.20	-0.16	1.10
Turkey	-0.22	0.92	-0.70	1.17	-0.57	0.92
United Kingdom	0.16	1.03	-0.14	1.17	-0.00	1.08
United States	0.02	1.19	-0.35	1.24	-0.01	1.19

Table C.5: Estimated $\hat{\gamma}_{jk}$ parameters for $k = 1, ..., 18$ for Model (4.9) on the PISA 2018 reading data.

1	0.34	0.18	-0.14	-0.02	0.23	0.07	0.21	0.02	0.24	0.16	0.22	-0.03	0.23	-0.03	-0.05	0.14	0.39	0.01
2	0.13	-0.09	0.01	-0.13	0.02	-0.20	-0.07	-0.01	-0.06	0.01	0.14	-0.29	-0.12	-0.47	-0.52	-0.48	-0.05	0.05
3	-0.01	0.05	0.05	0.37	0.63	0.18	0.16	0.20	-0.01	-0.07	0.09	0.09	0.00	-0.05	-0.03	0.44	0.69	-0.25
4	-0.35	0.04	-0.04	-0.40	-0.20	0.59	-0.16	0.16	-0.23	0.22	-0.15	-0.76	0.37	0.05	-0.01	-0.63	0.04	0.96
5	-0.11	0.08	-0.01	0.07	0.22	0.40	0.15	-0.30	-0.25	0.24	-0.15	0.07	0.33	0.31	-0.18	0.19	0.38	-0.03
6	0.63	0.87	0.03	0.89	1.24	0.43	-0.42	-0.51	0.91	1.29	0.88	1.14	0.36	0.35	-0.27	1.06	1.30	0.23
7	0.71	0.47	0.04	0.82	1.46	1.05	0.21	0.39	0.51	0.74	0.55	0.52	0.42	0.67	0.06	0.78	1.03	0.23
8	1.24	0.96	0.14	0.79	1.22	1.49	0.79	-0.07	0.84	0.97	1.30	0.90	1.01	0.68	-0.02	0.39	1.44	0.15
9	0.49	-0.05	0.08	0.40	0.75	0.20	-0.26	0.43	0.14	-0.04	0.36	-0.02	0.28	-0.08	0.27	-0.11	0.54	-0.23
10	1.07	0.81	0.48	0.43	1.31	0.65	0.55	0.80	-0.31	0.70	0.51	-0.31	-0.27	0.80	0.90	0.36	0.25	0.00
11	-0.52	-1.09	-0.16	-0.49	-0.98	-0.44	-0.69	0.33	-0.50	-1.02	0.06	0.20	0.10	0.44	0.81	-0.20	-1.06	-0.45
12	0.10	0.05	-0.12	0.60	0.57	0.58	-0.25	0.00	-0.10	-0.53	-0.25	0.55	0.36	0.03	0.39	0.22	-0.46	0.32
13	1.34	0.68	-0.05	0.75	0.80	1.13	0.35	0.63	0.46	0.52	1.07	1.39	0.76	0.85	-0.51	0.68	-0.07	2.38
14	-0.25	0.28	0.18	-0.32	-0.86	-1.23	0.44	-0.05	-0.15	-0.61	-0.25	0.33	-0.58	-0.42	0.10	0.01	-0.41	0.83
15	0.02	0.35	0.07	0.12	0.32	0.22	-0.06	-0.43	-0.25	0.21	-0.11	0.34	0.25	0.06	1.10	0.12	-0.09	0.46
16	-0.68	-0.62	-0.28	-0.09	-0.69	-0.76	-0.39	-1.11	-0.73	-0.38	-0.91	-0.02	-0.93	-0.30	0.10	-0.83	-0.19	-0.47
17	0.16	-0.04	-0.20	0.22	-0.37	0.05	0.17	-0.76	-0.25	-0.81	0.05	0.40	0.45	0.36	0.11	0.69	0.93	0.97
18	-0.09	0.11	-0.21	0.02	0.79	0.09	0.20	-0.27	0.25	-0.02	0.10	0.48	1.17	-0.37	-0.23	1.04	0.51	-0.40
19	-0.99	0.16	-0.47	0.22	-0.28	-0.22	-0.74	-0.29	0.14	-1.07	-0.89	-0.07	-0.92	-0.23	-0.06	-0.13	-0.79	0.26
20	0.50	-0.36	0.09	0.67	0.58	0.11	0.32	0.59	0.19	-0.07	0.13	-0.14	0.05	0.47	0.18	-0.26	0.26	-0.60
21	0.52	-0.20	0.01	-0.07	-0.38	-0.10	-0.17	0.05	-0.71	0.18	0.21	0.40	-0.03	-1.05	-0.16	-0.20	-0.41	0.03
22	-0.43	-0.51	0.07	0.13	-0.13	-0.43	-0.33	-0.14	-0.24	0.22	-0.26	-1.15	-0.29	-0.28	-0.05	0.27	0.39	-1.10
23	0.30	-0.60	-0.00	0.57	0.61	0.32	0.11	0.06	0.60	-0.62	0.43	0.50	-0.05	-0.91	-0.42	0.48	0.41	0.12
24	0.15	0.11	0.18	-0.82	-0.81	-0.38	0.40	0.00	-0.05	0.01	0.00	0.19	0.65	0.05	0.50	-0.11	-0.50	0.25
25	-0.88	-0.58	-0.24	-0.30	0.04	-0.87	-0.51	-0.20	-0.56	-0.78	-0.72	-0.39	-0.21	-0.51	0.30	-0.87	-0.73	0.76
26	-0.03	0.33	-0.05	-0.00	0.03	0.54	0.36	0.24	0.14	0.71	-0.11	0.51	0.35	0.33	-0.47	0.20	-0.08	-0.42
27	0.04	0.13	0.28	-0.56	-0.01	-0.16	0.42	0.60	-0.20	1.15	-0.70	0.09	0.80	0.23	0.36	0.64	0.15	1.14
28	0.36	0.04	0.04	0.20	0.32	0.27	0.01	-0.75	-0.62	0.22	0.79	0.60	0.96	0.33	-0.16	-0.14	0.45	0.55
29	0.16	-0.24	0.05	0.02	0.28	0.10	0.00	-0.07	0.55	0.20	-0.06	0.18	-0.07	-0.22	0.14	0.25	0.73	0.15
30	-0.54	-0.35	-0.18	0.03	-0.01	-0.24	-0.50	0.02	-0.53	-0.06	-0.44	0.50	0.14	-0.39	-0.38	-0.14	-0.27	0.03
31	0.75	0.00	-0.09	-0.76	-0.83	0.09	-0.46	0.40	0.00	-0.05	0.59	-0.77	0.35	-0.13	-0.39	-0.07	-0.49	-0.20
32	-0.04	-0.06	-0.11	-0.59	0.47	-0.39	-0.33	-0.76	-0.21	-0.24	0.09	0.67	-0.69	0.31	0.08	0.10	0.00	-0.17
33	0.04	-0.08	0.27	0.00	1.00	0.00	-0.11	0.00	-0.27	0.44	0.69	0.82	-0.18	0.45	0.02	0.51	0.35	0.39
34	-0.32	-0.23	0.04	0.09	0.24	0.50	0.34	-0.37	1.29	0.38	-0.62	-0.09	-0.23	-0.05	0.43	-0.34	0.23	-0.47
35	-0.76	-0.87	0.09	0.14	-0.83	-1.10	0.36	0.03	-0.27	0.51	0.23	0.32	-0.64	-0.53	-0.04	-0.45	-0.53	-1.20
36	0.28	-0.04	-0.02	-0.11	-0.28	0.04	0.26	0.01	0.49	-0.02	0.36	0.09	0.34	0.07	0.17	-0.30	0.22	-0.00
37	-0.59	-0.05	-0.02	-0.73	-0.78	-1.57	-0.28	-1.86	-1.19	-0.60	-0.40	0.43	-0.96	-1.63	-0.22	-0.07	-0.45	-1.23
38	-0.19	0.17	-0.12	-0.38	0.09	-0.36	-0.12	-0.09	0.06	0.00	-0.11	-0.18	-0.56	0.09	-0.09	-0.07	-0.77	-0.07
39	-0.21	0.15	-0.19	-0.00	-0.29	0.34	0.21	-0.46	-0.35	0.10	-1.00	-0.26	-0.23	-0.35	0.08	-0.22	-0.12	-0.34
40	0.15	0.42	0.03	-0.06	-0.13	-0.88	-0.18	0.43	0.18	0.63	0.85	-1.17	0.38	0.13	0.17	0.92	0.22	0.66
41	0.40	1.04	-0.22	-0.17	0.66	0.09	0.25	0.72	0.49	0.53	0.60	0.35	0.35	0.36	-0.35	0.78	0.41	0.59
42	-0.37	0.65	0.03	0.20	-0.21	-0.33	0.64	0.33	0.43	0.22	-0.15	-0.20	0.47	0.31	-0.09	0.19	0.00	0.74
43	-0.75	-0.67	0.00	-0.55	-0.51	-1.02	-0.52	-1.16	-0.75	-0.93	-0.39	-0.83	-0.70	-0.09	-0.22	-0.53	-1.18	-1.30
44	0.51	0.70	0.23	-0.09	0.25	0.39	0.59	0.41	0.83	0.68	0.38	0.82	0.17	0.53	0.01	0.91	0.32	-0.66
45	-0.53	-0.52	0.06	0.55	-0.17	-0.10	-0.72	-0.39	-0.21	-0.08	-0.14	-0.02	0.47	-0.91	0.13	0.07	-0.28	-0.97
46	-1.10	-0.73	0.00	-0.75	-1.15	0.57	-0.55	0.38	0.27	-0.53	-0.98	-0.22	-0.27	-0.36	0.41	-1.29	-0.01	-0.68
47	-1.26	-0.18	0.37	-0.79	-1.21	-1.15	-0.77	-1.14	-1.14	-0.72	-1.51	-1.08	-1.33	-0.45	0.34	-0.72	-1.27	-0.86
48	-1.30	-0.78	-0.39	-1.14	-0.74	-0.51	-1.55	-1.63	-1.10	-0.41	-1.52	-1.13	-1.10	-1.70	-0.93	-0.40	-1.06	-0.21
49	-0.34	-1.04	-0.51	-1.22	-1.92	-0.87	-0.63	-0.96	-0.99	-1.19	-1.03	-1.07	-0.75	-1.22	-0.10	-0.58	-0.67	-1.10
50	0.42	-0.11	-0.32	-1.47	-1.96	-0.46	-0.15	-0.41	-0.72	-0.29	0.29	-1.28	-0.57	-0.63	-0.28	-0.35	-0.79	-0.04
51	-0.38	-0.46	-0.05	-0.10	-0.86	-1.07	-0.05	-0.45	-0.94	-0.14	-0.26	-0.25	-0.54	-0.81	-0.37	-0.48	-0.97	-0.99
52	-0.15	0.29	0.55	0.40	0.00	0.03	0.39	-0.64	0.46	0.28	-0.22	0.61	0.59	0.05	0.28	0.01	0.05	1.14
53	0.00	0.01	0.33	-0.20	0.22	-0.38	-0.21	0.02	0.21	0.12	0.06	-0.25	0.02	0.00	0.13	-0.46	-0.32	0.04
54	-0.07	-0.07	-0.06	-0.13	0.05	-0.21	-0.12	-0.20	-0.58	-0.29	-0.27	0.10	-0.27	-0.04	-0.46	-0.09	0.17	-0.34
55	1.40	1.11	0.28	0.87	0.94	0.72	1.17	0.48	0.09	1.04	0.50	1.16	0.62	0.82	0.63	1.40	0.82	0.73
56	0.67	0.08	-0.21	0.47	0.05	0.21	0.04	0.26	0.33	0.34	-0.18	-0.60	-0.64	0.13	0.17	0.53	0.07	0.82
57	-0.15	-0.04	-0.30	-0.43	-0.64	0.10	-0.58	-0.56	0.09	-0.10	-1.50	0.47	-0.53	-1.22	0.14	-0.08	-0.36	-0.35
58	0.78	0.18	0.17	0.26	0.70	0.63	0.84	-0.29	0.82	-0.18	0.63	0.25	0.26	0.76	-0.39	0.33	-0.30	1.10
59	0.83	0.86	0.54	0.91	0.21	-0.16	0.40	0.52	0.30	0.58	0.80	-0.34	0.16	0.25	-0.04	0.64	1.05	0.63
60	0.37	0.16	-0.21	-0.52	-0.56	-0.30	0.39	-0.66	0.15	0.07	-0.12	-0.46	-0.01	-0.14	-0.20	-0.00	-0.50	-0.20
61	0.12	0.45	-0.35	-0.19	-0.19	0.41	0.42	0.16	0.57	-0.75	0.38	-0.15	0.52	0.43	-0.07	-1.16	-0.15	0.77
62	-0.56	-0.21	-0.20	-1.60	-0.46	-0.66	0.20	-0.14	-0.67	-0.66	-0.30	-0.35	-0.59	-0.67	-0.14	-0.52	-1.14	-0.41
63	-0.85	-1.17	-0.21	0.29	-0.37	-0.59	-0.48	0.13	0.51	-0.61	-0.80	0.00	0.02	-0.58	0.00	-0.10	-0.32	-0.93

Table C.6: Estimated $\hat{\gamma}_{jk}$ parameters for $k = 19, ..., 36$ for Model (4.9) on the PISA 2018 reading data.

Item Index	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	-0.22	0.14	0.00	0.12	0.12	0.17	-0.11	0.01	0.01	0.04	0.15	-0.02	0.14	0.03	0.35	-0.03	0.02	-0.04
2	0.29	-0.58	-0.46	0.05	-0.10	0.65	-0.32	-0.45	-0.46	-0.11	-0.03	-0.22	0.19	-0.62	-0.00	-0.45	-0.21	-0.08
3	-0.23	0.55	0.37	0.09	0.62	0.27	0.03	0.15	0.19	0.43	-0.16	0.10	0.15	0.03	-0.01	0.70	-0.05	0.05
4	-0.88	0.39	-0.06	0.30	-0.16	-0.50	0.17	0.34	0.35	-0.10	0.61	0.13	-0.18	0.22	0.18	-0.88	-0.07	0.01
5	0.00	0.04	-0.31	0.04	0.08	-0.21	0.31	-0.15	-0.27	0.15	0.14	-0.30	0.05	0.31	0.06	0.12	-0.04	-0.08
6	0.44	0.71	0.30	0.91	0.77	1.15	-0.12	0.02	0.23	0.53	0.49	0.50	0.55	0.16	0.82	1.49	-0.32	0.17
7	0.32	0.77	0.86	0.92	0.32	0.71	0.11	0.28	0.75	0.85	0.82	0.19	-0.27	0.63	0.82	0.54	-0.07	0.12
8	0.67	0.78	0.79	1.47	0.61	0.78	0.00	0.40	0.63	1.03	1.02	1.01	1.03	0.90	1.37	0.68	0.14	0.18
9	0.11	0.71	0.72	0.33	0.05	-0.42	-0.02	-0.15	0.40	0.78	0.46	0.52	0.16	0.08	0.36	0.00	-0.07	0.28
10	-0.22	0.74	0.65	1.00	1.38	1.14	0.26	0.00	0.20	0.51	0.55	0.68	0.10	-0.51	0.34	1.24	0.14	-0.00
11	-0.10	-0.11	-0.61	0.00	-0.39	-1.02	-0.04	-0.06	-0.48	-0.46	-1.13	-1.05	0.24	0.08	-0.50	0.31	0.02	-0.22
12	1.45	-0.14	0.11	-0.14	0.26	-0.14	0.45	0.28	-0.50	0.22	0.03	-0.45	-0.07	0.39	-0.15	0.27	0.04	0.10
13	0.81	1.05	1.33	1.01	0.78	1.31	-0.30	1.01	1.25	0.80	1.08	1.16	0.68	0.56	1.40	0.23	0.05	-0.05
14	0.78	-0.31	-0.10	-0.11	-0.16	0.68	-0.20	0.15	-0.58	0.25	-1.26	-0.24	0.34	0.16	-0.12	0.03	-0.19	-0.17
15	-0.07	0.04	0.29	0.08	-0.02	0.34	0.19	-0.16	-0.07	0.13	0.15	0.08	0.15	0.05	0.10	0.47	0.18	0.26
16	-0.30	-1.19	-1.00	-0.09	-0.75	-0.71	-0.28	0.13	-0.38	-0.25	-1.29	-0.81	-0.54	-0.11	-0.57	-0.27	0.05	0.16
17	1.93	0.16	0.52	-0.01	0.55	0.83	-0.03	0.95	0.50	0.81	0.69	0.10	0.38	0.89	-0.46	0.77	-0.34	-0.01
18	1.27	0.97	0.18	0.25	0.21	0.14	-0.06	0.51	0.54	0.76	0.52	0.31	0.64	0.02	0.02	0.28	-0.29	0.41
19	0.78	-0.06	-0.51	-0.11	0.01	0.29	-0.20	0.47	0.22	-0.07	0.31	0.24	-0.45	-0.90	-0.73	-0.63	-0.13	-0.56
20	0.20	0.23	-0.47	0.20	0.87	-0.32	0.03	-0.24	-0.48	0.71	0.33	0.52	0.21	-0.17	0.16	-0.30	-0.22	0.09
21	-0.22	-0.33	-0.15	-0.21	0.13	-0.98	-0.25	-0.41	0.25	0.58	-0.38	-0.76	-0.26	-0.26	-0.09	0.39	-0.10	-0.40
22	0.56	0.53	0.42	-0.37	-0.60	-0.57	0.15	-0.25	0.54	0.48	-0.50	0.13	0.11	-0.10	-0.31	0.06	0.37	-0.11
23	1.21	0.20	0.08	-0.12	-0.07	-0.35	0.22	0.54	0.72	-0.36	0.05	0.09	0.43	0.64	0.17	0.57	-0.53	0.07
24	0.43	0.08	-0.52	0.39	-0.43	1.41	0.23	-0.17	-1.52	0.00	-0.24	0.30	-0.91	0.13	0.36	-0.83	0.34	0.07
25	1.00	-0.38	-0.53	-0.77	-0.47	0.32	0.04	-0.10	-0.84	-0.65	-1.18	-1.58	-0.64	-0.52	-0.92	-0.30	0.03	-0.26
26	-0.01	0.44	0.81	0.09	-0.42	-0.11	-0.23	-0.19	0.84	-0.22	0.44	0.26	0.24	-0.01	-0.03	0.14	-0.12	0.38
27	0.23	0.63	0.19	0.24	-0.08	-0.30	0.37	0.09	0.00	0.39	-0.20	0.22	0.18	-0.04	0.13	1.24	-0.07	0.23
28	0.86	-0.39	-0.03	0.34	0.22	0.00	0.01	-0.78	-0.30	0.22	0.40	-0.20	0.39	-0.56	0.19	0.33	0.11	0.41
29	0.06	0.12	0.27	0.05	0.12	-0.83	0.43	-0.15	0.08	-0.14	0.30	-0.68	-0.04	-0.19	-0.04	0.13	0.13	0.09
30	-0.48	-0.39	-0.55	-0.30	-0.20	0.02	-0.02	-0.70	0.10	0.10	0.06	-0.15	-0.12	-0.37	-0.48	-0.26	-0.12	-0.22
31	-0.62	0.13	-0.62	0.26	-0.78	0.57	-0.01	0.13	0.33	-0.39	-0.14	0.65	-0.70	0.13	0.48	-0.48	0.04	-0.16
32	-0.06	-0.60	-0.10	-0.12	-0.36	-0.11	0.00	0.45	-0.26	-0.19	-0.80	-0.55	-0.48	-0.60	0.03	-0.61	0.05	-0.25
33	-0.20	-0.11	0.31	0.24	0.05	-0.03	0.04	0.00	0.29	-0.13	-0.00	-0.25	0.14	0.01	0.11	-0.04	0.40	-0.02
34	-0.69	-0.04	0.22	-0.28	0.23	-0.00	0.06	0.58	-0.17	0.18	0.14	-0.16	0.02	0.22	-0.28	-0.36	0.21	0.11
35	-0.79	0.03	-0.55	-0.19	-0.92	-1.20	-0.45	0.28	-0.61	-0.50	-1.08	-0.35	-1.11	0.35	-0.16	0.55	0.08	0.17
36	-0.92	0.22	-0.55	0.00	0.24	-0.11	-0.05	0.41	0.01	-0.05	0.32	0.24	-0.18	0.13	0.22	-0.26	0.00	0.10
37	-0.28	-1.04	-1.09	-0.44	-0.71	-0.21	-0.63	-0.05	-0.96	-0.60	-1.55	-1.10	-0.66	-0.42	0.15	-0.91	-0.17	-0.52
38	-0.04	-0.50	-0.01	-0.01	-0.02	-0.28	0.16	0.05	-0.76	-0.21	-0.45	-0.13	0.00	0.23	0.16	-0.29	-0.10	-0.01
39	-0.41	-0.18	-0.43	-0.69	-0.20	-0.32	-0.04	-0.56	0.06	0.14	0.19	0.23	-0.29	-0.41	-0.29	-0.05	-0.25	-0.16
40	-0.10	0.72	0.19	0.13	0.45	0.78	-0.67	-0.34	1.10	0.37	0.82	0.37	0.28	0.47	0.74	0.55	-0.23	-0.47
41	0.79	0.58	0.53	0.77	0.17	1.19	-0.28	0.97	0.23	0.23	0.53	0.21	0.51	0.41	0.56	0.34	-0.27	-0.84
42	-0.11	-1.07	0.28	0.53	0.02	0.75	0.18	0.09	-0.15	0.72	0.27	0.61	0.27	-0.42	0.09	-0.19	0.38	0.19
43	-0.37	-0.59	-0.74	-0.36	-0.67	-0.07	0.23	0.07	-0.69	-0.98	-0.65	-0.59	-0.66	-0.35	-0.82	-0.19	0.06	0.16
44	-0.17	0.08	0.13	0.79	0.00	0.61	0.01	0.48	0.53	0.51	0.05	0.33	0.26	0.52	0.82	0.51	0.02	0.02
45	-1.07	-0.23	0.14	-0.99	-0.49	0.00	-0.37	-0.46	0.09	-0.63	-0.07	-1.17	-0.41	-0.37	-0.52	0.51	-0.48	-0.13
46	-1.36	-0.52	0.10	-0.98	-1.52	0.04	-0.23	-0.82	0.42	-0.47	0.48	-0.61	-0.69	-1.07	-0.96	-0.74	0.16	0.40
47	-0.49	-1.34	-1.11	-0.76	-1.10	-0.17	-0.07	-0.66	-1.04	-1.20	-1.36	-0.74	-0.99	-0.60	-1.19	-1.08	0.12	-0.10
48	-1.08	-1.54	-1.33	-1.00	-0.84	-0.81	-0.36	-0.94	-1.97	-1.61	-0.46	-1.11	-0.70	-1.19	-0.89	-1.08	-0.20	-0.33
49	-0.23	-0.66	-1.04	-0.78	-1.49	-0.51	-0.28	-0.70	-0.86	-1.16	-0.90	-0.11	-0.62	-0.52	-1.12	-1.17	-0.15	-0.32
50	0.40	-0.09	-0.75	0.01	-1.80	0.77	-0.12	-0.29	-0.33	-0.54	-0.39	-0.20	-0.15	-0.22	0.14	-0.76	-0.18	-0.37
51	0.96	-0.13	-0.66	-0.56	-0.43	0.16	-0.22	-0.35	-0.84	-0.07	-1.13	0.51	-0.70	-0.20	-0.30	-0.06	0.01	-0.34
52	1.60	0.01	0.55	-0.44	0.04	0.22	0.43	-0.17	-0.18	0.46	0.13	-0.37	0.22	-0.11	-0.04	-0.19	0.16	0.28
53	0.71	0.06	0.24	-0.31	0.20	0.01	0.18	-0.33	-0.44	-0.32	-0.10	0.00	0.03	0.04	-0.08	-0.21	0.12	0.20
54	1.06	-0.48	-0.15	-0.11	-0.01	-0.12	-0.09	-0.17	-0.32	-0.13	0.00	-0.11	-0.04	-0.30	-0.40	0.17	-0.31	0.13
55	0.83	0.12	0.16	0.79	0.72	0.95	0.10	0.43	1.03	1.75	1.05	0.36	1.01	0.64	0.66	1.31	0.34	0.14
56	0.22	0.00	-0.02	-0.02	-0.03	-0.09	-0.05	-0.37	-0.92	0.40	-0.09	0.05	0.16	-0.06	-0.06	0.69	-0.10	0.07
57	-0.24	-0.58	0.12	-0.10	-0.44	-1.02	0.04	-0.69	-0.03	-0.28	-0.65	-0.47	-0.06	-0.85	-0.22	0.94	-0.06	-0.35
58	1.36	0.32	0.43	0.29	0.52	0.89	-0.11	0.29	-0.35	-0.03	0.64	0.82	-0.49	0.00	0.53	0.19	-0.21	0.64
59	0.23	0.72	-0.14	0.56	0.10	0.46	-0.38	0.23	0.17	1.13	-0.29	0.31	0.25	0.19	1.12	0.59	0.14	-0.23
60	0.01	-0.86	-0.86	0.07	-0.20	-0.13	0.02	0.38	-1.23	-0.20	-0.78	-0.33	-0.01	0.05	-0.03	-0.96	-0.28	-0.50
61	-0.93	-0.55	0.43	0.32	-1.22	0.27	-0.15	0.90	0.41	0.52	0.35	0.10	-0.64	0.12	0.00	0.25	-0.45	-0.65
62	-0.65	-0.59	-0.58	-0.98	-0.30	0.38	0.20	-0.59	0.25	-0.06	-0.65	-0.60	-0.62	-0.42	-0.95	-0.52	0.25	-0.35
63	0.68	-0.52	-0.16	-1.12	0.45	-1.23	-0.24	-0.13	-0.16	0.18	-0.50	-0.33	0.08	-1.02	-0.95	-0.58	-0.33	-0.25

Table C.7: Estimated $\hat{\gamma}_{jk}$ parameters for $k = 1, ..., 18$ for Model (4.9) on the PISA 2018 math data.

Item Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	-0.04	-0.05	-0.12	0.07	-0.00	0.00	0.03	-0.12	-0.01	-0.05	0.03	-0.07	-0.16	-0.09	-0.26	-0.65	-0.14	-0.12
2	0.01	-0.15	-0.16	-0.23	-0.64	0.18	0.18	0.24	0.54	-0.05	0.27	0.06	0.29	0.16	-0.49	-0.26	-0.09	0.74
3	0.16	-0.07	0.12	0.31	-0.04	0.39	-0.15	0.28	-0.08	0.11	0.09	0.50	0.20	0.34	0.37	-0.18	0.27	-0.75
4	-0.36	-0.19	0.11	-0.01	-0.57	-0.18	0.18	0.05	0.37	0.19	-0.23	-0.26	0.22	-0.21	0.35	-0.69	-0.32	-0.39
5	0.09	0.10	0.17	-0.38	-0.67	-0.03	-0.17	0.04	0.11	-0.05	-0.08	-0.19	0.44	-0.00	0.34	0.34	0.67	-0.31
6	0.05	-0.06	0.00	-0.03	0.10	-0.59	-0.14	-0.33	-0.36	-0.73	-0.02	0.31	-0.17	-0.47	0.01	0.12	-0.27	-0.28
7	-0.32	-0.26	-0.18	-0.96	-0.80	-0.68	-0.45	-0.64	-0.21	-0.94	-0.45	-0.34	-0.31	-0.07	0.07	0.19	-0.78	0.38
8	0.07	0.00	0.01	0.59	0.48	0.50	-0.90	-0.12	-0.19	0.01	-0.14	0.34	0.65	0.00	-0.22	0.04	0.09	0.47
9	0.02	0.26	-0.09	0.15	0.09	0.05	0.25	0.22	0.39	-0.17	0.00	-0.10	0.27	-0.03	0.00	-0.16	-0.27	0.18
10	-0.17	-0.07	-0.03	-0.21	-0.18	0.06	0.18	0.15	0.06	-0.04	0.10	-0.14	0.21	-0.11	-0.08	-0.10	-0.05	0.21
11	-0.23	-0.09	-0.18	0.11	0.30	0.05	-0.36	-0.04	-0.40	-0.48	-0.16	-0.14	-0.21	0.17	-0.15	0.00	-0.20	-0.71
12	0.31	0.14	0.05	0.31	0.24	0.32	-0.29	0.00	0.04	-0.03	0.15	0.14	0.40	0.40	0.13	0.12	0.22	-0.11
13	-0.88	-0.63	-0.56	-0.32	-0.35	-1.15	0.12	-0.87	-0.55	-0.14	-0.23	-0.40	-0.09	0.07	-0.12	-1.17	-0.44	-0.78
14	0.18	0.01	-0.03	-0.15	-0.35	-0.15	0.07	-0.18	0.23	-0.26	0.34	-0.18	0.00	-0.22	0.09	-0.05	-0.48	0.12
15	0.23	0.09	-0.04	0.16	0.36	-0.05	-0.09	0.12	0.12	0.10	0.33	0.09	0.16	0.02	-0.08	0.18	0.79	-0.06
16	0.30	0.15	0.08	-0.00	0.20	0.29	-0.02	0.22	0.22	0.54	0.33	0.60	0.39	-0.45	0.00	-0.22	0.04	-0.17
17	0.28	0.43	0.22	0.18	0.09	0.54	0.04	0.25	0.15	-0.07	0.25	0.40	0.14	-0.01	0.02	0.54	0.18	-0.02
18	0.32	0.39	0.21	-0.11	-0.04	0.29	-0.17	0.22	-0.02	0.05	0.30	0.24	0.27	0.16	-0.05	0.13	0.03	0.26
19	-0.02	0.30	-0.12	0.01	-0.06	0.15	-0.02	-0.31	0.35	-0.11	0.03	-0.32	0.37	0.09	-0.18	-0.17	-0.07	0.42
20	0.00	-0.24	0.06	0.00	0.04	0.12	0.09	-0.27	0.21	0.11	-0.12	0.22	-0.15	-0.35	-0.03	-0.05	-0.10	0.08
21	-0.73	-0.54	-0.62	-0.41	-0.83	-1.15	-0.40	-1.43	-0.16	0.27	-0.08	-0.20	-0.53	-0.24	0.28	-0.35	-0.40	-0.91
22	-0.61	-0.50	0.03	-0.40	-0.72	-0.44	0.03	-0.66	-0.33	-0.40	-0.74	-0.26	-0.13	-0.85	0.14	-1.09	-0.61	-1.35
23	-0.83	-0.42	-0.08	0.58	-0.31	0.29	-0.01	0.90	0.14	-0.24	-0.81	0.29	0.43	0.11	0.11	-0.06	-0.35	0.08
24	-0.19	0.07	0.14	0.16	0.39	0.00	0.28	0.13	0.70	0.00	0.04	0.51	-0.27	0.05	0.05	0.07	0.27	-0.17
25	0.09	0.12	0.08	0.18	0.61	-0.01	0.06	0.24	-0.16	0.26	0.34	0.02	0.29	0.15	0.43	0.50	-0.11	-1.79
26	-0.28	-0.18	0.08	-0.14	-0.66	-0.02	-0.52	-0.79	0.06	-0.61	-0.12	-0.40	-0.14	0.02	-0.33	-0.98	0.08	-0.07
27	-0.33	0.37	0.26	0.17	0.33	0.91	0.31	1.04	1.30	0.46	0.00	0.84	1.40	1.15	0.19	0.22	0.97	1.20
28	0.54	0.27	-0.04	-0.16	0.21	0.40	0.76	0.58	0.28	0.45	0.28	0.20	0.26	-0.47	0.36	0.17	0.28	-0.64
29	-0.65	-0.11	-0.04	0.38	-0.27	-0.44	-0.41	-0.72	-1.02	-0.26	-0.35	-0.51	-0.04	-0.33	0.26	-0.26	-0.09	-0.19
30	0.32	0.43	0.06	0.71	0.63	0.29	0.06	-0.22	0.75	0.63	0.20	0.60	0.53	0.29	0.18	0.02	0.69	0.17
31	0.16	0.40	0.41	-0.11	0.38	0.44	1.44	0.43	1.30	0.30	0.51	0.45	-0.22	0.92	0.13	0.70	0.03	1.08
32	-0.37	-0.44	0.21	0.25	0.32	-0.05	-0.08	-0.48	-0.20	-0.18	0.07	-0.38	-0.13	-0.16	-0.17	-0.29	-0.57	-0.20
33	-0.63	-0.49	0.15	0.68	1.30	-0.17	0.01	-0.28	-0.29	-0.34	-0.40	0.23	-0.41	0.22	-0.37	0.36	0.13	0.57
34	-1.07	-0.48	-0.58	-0.27	-0.57	-0.33	-0.04	-0.30	-0.63	-0.43	-0.83	-0.64	-0.13	-0.13	-0.00	-0.16	-0.58	0.36
35	-0.59	-0.11	-0.15	-0.56	-0.89	-0.35	-0.35	0.29	-0.11	0.43	-0.20	-0.47	0.03	-0.57	0.33	-0.11	0.11	-2.37
36	0.20	0.50	0.17	0.24	0.19	0.43	0.54	0.40	0.08	0.23	0.22	-0.10	0.30	0.19	0.39	0.13	0.02	-0.43
37	-0.84	-0.09	-0.21	-0.15	-0.53	-0.47	0.20	-0.73	0.30	0.27	-0.41	-0.62	-0.40	-0.92	0.01	-0.09	-0.11	0.00
38	0.02	0.24	-0.31	-0.31	0.23	0.07	-0.00	0.06	0.00	0.22	0.42	0.00	0.15	-0.03	-0.22	0.63	-0.16	-0.60
39	0.04	0.11	-0.26	0.11	0.31	-0.00	0.09	0.02	-1.08	0.29	0.07	0.39	-0.21	0.17	-0.26	0.72	0.00	0.25
40	0.18	0.04	0.22	-0.18	-0.32	-0.00	0.30	0.45	0.01	0.40	0.12	0.10	0.13	0.11	-0.16	-0.31	0.13	0.50
41	-0.61	0.36	0.09	-0.06	-0.20	0.34	-0.62	0.34	-1.31	0.77	-0.62	0.33	-0.36	-0.59	-0.10	0.90	0.39	1.67
42	-0.28	-0.31	-0.34	-0.82	0.00	0.13	0.07	-0.07	-0.40	-0.16	-0.17	-1.03	-0.15	-0.27	-1.51	-0.18	-0.42	0.65
43	0.19	0.19	-0.09	-0.10	0.21	0.30	0.22	0.09	0.13	0.40	0.30	0.27	0.17	0.14	0.02	0.67	0.35	0.38

Table C.8: Estimated $\hat{\gamma}_{jk}$ parameters for $k = 19, ..., 36$ for Model (4.9) on the PISA 2018 math data.

Item Index	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	-0.24	-0.14	-0.09	-0.08	-0.08	0.19	-0.09	-0.08	-0.08	-0.03	-0.08	-0.13	-0.06	0.08	-0.04	-0.33	-0.11	-0.07
2	0.38	0.27	0.00	0.13	-0.24	-0.16	0.01	0.10	0.47	0.18	0.19	0.61	0.15	0.40	0.24	-0.17	-0.12	-0.22
3	0.13	0.09	-0.11	-0.01	-0.15	0.46	0.08	-0.14	-0.22	0.07	0.12	0.29	0.09	0.08	0.00	0.04	0.03	0.03
4	-0.26	0.12	-0.24	-0.04	-0.28	-0.63	-0.17	0.23	0.04	-0.29	-0.38	0.17	-0.17	0.39	-0.31	-0.13	0.02	0.03
5	-0.14	-0.00	0.26	0.03	-0.29	0.46	-0.16	0.09	-0.03	0.25	0.39	-0.06	0.57	0.03	-0.04	-0.38	0.36	0.01
6	-0.00	-0.26	-0.05	-0.12	-0.39	0.35	0.02	-0.17	-0.40	-0.12	-0.81	-0.07	-0.05	-0.34	-0.02	-0.55	0.06	0.00
7	-0.29	-0.98	-0.31	-0.36	-0.65	0.54	-0.06	0.28	-0.88	-0.64	-0.17	-0.71	-0.74	-0.13	-0.44	-0.54	-0.05	-0.03
8	0.39	0.00	0.32	0.21	0.05	-0.07	0.10	-0.17	0.01	0.10	0.08	0.29	0.42	-0.10	-0.19	0.11	0.03	0.16
9	-0.12	0.30	0.19	0.11	-0.04	0.54	0.15	0.21	0.17	-0.01	0.13	0.01	-0.09	0.17	0.00	-0.02	-0.05	-0.34
10	-0.01	0.00	0.15	-0.02	-0.30	0.09	-0.06	-0.06	-0.01	-0.29	-0.07	-0.04	-0.00	-0.06	-0.08	-0.19	-0.01	0.00
11	-0.27	0.04	0.12	-0.02	0.30	-0.05	-0.13	-0.29	-0.60	-0.11	-0.09	-0.48	-0.11	0.18	-0.38	-0.30	-0.01	0.10
12	0.20	0.27	0.27	0.38	0.17	0.13	-0.16	0.13	-0.23	0.08	0.43	-0.14	0.15	0.28	-0.03	0.00	0.25	0.08
13	-0.48	-0.59	-0.87	-0.65	0.17	-0.48	0.05	0.21	0.13	-0.68	-0.79	-0.53	-0.43	-0.39	-0.58	0.33	-0.34	-0.75
14	0.22	-0.10	-0.13	-0.09	-0.26	0.73	0.09	-0.03	-0.12	-0.04	-0.22	-0.01	-0.14	0.94	0.62	0.39	0.43	-0.14
15	-0.35	0.25	-0.11	0.30	0.04	-0.05	-0.06	-0.70	-0.12	0.54	-0.02	0.45	0.26	0.19	-0.03	0.22	-0.11	0.22
16	0.04	0.16	-0.01	0.23	0.18	0.35	-0.21	0.02	0.00	0.06	-0.09	-0.07	0.09	0.50	0.04	-0.00	0.15	0.10
17	0.64	0.04	0.57	0.22	0.28	0.96	-0.14	0.07	0.46	0.25	0.57	0.41	0.26	-0.04	0.32	0.15	-0.13	0.08
18	-0.07	0.02	0.26	0.16	0.04	0.54	-0.12	0.10	0.22	0.07	0.32	0.31	0.18	0.10	0.37	0.22	0.08	0.10
19	0.97	-0.10	-0.11	0.01	-0.23	0.95	-0.10	-0.37	0.00	-0.38	0.09	0.17	-0.17	-0.60	0.18	0.16	-0.04	-0.10
20	-0.17	0.11	0.08	-0.09	0.04	0.28	0.11	0.22	0.04	0.09	-0.13	0.07	0.02	-0.04	0.13	0.32	-0.15	-0.23
21	-0.62	-1.21	-1.35	-0.62	-0.37	-0.69	0.25	-0.52	-1.26	-0.44	-1.51	-1.13	0.13	0.06	-0.57	0.53	-0.23	-0.78
22	-1.93	-0.59	-0.43	-0.61	-0.60	-0.35	0.03	-0.14	-0.20	-0.39	-0.88	-0.16	-0.61	-0.19	-0.53	-0.29	-0.07	0.08
23	0.24	-0.06	0.20	-0.65	-0.08	-0.15	0.11	0.41	-0.08	-0.44	0.00	-0.61	-0.20	-0.14	-0.85	1.05	0.20	0.34
24	0.71	0.00	-0.08	-0.18	0.15	-0.01	0.17	0.36	-0.26	0.68	-0.38	-0.06	0.29	0.31	-0.29	0.32	-0.02	-0.18
25	-1.44	0.08	0.12	0.36	0.37	0.52	0.04	0.30	0.45	0.11	-0.06	-0.06	0.44	0.21	-0.05	0.28	0.15	-0.23
26	0.79	-0.01	0.26	-0.10	-0.25	-0.25	-0.38	-0.27	0.25	0.08	0.01	-0.00	-0.59	-0.29	0.50	0.86	0.00	-0.62
27	1.57	1.16	1.71	0.26	0.97	0.60	0.06	1.18	1.38	-0.19	0.72	1.52	0.78	0.80	0.13	1.87	0.06	0.26
28	-0.51	0.61	-0.15	0.43	0.39	-0.04	0.28	0.48	0.51	0.42	0.26	0.34	0.30	0.35	0.38	0.13	0.45	0.57
29	0.34	-0.28	-0.42	-0.41	0.11	-0.42	0.05	-0.35	0.55	0.33	0.02	0.30	0.00	-1.00	-0.60	-0.24	0.46	0.18
30	0.24	0.63	0.25	0.07	0.78	0.42	-0.04	0.20	0.58	0.37	0.29	0.00	0.54	0.45	0.02	0.99	0.17	0.38
31	0.47	0.65	0.53	0.38	0.22	0.99	0.02	0.85	0.41	-0.08	0.43	0.63	0.38	0.68	0.46	0.28	0.57	0.17
32	-0.27	-0.24	-0.46	-0.10	0.00	-0.17	0.20	-0.10	-0.58	-0.42	-0.35	-0.44	-0.01	-0.23	-0.35	-0.61	-0.33	0.24
33	0.71	-0.13	-0.41	-0.03	1.05	-0.15	-0.22	-0.10	-0.14	-0.02	0.13	0.11	-0.42	-0.01	0.01	0.73	-0.27	-0.28
34	0.69	-0.67	-0.24	-0.83	-0.37	-0.16	0.01	-0.14	-0.12	-0.03	-0.15	-0.24	-0.74	-0.12	-1.05	-0.07	0.27	-0.08
35	-0.02	-2.17	0.00	0.14	-0.44	-0.27	-0.51	-1.78	-0.24	0.35	-0.80	-0.06	0.21	-0.94	-0.62	-0.17	-0.06	-0.28
36	-0.38	0.18	0.39	0.00	0.32	1.14	-0.10	0.34	0.17	-0.18	0.42	0.33	0.23	0.30	0.10	0.03	0.29	-0.05
37	-1.13	-0.48	-0.29	-0.62	-0.40	0.38	0.18	-0.18	-0.55	0.39	-0.82	-0.74	-0.18	-0.46	-0.26	-1.34	0.26	0.12
38	-0.70	0.02	-0.03	0.22	-0.06	-0.07	0.17	-0.24	0.07	-0.27	0.43	0.02	-0.18	-0.11	0.27	-0.60	-0.25	-0.15
39	-0.12	-0.12	0.08	0.65	0.29	0.00	0.00	0.00	-0.08	0.00	0.00	0.00	0.14	0.00	0.31	-0.20	-0.13	-0.16
40	0.47	0.86	0.64	0.20	0.19	0.20	0.05	0.24	0.55	0.21	-0.33	0.06	0.00	0.38	0.53	-0.15	0.19	0.01
41	1.80	1.08	0.64	0.54	-0.22	-0.93	-0.11	0.20	0.22	0.93	0.11	-0.40	-0.91	-0.96	-0.09	0.17	0.31	0.15
42	0.21	0.28	-0.03	-0.36	0.45	-0.21	-0.05	0.52	0.20	-0.48	-0.19	0.24	-0.55	0.65	0.46	-0.05	-0.05	-0.52
43	0.35	0.08	-0.14	0.40	0.38	0.38	0.21	0.28	0.05	0.37	-0.04	0.22	0.17	0.15	0.15	0.02	0.11	-0.10

Table C.9: Estimated $\hat{\gamma}_{jk}$ parameters for $k = 1, ..., 18$ for Model (4.9) on the PISA 2018 science data.

Item Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0.09	-0.11	-0.09	-0.11	0.04	0.03	0.03	0.16	0.28	-0.28	0.17	0.03	0.24	-0.09	0.26	0.10	0.04	0.21
2	-0.38	-0.49	-0.10	-0.33	-0.52	-0.11	-0.12	-0.06	-0.14	-0.29	-0.14	-0.55	-0.64	-0.54	-0.12	-0.21	-0.42	-1.02
3	0.61	0.48	0.13	0.84	0.67	0.81	0.69	0.98	0.53	0.82	0.92	1.05	0.65	0.59	0.08	0.54	1.03	0.10
4	0.19	0.65	0.38	-0.41	-0.19	0.47	0.00	0.91	0.07	0.49	0.03	0.58	0.76	0.07	0.51	0.05	0.63	0.20
5	0.11	-0.65	-0.12	-0.56	-0.46	-0.11	-0.44	-0.14	0.92	-0.86	0.19	-0.72	-0.13	0.02	0.14	-0.03	-0.73	0.11
6	0.22	-0.27	-0.12	0.32	0.33	0.34	-0.37	-0.15	0.65	-0.26	0.20	0.32	0.42	-0.04	-0.13	-0.22	0.45	0.05
7	-0.20	-0.15	0.02	0.29	0.14	-0.29	-0.46	0.02	-0.40	0.01	-0.09	0.12	0.03	0.00	-0.17	-0.12	-0.06	-0.16
8	0.07	-0.03	0.03	0.08	0.16	-0.49	-0.30	-0.16	-0.41	0.45	-0.05	-0.20	0.26	-0.37	-0.39	-0.37	0.14	0.23
9	0.17	-0.11	0.07	-0.79	-0.92	0.10	0.46	-0.48	0.33	-0.07	0.01	-0.76	-0.50	0.06	0.30	-0.62	-0.36	0.38
10	-0.22	0.18	0.03	0.07	-0.16	0.46	0.35	-0.01	0.02	0.29	0.06	-0.31	0.44	-0.20	-0.55	0.10	0.06	0.10
11	0.02	-0.06	0.20	-0.18	-0.03	-0.18	0.04	-0.18	0.24	-0.14	0.02	-0.21	-0.14	0.05	-0.02	-0.22	-0.29	-0.19
12	0.04	0.09	0.07	-0.27	-0.16	-0.32	-0.29	0.37	0.34	-0.06	-0.01	0.19	0.02	-0.26	0.03	0.29	0.07	-0.05
13	-0.27	-0.16	-0.10	0.01	-0.15	-0.21	-0.05	-0.07	0.16	-0.17	-0.33	-0.25	0.10	-0.24	-0.22	-0.09	-0.22	-0.58
14	-0.09	-0.11	-0.05	-0.14	-0.05	0.13	-0.15	-0.21	-0.20	-0.28	-0.02	-0.13	0.03	-0.28	-0.25	-0.15	0.28	0.07
15	0.11	0.18	0.10	0.00	-0.06	-0.25	0.08	-0.32	-0.31	0.09	0.02	0.29	-0.03	0.24	0.09	0.00	0.08	0.06
16	0.23	0.26	-0.09	0.17	0.68	0.65	0.23	0.11	-0.03	-0.26	0.21	-0.09	0.29	0.43	0.22	-0.23	0.31	-0.43
17	-0.73	-0.12	-0.03	0.15	0.16	-0.83	-0.69	0.14	-0.36	-0.51	-0.59	-0.14	-0.27	-0.22	0.18	0.44	-0.42	-0.39
18	-0.11	-0.20	-0.28	-0.45	-0.34	0.23	-0.11	0.40	-0.06	-0.40	-0.08	0.29	0.43	-0.04	0.25	0.11	0.18	0.54
19	-0.05	0.20	0.02	0.08	0.01	0.53	0.09	0.55	-0.03	0.21	-0.00	0.49	0.34	0.36	0.13	-0.38	0.15	0.03
20	-0.06	-0.35	0.05	-0.43	0.08	-0.25	0.26	-0.18	0.22	-0.50	-0.02	-0.25	0.03	-0.38	0.17	-0.15	-0.06	-0.06
21	0.00	-0.09	0.29	-0.30	-0.16	-0.19	0.14	0.17	0.29	0.11	-0.30	-0.08	-0.17	0.47	0.07	0.14	-0.27	0.12
22	0.83	0.40	-0.03	0.20	-0.16	0.23	0.14	0.19	0.65	0.23	0.57	0.34	0.37	1.03	0.16	0.08	0.20	0.74
23	-0.73	-0.03	0.21	-0.23	0.13	-0.19	-0.35	-0.18	-0.27	0.09	-0.40	-0.08	-0.14	-1.81	0.04	0.05	-0.67	-0.21
24	-0.10	0.16	0.32	0.63	0.42	-0.78	-0.75	-0.12	0.12	-0.00	-0.06	-0.90	-0.39	0.22	0.24	-0.25	-0.30	-0.22
25	0.03	0.27	-0.19	-0.11	0.15	0.49	0.40	0.27	0.33	0.28	0.02	0.08	0.36	0.05	-0.01	-0.36	0.38	-0.45
26	-0.49	-0.67	-0.42	-0.25	-0.22	-0.38	-0.64	-0.23	0.23	-0.96	-0.51	-0.38	-0.27	-0.24	0.38	-0.29	-0.58	0.57
27	0.45	0.27	-0.13	0.54	0.14	0.39	0.20	0.17	0.50	0.08	0.51	0.24	0.00	0.56	-0.17	0.53	0.51	0.17
28	0.26	-0.01	-0.01	-0.74	-0.85	0.53	0.37	0.21	0.60	-0.38	0.23	-0.13	-0.18	0.08	0.57	0.25	0.37	-0.27
29	0.20	0.23	0.06	0.11	-0.30	-0.02	0.40	0.50	0.39	0.15	0.04	0.05	0.10	0.20	-0.05	-0.64	0.48	1.06
30	0.93	0.52	0.09	0.32	0.20	0.48	0.25	0.31	0.14	0.49	0.67	0.22	0.96	-0.31	0.00	0.03	0.07	0.45
31	0.74	0.53	-0.14	0.20	0.22	0.10	0.41	0.53	0.30	0.69	0.52	0.50	0.13	0.01	-0.07	0.19	0.18	1.47
32	-0.06	0.12	-0.06	0.22	0.32	0.07	-0.10	0.17	0.26	0.46	0.11	0.32	0.05	0.11	0.25	0.00	0.04	0.00
33	-0.50	-0.87	-0.36	0.43	0.73	-0.09	-0.76	-0.48	0.21	-0.89	-0.66	-0.18	-0.31	-0.08	-0.21	0.15	-0.81	-0.39
34	-0.25	-0.35	-0.19	-0.22	-0.29	-0.52	-0.53	-0.65	-0.58	-0.28	-0.39	-0.39	-0.45	-0.47	-0.01	-0.19	-0.69	-0.47
35	-0.06	0.13	-0.08	0.16	0.26	0.03	-0.08	0.01	0.11	-0.01	-0.02	-0.30	-0.20	0.27	0.06	0.60	0.00	-0.32
36	0.21	-0.17	-0.06	0.57	-0.03	0.38	-0.32	-0.31	0.21	-0.19	0.23	0.66	-0.32	0.48	-0.50	-0.52	-0.10	1.26
37	-0.37	-0.42	-0.02	0.02	0.19	-0.36	-0.10	-0.60	-0.13	-0.27	-0.43	-0.07	-0.21	-0.13	-0.04	-0.35	-0.06	-0.51
38	-0.24	0.01	-0.18	0.11	0.33	0.12	-0.35	-0.51	-0.41	-0.06	-0.06	0.19	-0.21	0.10	-0.09	0.27	-0.01	0.28
39	-0.49	-0.82	-0.17	0.03	-0.09	-0.18	-0.48	-0.06	-0.26	-0.56	-0.39	0.12	-0.38	-0.70	-0.14	-0.24	-0.08	-0.52
40	-0.30	-0.44	0.03	-0.41	-0.97	-0.33	-0.05	-0.31	-0.47	-0.13	-0.23	-0.54	-0.43	-0.10	0.18	-0.50	-0.07	-0.01
41	0.13	-0.45	0.01	-0.31	-0.53	-0.15	-0.02	-0.74	-0.16	0.09	0.16	-0.36	0.18	-1.02	0.32	-0.73	0.57	-0.14
42	0.14	0.15	-0.04	0.35	0.20	0.82	0.02	0.69	-0.22	0.01	0.20	0.96	1.00	0.10	-0.33	0.92	1.24	0.30
43	0.22	-0.09	-0.24	0.13	-0.06	-0.11	-0.23	0.05	-0.34	0.00	0.00	-0.12	0.02	-0.16	-0.10	0.05	-0.10	0.03
44	-0.13	0.10	0.03	-0.21	-0.16	-0.37	0.27	0.08	-0.31	0.38	0.00	0.28	-0.07	0.26	-0.25	-0.15	0.07	-0.68
45	0.65	0.24	0.09	-0.01	0.00	0.15	-0.02	0.13	0.40	0.23	0.52	-0.06	0.34	-0.02	-0.10	0.26	-0.29	0.30
46	-0.01	-0.22	-0.19	-0.46	-0.96	0.33	0.21	0.40	0.07	-0.63	-0.09	0.08	0.07	-0.41	-0.10	0.05	-0.21	0.29
47	-0.09	-0.09	-0.16	-0.49	-0.09	-0.48	0.28	-0.42	-0.10	-0.47	-0.25	-0.41	-0.03	-0.29	0.07	-0.25	-0.44	-0.20
48	-0.09	0.39	0.05	0.14	-0.16	0.29	-0.81	-0.09	-0.70	0.30	-0.27	-0.51	-0.02	0.28	-0.20	-0.50	-0.10	0.09
49	0.08	-0.23	0.05	0.22	0.09	-0.28	-0.11	-0.12	-0.19	-0.06	-0.07	-0.15	0.10	-0.27	0.20	-0.43	-0.17	-0.65
50	-0.03	0.31	-0.11	0.29	0.21	0.03	0.11	-0.13	-0.05	0.11	-0.01	-0.24	0.42	-0.22	-0.03	0.28	0.33	0.22
51	0.08	0.27	0.05	-0.51	-0.78	0.32	0.18	-0.18	0.66	0.19	-0.06	-0.01	-0.19	-0.34	0.01	-0.48	0.07	-0.05
52	0.44	-0.10	0.08	-0.37	-0.21	0.47	0.23	0.58	0.53	0.02	0.37	0.24	0.44	0.39	0.13	0.04	0.24	-0.19
53	0.12	0.18	0.08	0.11	0.34	-0.19	0.01	-0.21	0.29	0.11	0.26	-0.24	-0.13	0.22	0.09	0.14	0.25	-0.19
54	-0.43	0.04	0.25	-0.59	0.05	-0.29	0.10	-0.05	-0.30	-0.00	-0.13	-0.29	-0.06	0.09	0.05	-0.25	-0.42	-0.44
55	-0.21	-0.06	-0.01	-0.33	0.06	0.00	-0.63	-0.26	-0.01	-0.07	-0.11	0.17	0.24	-0.79	-0.03	0.00	-0.17	0.28
56	0.02	0.18	0.09	0.19	0.13	0.07	0.24	0.09	0.21	-0.15	0.14	-0.00	-0.09	0.41	-0.03	-0.28	-0.03	0.27
57	-0.19	-0.03	0.10	-0.04	0.04	-0.13	-0.07	-0.49	-0.28	-0.00	0.15	-0.55	-0.25	-0.40	0.34	-0.45	0.06	-0.06
58	0.16	0.22	0.06	0.20	-0.14	0.01	0.05	-0.09	0.18	0.11	0.19	-0.03	0.15	0.14	0.32	0.02	0.02	0.27
59	0.37	-0.19	0.31	0.58	0.38	-0.72	0.15	-0.14	0.39	0.42	-0.03	-0.36	-0.23	0.20	0.09	-0.19	-0.46	0.69
60	0.10	0.16	-0.12	0.26	0.19	0.27	-0.26	0.17	-0.12	0.84	0.18	0.32	0.22	0.26	-0.02	0.02	0.52	0.19
61	0.65	0.39	-0.15	-0.10	-0.01	0.12	0.48	0.40	0.53	0.40	0.51	0.52	0.32	0.10	0.02	0.30	0.24	-0.14
62	-0.05	-0.44	-0.36	-0.51	0.09	-0.34	-0.15	-0.32	-0.52	-0.35	0.01	0.07	-0.41	-1.38	0.19	-0.21	0.24	-0.70
63	0.75	-0.13	0.04	0.43	0.77	0.10	-0.15	-0.06	-0.12	0.01	0.72	0.33	-0.08	-0.20	-0.04	-0.07	0.41	-0.58
64	-0.42	0.23	-0.31	0.46	0.73	-0.18	0.36	0.12	0.02	0.41	-0.73	0.27	0.17	0.04	-0.15	-0.02	0.53	0.54
65	-0.13	-0.37	-0.03	-0.42	-0.48	0.00	0.20	-0.19	0.57	-0.34	-0.23	0.37	-0.15	-0.41	-0.23	0.14	0.03	0.62
66	-0.34	-0.32	-0.20	-0.10	-0.39	-0.43	-0.29	-0.31	-0.24	-0.34	-0.14	-0.32	-0.17	-0.66	-0.20	-0.35	-0.43	-0.52
67	-0.24	-0.47	-0.13	-0.32	0.16	-0.62	0.19	-0.15	-0.11	-0.65	-0.40	-0.23	-0.26	-0.76	-0.04	-0.03	-0.13	-0.01
68	-0.15	-0.33	0.09	-0.00	-0.50	-0.93	-0.82	-0.49	-0.60	-0.35	-0.52	0.05	-0.44	-0.80	-0.02	-0.17	-0.55	-0.82
69	-1.25	0.23	-0.07	-0.49	-1.23	-1.05	-0.99	-0.58	-1.52	-0.63	-1.22	-0.48	-0.23	-0.71	-0.55	0.24	-1.03	0.28
70	0.25	0.00	0.00	-0.21	-0.14	0.15	0.14	0.22	0.68	-0.05	0.27	0.29	0.52	-0.51	0.17	-0.20	0.20	-0.12
71	0.06	-0.38	-0.18	0.25	0.07	0.02	-0.22	-1.03	-0.49	-0.72	-0.08	-0.13	-0.42	-0.16	0.20	-0.10	-0.62	-0.36
72	-0.01	0.10	-0.08	-1.09	-0.56	-1.28	0.01	0.00	-0.25	-0.33	0.02	-0.42	-0.56	-0.68	0.23	0.44	0.19	-0.22
73	0.41	0.43	0.13	0.16	0.35	0.53	0.01	0.58	0.41	0.27	0.14	0.58	0.76	0.15	-0.29	0.53	0.35	0.66
74	0.15	-0.06	0.16	0.03	-0.01	0.46	0.22	0.42	0.00	0.14	-0.05	0.16	0.56	0.28	0.06	0.09	-0.16	-0.43
75	-0.14	0.27	0.05	-0.28	-0.01	-0.38	-0.06	0.08	-0.11	0.43	0.05	0.16	0.02	0.20	-0.27	0.37	-0.17	0.55
76	-0.01	0.04	-0.09	-0.25	-0.07	0.10	-0.41	0.23	-0.30	0.16	-0.37	0.15	0.10	-0.20	-0.33	-0.16	-0.50	0.24

Table C.10: Estimated $\hat{\gamma}_{jk}$ parameters for $k = 19, ..., 36$ for Model (4.9) on the PISA 2018 science data.

Item Index	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	0.14	0.09	0.41	-0.11	-0.39	-0.25	-0.06	0.04	0.07	0.38	0.08	0.21	-0.07	-0.17	-0.11	-0.11	0.16	0.00
2	-1.07	-0.05	0.11	-0.54	-1.28	-0.25	0.09	-0.14	0.25	-0.21	-0.27	0.06	-0.35	-0.23	-0.45	-0.77	-0.17	-0.01
3	1.21	0.90	0.95	0.69	0.73	0.28	-0.07	1.14	0.68	0.39	0.23	0.68	0.65	0.68	0.75	0.20	-0.29	-0.22
4	0.46	0.67	0.82	0.54	0.00	0.35	-0.31	0.00	0.55	0.29	0.81	0.77	0.40	-0.23	0.29	1.08	0.11	0.16
5	-0.46	-0.18	-0.19	-0.24	-0.12	-0.34	0.10	-0.15	-0.29	-1.18	-0.15	-0.36	-0.35	-0.08	-0.32	-0.22	0.19	-0.29
6	0.34	0.18	-0.21	-0.23	0.25	0.17	-0.25	-0.00	0.49	0.07	0.42	0.32	-0.14	-0.12	-0.29	0.24	-0.12	-0.42
7	-0.41	-0.33	0.16	-0.18	0.04	-0.09	-0.26	-0.18	-0.21	0.00	0.04	-0.25	0.19	-0.54	-0.12	-0.02	-0.31	0.09
8	-0.42	-0.49	-0.34	0.00	0.26	-0.22	-0.06	-0.02	-0.22	-0.34	-0.19	-0.04	-0.18	-0.36	-0.05	0.18	-0.59	0.21
9	-0.18	0.07	0.00	-0.28	-0.44	0.28	0.16	0.12	-0.04	-0.68	0.15	-0.12	-0.45	0.06	0.21	-1.32	0.05	-0.07
10	0.03	0.36	0.52	-0.01	0.07	0.15	-0.10	0.50	0.40	0.23	0.22	0.33	0.05	-0.01	0.21	0.03	-0.61	0.30
11	-0.71	-0.33	-0.21	0.01	-0.20	-0.08	0.18	-0.09	-0.30	-0.18	-0.12	-0.44	-0.24	-0.18	0.26	-0.17	-0.10	-0.00
12	-0.49	0.09	-0.13	0.14	-0.63	-0.19	0.14	-0.33	0.19	-0.32	-0.04	0.05	-0.23	-0.09	0.05	0.13	0.05	-0.08
13	-0.38	0.02	0.06	-0.15	-0.21	-0.09	-0.05	-0.15	-0.41	-0.07	-0.19	-0.14	-0.00	0.14	-0.17	0.01	-0.26	-0.13
14	0.13	-0.25	-0.12	-0.23	-0.02	0.12	0.00	-0.14	-0.39	0.20	-0.08	-0.14	-0.03	-0.22	-0.03	0.06	-0.35	-0.11
15	0.17	-0.19	-0.43	0.23	0.11	0.26	-0.06	0.17	0.02	-0.25	0.14	-0.07	-0.07	0.31	0.07	0.24	-0.12	0.11
16	-0.54	0.21	0.39	-0.08	0.72	0.24	0.13	0.38	0.35	0.07	0.20	0.20	0.48	-0.07	0.04	-0.01	0.17	-0.01
17	-0.75	0.09	-0.25	-0.46	0.10	-0.02	0.16	0.01	0.06	-0.09	-0.34	-0.08	0.17	-0.05	-0.56	0.31	-0.03	-0.26
18	0.67	-0.23	-0.23	0.12	-0.05	0.36	-0.17	0.03	0.70	0.29	-0.20	-0.53	-0.01	-0.48	0.04	-0.64	0.22	-0.20
19	-0.30	0.35	0.13	0.07	-0.10	0.05	-0.19	0.51	0.37	0.38	0.37	0.59	0.44	0.27	0.00	0.06	0.05	0.00
20	-0.14	0.31	-0.19	0.05	-0.22	0.09	-0.10	0.35	0.06	-0.20	-0.19	0.02	-0.04	0.31	-0.16	0.12	-0.02	-0.00
21	-0.47	-0.40	-0.02	-0.00	-0.20	-0.16	0.09	0.40	0.06	0.06	0.29	0.00	-0.11	0.22	0.11	-0.52	0.13	0.10
22	0.15	0.05	-0.18	0.40	-0.01	0.92	-0.25	0.52	-0.01	0.29	0.52	0.46	0.30	0.58	0.40	0.46	0.08	0.20
23	-0.34	-0.05	-0.05	-0.58	0.18	0.14	-0.02	-0.35	-0.24	-0.10	-0.40	-0.44	-0.11	0.04	-0.29	-0.55	0.12	0.10
24	-0.68	-0.51	0.18	-0.13	0.37	-0.29	-0.13	-0.09	-0.70	0.19	-0.10	-0.31	0.45	0.26	-0.14	-0.28	0.12	-0.29
25	-0.23	0.00	0.22	-0.03	0.14	0.02	-0.04	0.43	0.43	0.46	-0.04	0.41	0.20	0.53	0.14	0.05	-0.19	-0.20
26	0.35	-0.37	-0.25	-0.55	-0.25	-0.64	-0.12	-0.35	-0.39	-0.82	-0.62	-0.28	-0.99	-0.26	-0.25	-0.72	0.09	0.03
27	-0.40	-0.03	0.76	0.51	-0.26	0.31	0.07	0.40	0.24	-0.03	0.02	-0.33	0.25	0.25	0.30	0.06	-0.24	-0.10
28	0.29	0.11	0.49	0.16	-0.40	0.63	0.03	0.46	0.39	0.29	0.51	0.67	0.14	0.83	0.27	0.80	0.62	-0.03
29	0.54	0.48	-0.32	-0.19	-0.08	-0.05	0.11	0.05	0.74	0.13	0.16	-0.14	0.10	-0.31	-0.21	-0.82	-0.02	-0.24
30	-0.38	0.44	0.69	0.23	0.27	0.25	0.03	0.10	0.38	0.11	0.59	0.58	0.52	0.42	0.61	0.44	0.03	0.04
31	0.66	0.36	0.47	0.50	0.82	0.39	-0.00	0.33	0.45	0.72	0.26	0.32	0.18	0.46	0.26	0.00	0.08	-0.05
32	0.06	0.24	0.37	0.43	0.29	-0.26	0.22	0.10	0.26	0.24	-0.25	0.10	0.24	0.08	-0.00	0.28	0.11	-0.02
33	-0.20	0.12	0.38	-0.58	0.22	-0.65	-0.21	-0.23	0.26	-0.05	-0.02	-0.02	-0.37	-0.37	-0.89	1.13	-0.15	0.01
34	0.19	-0.12	-0.44	-0.21	-0.29	-0.35	-0.07	-0.47	-0.41	-0.22	-0.18	-0.11	-0.15	-0.20	-0.38	-0.32	-0.09	0.08
35	-0.60	0.05	0.33	0.06	-0.12	0.71	0.09	-0.08	0.00	-0.31	-0.22	0.06	0.03	0.27	-0.19	0.33	0.00	-0.07
36	1.11	-0.43	-0.85	0.23	0.32	1.08	-0.14	-0.12	-0.71	0.05	-0.13	-0.03	0.13	0.10	0.06	-0.63	-0.24	0.11
37	-0.87	-0.19	-0.36	-0.34	0.31	-0.59	0.10	-0.12	-0.04	-0.29	-0.28	-0.38	-0.25	0.10	-0.50	-0.10	-0.10	0.09
38	0.13	-0.38	-0.20	0.02	0.14	-0.61	-0.12	0.03	0.00	-0.15	0.15	-0.50	0.09	-0.10	-0.22	0.08	-0.23	-0.10
39	-0.91	-0.22	-0.27	-0.30	-0.25	-0.54	0.01	0.02	0.05	-0.54	-0.06	-0.65	-0.20	0.13	-0.33	-0.31	-0.22	-0.00
40	0.05	-0.40	-0.01	-0.73	-0.54	0.06	-0.26	0.40	-0.47	0.38	-0.28	0.04	-0.06	0.40	-0.44	-0.40	-0.19	0.09
41	-1.14	-0.22	-0.26	-0.11	-0.55	-0.22	0.05	-0.12	-0.12	-0.02	-0.33	0.02	-0.25	-0.05	-0.04	-0.29	0.05	-0.11
42	0.08	0.27	0.54	0.26	0.28	0.14	-0.19	0.10	0.90	0.59	1.22	0.52	0.67	1.10	0.35	1.32	-0.24	-0.39
43	0.22	-0.30	0.15	-0.01	-0.07	-0.57	0.41	-0.05	0.00	-0.19	-0.28	0.37	-0.30	0.06	0.13	-0.27	-0.01	-0.16
44	-1.10	-0.35	-0.16	0.34	0.13	-0.26	0.00	0.15	-0.60	-0.13	0.05	-0.48	0.10	-0.05	0.29	-0.20	-0.20	0.06
45	0.28	0.28	1.03	0.27	0.24	0.09	0.16	0.54	0.08	-0.25	0.05	0.05	0.12	0.21	0.26	0.62	-0.10	0.16
46	-0.07	0.18	0.23	-0.28	-0.67	0.33	-0.16	-0.03	-0.07	-0.21	0.00	0.29	-0.27	0.00	-0.19	0.41	-0.39	-0.25
47	-0.61	-0.49	-0.16	-0.52	0.16	0.21	-0.13	-0.26	-0.23	-0.14	-0.55	-0.27	-0.29	0.02	-0.45	-0.20	-0.07	0.18
48	0.25	0.05	-0.70	-0.06	-0.01	0.59	-0.18	-0.24	-0.29	0.69	-0.36	-0.17	0.26	-0.07	-0.14	-0.40	0.03	0.24
49	-0.75	0.14	0.22	0.05	0.25	-0.32	0.12	-0.05	-0.07	-0.05	-0.26	-0.11	-0.02	-0.23	-0.12	0.08	0.04	0.08
50	-0.05	0.19	-0.15	-0.12	0.14	0.22	-0.03	0.06	0.35	0.32	0.62	0.14	0.35	-0.08	0.13	0.13	-0.06	-0.24
51	-0.07	-0.72	-0.44	0.05	-0.06	-0.02	-0.00	-0.00	-0.12	-0.33	0.18	-0.07	-0.23	-0.14	-0.02	-0.31	0.08	-0.21
52	0.13	0.48	0.58	0.34	-0.35	-0.25	0.13	0.23	0.44	-0.06	0.13	0.64	0.17	0.19	0.33	0.65	-0.11	0.07
53	0.53	-0.27	0.31	0.14	0.42	0.04	-0.01	0.26	-0.18	0.22	-0.45	-0.53	0.31	0.02	0.06	0.09	0.34	-0.05
54	0.00	-0.17	0.04	0.02	0.02	0.00	0.23	-0.20	-0.23	-0.29	-0.30	-0.33	-0.19	-0.21	-0.13	-0.19	0.14	0.14
55	0.29	-0.24	-0.63	-0.16	-0.23	-0.38	-0.07	-0.10	-0.25	0.12	0.22	-0.08	-0.13	-0.12	-0.05	0.13	0.24	0.08
56	0.10	0.06	0.19	0.06	-0.03	0.09	0.19	0.27	-0.25	0.20	-0.01	0.14	0.22	0.23	0.24	-0.24	0.19	-0.12
57	0.02	-0.59	-0.04	0.15	-0.06	0.11	-0.00	-0.25	-0.38	-0.05	-0.06	-0.03	0.12	0.09	-0.20	-0.13	-0.31	0.30
58	-0.27	-0.04	0.23	0.25	-0.38	0.49	-0.11	0.29	-0.18	0.13	0.20	0.23	0.30	0.05	0.26	0.46	0.01	-0.03
59	0.09	-0.47	-0.46	-0.17	0.72	-0.52	0.00	0.43	-0.44	0.01	-0.20	-0.04	0.55	0.47	0.09	0.15	-0.34	-0.15
60	0.00	0.19	0.00	0.27	-0.24	0.07	0.06	-0.18	0.42	0.44	0.42	0.41	0.38	-0.66	0.17	-0.11	0.37	-0.03
61	0.02	-0.23	-0.06	0.48	0.25	0.33	0.10	0.59	0.19	0.41	0.45	0.47	0.33	0.63	0.33	-0.70	0.54	-0.02
62	-1.18	-0.55	0.08	-0.24	0.26	-0.29	0.08	-0.85	-0.21	-0.35	-0.12	0.10	-0.10	-0.27	0.03	-0.19	0.10	-0.04
63	-0.51	0.19	0.07	0.16	0.11	-0.22	0.25	-0.20	-0.01	0.43	0.13	0.11	0.54	-0.23	0.42	-0.74	0.00	-0.07
64	0.48	0.12	0.14	-0.26	0.45	0.06	-0.11	-0.22	0.29	0.02	0.00	0.24	0.33	0.48	-0.10	0.90	-0.20	-0.03
65	1.46	0.03	0.27	-0.50	-0.29	-0.03	0.06	0.01	0.05	0.16	0.06	0.69	-0.46	-0.11	0.04	0.64	-0.16	-0.08
66	-0.69	-0.31	-0.45	-0.47	-0.19	-0.22	0.32	-0.44	-0.47	-0.14	0.04	-0.16	-0.21	-0.30	-0.08	-0.08	0.04	-0.24
67	0.09	-0.77	-0.09	-0.24	-0.41	-0.14	0.01	-0.35	-0.04	-0.07	-0.58	-0.52	-0.00	-0.36	-0.29	0.39	0.08	0.05
68	-0.29	-0.20	-0.01	-0.42	-0.32	-0.75	0.00	-0.11	-0.57	0.26	-1.35	0.33	-0.27	-0.68	-0.44	-0.50	0.44	0.77
69	-0.58	-0.62	-0.91	-0.91	-0.97	-0.31	-0.18	-0.73	-0.60	-0.23	-1.32	-0.67	-0.29	-0.67	-0.79	-0.80	0.14	0.45
70	0.52	0.25	0.35	0.01	-0.10	0.29	0.18	0.01	-0.37	0.01	-0.03	0.05	-0.06	0.12	0.12	-0.32	-0.41	0.08
71	-0.42	-0.80	-0.36	-0.51	0.32	0.23	0.31	-0.22	-0.53	-0.49	-0.33	-0.97	-0.42	-0.40	-0.43	-0.38	-0.11	-0.00
72	-0.69	-0.81	-0.66	0.00	-0.67	0.04	0.01	-0.64	-0.33	-0.23	-0.77	-0.25	-0.56	-0.11	0.07	-0.25	0.31	0.07
73	0.67	0.25	0.16	0.30	0.17	0.59	-0.17	0.29	0.25	0.34	0.78	0.68	-0.05	0.48	0.65	0.42	-0.34	0.10
74	0.42	0.25	0.13	0.07	0.11	0.33	-0.07	0.31	0.11	-0.01	0.41	0.40	-0.12	-0.22	0.10	0.28	-0.13	0.25
75	0.17	0.24	-0.05	0.19	0.03	-0.07	-0.07	0.10	0.06	0.31	0.05	-0.14	0.00	0.15	0.06	0.34	0.20	0.07
76	0.10	-0.13	0.06	-0.17	0.10	-0.03	-0.11	-0.13	0.14	0.06	0.48	-0.11	0.06	0.12	-0.04	0.26	-0.05	0.00

APPENDIX D

Appendix of Chapter 5

Appendix D includes additional simulation results validating the inter- and intra-table inference procedures and the proofs for the theoretical results developed in Chapter V.

D.1 Additional Simulation Results

This section includes additional simulation results for validating inference procedures developed in Theorem V.11 for inter-table merging and Theorem V.12 for intra-table merging.

D.1.1 Inter-Table Merging

We seek to numerically examine the results given in Theorem V.11. We consider the same setting as in Chapter V, $H_0 : P_1 = 0.1, P_2 = 0.1, P_3 = 0.8$, and $D_j \sim$

Multinomial$(n_j, P_1 = 0.1, P_2 = 0.1, P_3 = 0.8)$ for $j = 1, 2$ and $n_1/n_2 = 3/7$. Consider merging two tables D_1^* and D_2^*, with total counts n_1 and n_2 respectively. Let the merged private table $D^* = D_1^* + D_2^*$ and consider sample sizes: $n = n_1 + n_2 = 100, 1000$. Setting the significance level to be 0.05, we evaluate 500 empirical test statistics. For private data obtained by five DP methods, Opt, Lap, TLap, GDP and TGDP, we check whether the empirical type I errors can be controlled as indicated by the results in Theorem V.11. The reported average empirical type I error rates under the "Inter-Table Merging" scenario in Table D.1 are controlled fairly well at around 5% for all the mechanisms. We note that even for small sample sizes of $n_1 = 30$ and $n_2 = 70$ and a high privacy requirement of $\epsilon = 0.25$, the type I errors can be well controlled.

		$n = 100$			$n = 1000$		
		$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
Inter-Table Merging	Opt	0.040	0.036	0.054	0.050	0.050	0.052
	Lap	0.056	0.042	0.050	0.060	0.044	0.050
	TLap	0.024	0.026	0.030	0.060	0.044	0.050
	GDP	0.042	0.052	0.054	0.034	0.040	0.042
	TGDP	0.020	0.018	0.024	0.034	0.040	0.038
Intra-Table Merging	Opt	0.031	0.050	0.052	0.051	0.051	0.053
	Lap	0.058	0.052	0.051	0.052	0.050	0.053
	TLap	0.037	0.049	0.050	0.052	0.050	0.053
	GDP	0.051	0.051	0.052	0.050	0.050	0.051
	TGDP	0.026	0.037	0.047	0.050	0.050	0.051

Table D.1: Mean empirical type I errors out of 500 simulated samples under two study scenarios across different sample sizes $n = 100, 1000$ and different privacy regimes $\epsilon = 0.25, 0.5, 0.75$. The two scenarios are "Inter-Table Merging": the inference methods proposed for the merging circumstances considered in Section 5.4.1 of Chapter V and "Intra-Table Merging": the inference methods proposed for the merging circumstances considered in Section 5.4.2 of Chapter V.

Similarly, empirical powers are evaluated and compared. The results are summarized in Figure D.1 below. Overall, the testing procedures proposed in Section 5.4.1

of Chapter V for the Opt and Lap methods yield good statistical power when we merge the tables. Furthermore, we note that when sample sizes and ϵ are small ($n_1 + n_2 = 100$, $\epsilon = 0.25$), the Opt procedure yields visibly better power than the field standard Lap/TLap mechanisms. In practice, data sets with smaller sample sizes are prevalent and, more importantly, they are more prone to privacy risks and the possibility of being merged is high, and as such might require a smaller and stricter ϵ. The Opt procedure stands out in these settings.

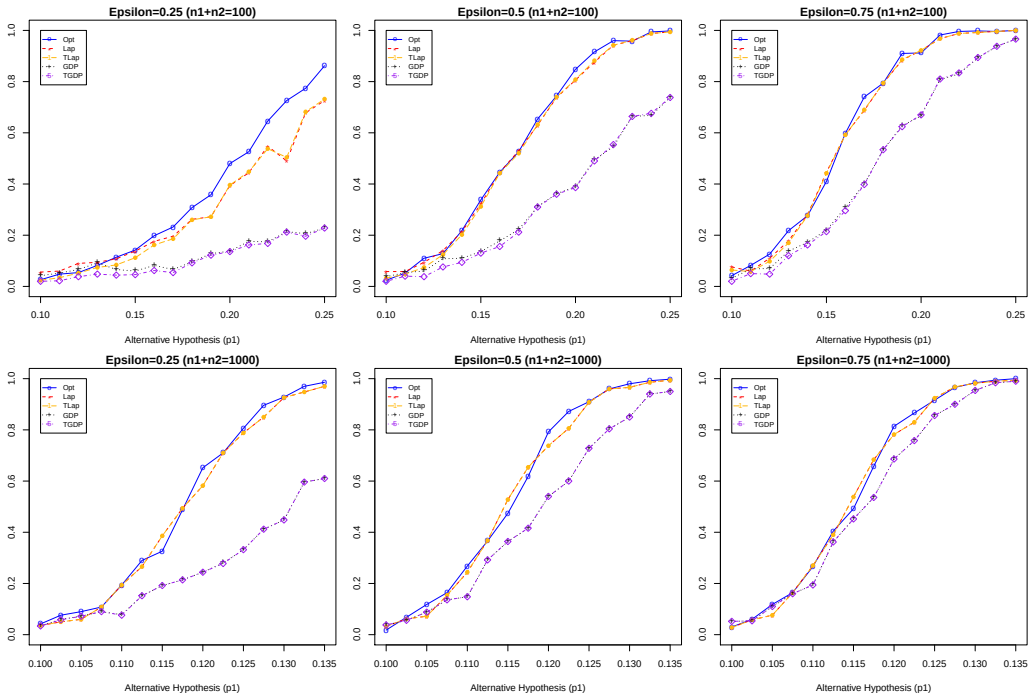

Figure D.1: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP, on the inter-table merged private data sets.

D.1.2 Intra-Table Merging

Next, we examine the numerical performance of the procedures presented in Section 5.4.2 of Chapter V. Assuming $D \sim \text{Multinomial}(n, P_1 = 0.1, P_2 = 0.1, P_3 = 0.8)$, we merge the first two cells of D^* and consider the corresponding goodness-of-fit test of $H_0 : P_{m1} = 0.2, P_{m2} = 0.8$. Setting the significance level to be 0.05 and using 500 each per generated data sets with $n = 100$ and $n = 1000$, we check whether the empirical type I errors can be controlled for the proposed procedures applying to the corresponding D^*. From the results reported in Table D.1 "Intra-Table Merging" scenario, we see that the empirical type I errors are controlled fairly well at around 5% for all settings.

Empirical powers are also evaluated by considering alternative hypotheses $H_1 : P_{m1} = p_{m1} = p_1 + p_2, P_{m2} = 1 - p_{m1}$. We explore the cases with $p_1 = p_2 = 0.1, 0.11, ..., 0.25$ for $n = 100$ and $p_1 = p_2 = 0.1, 0.1025, , ..., 0.135$ for $n = 1000$. The results are summarized in Figure D.2. We observe small improvements from the Opt procedure over the Lap/TLap procedure when the sample sizes are small at $n = 100$ and the privacy requirement is high at $\epsilon = 0.25$; and the statistical powers on private data generated from the Opt and Lap/TLap are superior to those from the GDP/TGDP.

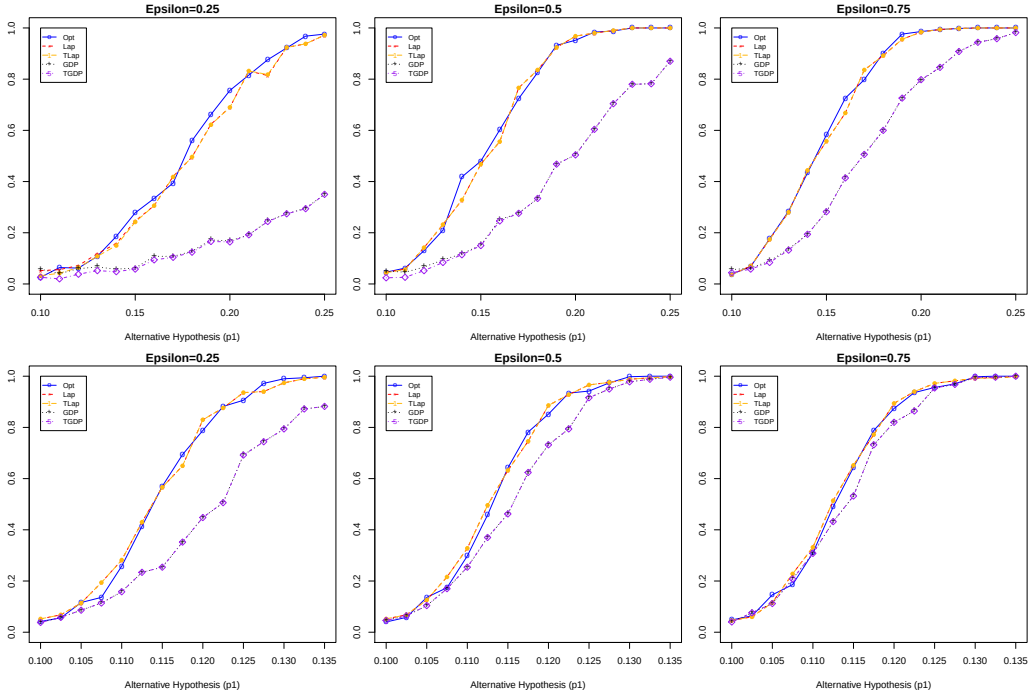

Figure D.2: Empirical power comparisons for five privacy procedures: Opt, Lap, TLap, GDP and TGDP, on the intra-table merged private data sets.

D.2 Proofs of the Theoretical Results

We first give four supporting lemmas that will be used in the proofs of the theorems.

Lemma D.1. *For any $\epsilon_n > 0$, if $\tilde{e}_i = r - i$ for $r = 0, ..., n$ with probability mass g_{ir} as defined in Section 5.3 of Chapter V. Then for any $i = 0, 1, ..., n$, and for any $a > 0$,*

$$P\Big(|\tilde{e}_i/\sqrt{n}| > a\Big) = O\Big(e^{-a\epsilon_n\sqrt{n}}\Big) \quad as \quad n \to \infty.$$

Proof. Note that in fact $\tilde{e}_i$ has a discretized $\text{Lap}(0, 1/\epsilon_n)$ distribution truncated at 0 and n. We start with the non-truncated but discretized Laplace distribution. Let $Y \sim$ discretized $\text{Lap}(0, 1/\epsilon_n)$ with

$$P(Y = k) = e^{-\epsilon_n |k|}/C, \qquad k = ..., -2, -1, 0, 1, 2, ...$$

where $C = \sum_{m \in \mathbb{Z}} e^{-\epsilon_n |m|} = \Big(1 + 2e^{-\epsilon_n}/(1 - e^{-\epsilon_n})\Big)$. Now consider the truncated version, $\tilde{e}_i$, where the truncated tail probabilities are simply added to the corresponding boundaries. For any $i = 0, ..., n$,

$$P\Big(\Big|\frac{\tilde{e}_i}{\sqrt{n}}\Big| > a\Big) = P(|\tilde{e}_i| > a\sqrt{n}) \leq P(|Y| > a\sqrt{n}) = O\Big(e^{-a\epsilon_n\sqrt{n}}\Big) \quad \text{as} \quad n \to \infty.$$

□

Lemma D.2 below characterizes the L_1 distance between the remapped index and the input index under the optimal remap x introduced in Section 5.3 of Chapter V.

Lemma D.2. *Let* $P(i \mid r') = \alpha^{|i-r'|}/\big(\sum_{i'=0}^{n} \alpha^{|i'-r'|}\big)$ *for* i *taking values in* $\{0, 1, ..., n\}$ *with* $0 < \alpha < 1$ *and for some fixed* $r' \in \{0, 1, ..., n\}$. *Let* r^* *be defined as*

$$r^* = \min_k \Big\{k = 0, ..., n : \sum_{i=0}^{k} \alpha^{|i-r'|}/\big(\sum_{i'=0}^{n} \alpha^{|i'-r'|}\big) \geq 1/2\Big\}. \tag{D.1}$$

Then,

$$|r' - r^*| = O\Big(\max\big\{-\log(1 + 1/2\alpha^{-1})/\log(\alpha), \quad \log(1/2)/\log(\alpha)\big\}\Big) \qquad \textit{as} \quad n \to \infty.$$

Proof. Note that

$$\begin{aligned}\sum_{i'=0}^{n}\alpha^{|i'-r'|} &= \sum_{i'=0}^{n-r'}\alpha^{i'}+\sum_{i'=0}^{r'}\alpha^{i'}-1\\ &= \frac{1-\alpha^{n-r'+1}}{1-\alpha}+\frac{1-\alpha^{r'+1}}{1-\alpha}-1\\ &= \frac{1+\alpha-\alpha^{n-r'+1}-\alpha^{r'+1}}{1-\alpha}.\end{aligned}$$

Consider $\sum_{i=0}^{r^*}\alpha^{|i-r'|}$. We can have either $r^*\le r'$ or $r^*>r'$. When $r^*\le r'$, we have

$$\begin{aligned}\sum_{i=0}^{r^*}\alpha^{|i-r'|} &= \sum_{i=0}^{r'}\alpha^{i}-\sum_{i=0}^{r'-r^*-1}\alpha^{i}\\ &= \frac{1-\alpha^{r'+1}}{1-\alpha}-\frac{1-\alpha^{r'-r^*}}{1-\alpha}\\ &= \frac{\alpha^{r'-r^*}-\alpha^{r'+1}}{1-\alpha}\end{aligned}$$

Then by (D.1), we have for some $0\le c\le 1/2$, such that

$$\begin{aligned}c &= \sum_{i=0}^{r^*}\alpha^{|i-r'|}\Big/\Big(\sum_{i'=0}^{n}\alpha^{|i'-r'|}\Big)-\frac{1}{2},\\ \sum_{i=0}^{r^*}\alpha^{|i-r'|} &= (0.5+c)\Big(\sum_{i'=0}^{n}\alpha^{|i'-r'|}\Big),\\ \alpha^{r'-r^*}-\alpha^{r'+1} &= (0.5+c)(1+\alpha-\alpha^{n-r'+1}-\alpha^{r'+1}),\\ |r'-r^*| &= \Big|\log\Big\{(0.5+c)(1+\alpha-\alpha^{n-r'+1}-\alpha^{r'+1})+\alpha^{r'+1}\Big\}/\log\alpha\Big|. \quad \text{(D.2)}\end{aligned}$$

Now we discuss the order of (D.2). As $n \to \infty$,

$$(\text{D.2}) = \begin{cases} O\Big(\log\{0.5(1+\alpha)\}/\log(\alpha) \Big) & \text{if} \quad r' = \omega(1), r' = o(n) \\ O\Big(\log\{0.5(1+\alpha-\alpha^{r'+1}) + \alpha^{r'+1}\}/\log(\alpha) \Big) & \text{if} \quad r' = O(1) \\ O\Big(\log\{0.5(1+\alpha-\alpha^{n-r'+1})\}/\log(\alpha) \Big) & \text{if} \quad r' = \Theta(n). \end{cases}$$
$$= O\Big(\log(0.5)/\log(\alpha) \Big).$$

Note that $c \to 0$ as $n \to \infty$. Now we consider $r^* > r'$,

$$\begin{aligned} \sum_{i=0}^{r^*} \alpha^{|i-r'|} &= \sum_{i=0}^{r'} \alpha^i + \sum_{i=0}^{r^*-r'} \alpha^i - 1. \\ &= \frac{1-\alpha^{r'+1}}{1-\alpha} + \frac{1-\alpha^{r^*-r'+1}}{1-\alpha} - 1 \\ &= \frac{1+\alpha-\alpha^{r'+1}-\alpha^{r^*-r'+1}}{1-\alpha}. \end{aligned}$$

Also from (D.1), we have

$$\begin{aligned} \sum_{i=0}^{r^*} \alpha^{|i-r'|} &= (0.5+c)\big(\sum_{i'=0}^{n} \alpha^{|i'-r'|} \big) \\ 1+\alpha-\alpha^{r'+1}-\alpha^{r^*-r'+1} &= (0.5+c)(1+\alpha-\alpha^{n-r'+1}-\alpha^{r'+1}) \\ \alpha^{r^*-r'+1} &= -(0.5+c)(1+\alpha-\alpha^{n-r'+1}-\alpha^{r'+1}) + 1 + \alpha - \alpha^{r'+1} \\ \alpha^{r^*-r'} &= -(0.5+c)(1+\alpha^{-1}-\alpha^{n-r'}-\alpha^{r'}) + 1 + \alpha^{-1} - \alpha^{r'} \\ |r'-r^*| &= \Big| \frac{\log\Big\{ -(0.5+c)(1+\alpha^{-1}-\alpha^{n-r'}-\alpha^{r'}) + 1 + \alpha^{-1} - \alpha^{r'} \Big\}}{\log \alpha} \Big|. \end{aligned} \tag{D.3}$$

Now we discuss the order of (D.3). As $n \to \infty$,

$$\text{(D.3)} = \begin{cases} O\Big(|\log(\frac{1}{2} + \frac{1}{2}\alpha^{-1}))/\log(\alpha)|\Big) & \text{if} \quad r' = \omega(1), r' = o(n) \\ O\Big(|\log(\frac{1}{2} + \frac{1}{2}\alpha^{-1} - \frac{1}{2}\alpha^{r'})/\log(\alpha)|\Big) & \text{if} \quad r' = O(1) \\ O\Big(|\log(\frac{1}{2} + \frac{1}{2}\alpha^{-1} - \frac{1}{2}\alpha^{n-r'})/\log(\alpha)|\Big) & \text{if} \quad r' = \Theta(n). \end{cases}$$
$$= O\Big(-\log(1 + 0.5\alpha^{-1})/\log(\alpha)\Big).$$

Hence, $|r' - r^*| = O\left(\max\left\{-\log(1 + 1/2\alpha^{-1})/\log(\alpha), \quad \log(1/2)/\log(\alpha)\right\}\right)$ as $n \to \infty$. □

Lemma D.3. *Assume Y_k^* and Y_{Tk}^* are from the ϵ_n-DP Laplace and the truncated ϵ_n-DP Laplace (at zero) mechanisms respectively, with the same underlying $Y_k \sim$ Bin(n, p_k) for $p_k \in (0, 1)$. Then for any $0 < \triangle_k < p_k$,*

$$P(Y_k^* \neq Y_{Tk}^*) = O\Big(e^{n(\triangle_k - p_k)\epsilon_n} + e^{-2n\triangle_k}\Big), \quad \text{as } n \to \infty.$$

Proof. Note that

$$Y_k^* = Y_k + err_k,$$
$$Y_{Tk}^* = Y_k + err_k + b_k,$$

where $err_k \sim \text{Lap}(0, 1/\epsilon_n)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -Y_k, \quad \text{(D.4)} \\ -(Y_k + err_k), & \text{otherwise.} \quad \text{(D.5)} \end{cases}$$

Note that for any $0 < \triangle_k < p_k$, we have

$$
\begin{aligned}
P(Y_k^* \neq Y_{Tk}^*) =& P(err_k < -Y_k) \\
=& P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} > \triangle_k) \cdot P(\frac{|Y_k - np_k|}{n} > \triangle_k) \\
&+ P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} \leq \triangle_k) \cdot P(\frac{|Y_k - np_k|}{n} \leq \triangle_k) \\
\leq& 2P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n}| > \triangle_k) \cdot e^{-2n\triangle_k} \\
&+ P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} \leq \triangle_k) \cdot 1 \qquad \text{(D.6)} \\
=& P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} \leq \triangle_k) + O\Big(e^{-2n\triangle_k}\Big) \\
=& P(err_k < -Y_k \mid np_k - n\triangle_k \leq Y_k \leq np_k + n\triangle_k) + O\Big(e^{-2n\triangle_k}\Big).
\end{aligned}
$$

where (D.6) follows from Hoeffding's inequality. Further note that

$$
\begin{aligned}
P(err_k < -Y_k \mid np_k - n\triangle_k \leq Y_k \leq np_k + n\triangle_k) &\leq P\big(err_k < n(\triangle_k - p_k)\big) \\
&= \frac{1}{2} e^{n(\triangle_k - p_k)\epsilon_n} \\
&= O\Big(e^{n(\triangle_k - p_k)\epsilon_n}\Big).
\end{aligned}
$$

Hence, we have

$$
P(Y_k^* \neq Y_{Tk}^*) = O\Big(e^{n(\triangle_k - p_k)\epsilon_n} + e^{-2n\triangle_k}\Big).
$$

□

Lemma D.4. *Assume Y_k^* and Y_{Tk}^* are from the (ϵ_n, δ)-DP Gaussian and the trun-*

cated (ϵ_n, δ)*-DP Gaussian (at zero) mechanisms respectively, with the same underlying* $Y_k \sim Bin(n, p_k)$ *for* $p_k \in (0,1)$*. Then for any* $0 < \triangle_k < p_k$*,*

$$P(Y_k^* \neq Y_{Tk}^*) = O\Big(\exp\Big\{ -2n\triangle_k \Big\} + \frac{1}{n\epsilon_n} \exp\Big\{ -\frac{n^2\epsilon_n^2(p_k - \triangle_k)^2}{4\ln\{1.25/\delta\}} \Big\} \Big).$$

Proof. Note that

$$Y_k^* = Y_k + err_k,$$
$$Y_{Tk}^* = Y_k + err_k + b_k,$$

where $err_k \sim \mathrm{N}\big(0, 2\ln\{1.25/\delta\}/\epsilon_n^2\big)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -Y_k, \quad \text{(D.7)} \\ -(Y_k + err_k), & \text{otherwise.} \quad \text{(D.8)} \end{cases}$$

Note that for any $Y \sim \mathrm{N}(0, \sigma^2)$, we have for any $t \in (0, \infty)$,

$$P(Y > t) \leq \frac{\sigma}{t\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}}. \quad \text{(D.9)}$$

To see this, note

$$\begin{aligned}
P(Y > t) &= \int_t^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy \\
&\leq \frac{1}{t\sqrt{2\pi\sigma^2}} \int_t^\infty y e^{-\frac{y^2}{2\sigma^2}} dy \qquad \text{since } \frac{y}{t} \geq 1 \text{ for } y \in [t, \infty) \\
&= \frac{1}{t\sqrt{2\pi\sigma^2}} \Big[-\sigma^2 e^{-\frac{y^2}{2\sigma^2}} \Big]_t^\infty \\
&= \frac{\sigma}{t\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}}.
\end{aligned}$$

Further note that for any $0 < \triangle_k < p_k$, we have

$$\begin{aligned}
P(Y_k^* \neq Y_{Tk}^*) =& P(err_k < -Y_k) \\
=& P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} > \triangle_k) \cdot P(\frac{|Y_k - np_k|}{n} > \triangle_k) \\
&+ P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} \leq \triangle_k) \cdot P(\frac{|Y_k - np_k|}{n} \leq \triangle_k) \\
\leq& 2P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n}| > \triangle_k) \cdot e^{-2n\triangle_k} \\
&+ P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} \leq \triangle_k) \cdot 1 \\
=& P(err_k < -Y_k \mid \frac{|Y_k - np_k|}{n} \leq \triangle_k) + O(e^{-2n\triangle_k}) \\
=& P(err_k < -Y_k \mid np_k - n\triangle_k \leq Y_k \leq np_k + n\triangle_k) + O\Big(e^{-2n\triangle_k}\Big).
\end{aligned}$$

where the ineqaulity in the third step can be obtained from Hoeffding's inequality.

By applying (D.9), we have

$$
\begin{aligned}
&P(err_k < -Y_k \mid np_k - n\triangle_k \le Y_k \le np_k + n\triangle_k) \\
&\le P\big(err_k < n(\triangle_k - p_k)\big) \\
&\le \frac{\sqrt{\ln\{1.25/\delta\}}}{\sqrt{\pi} n\epsilon_n (p_k - \triangle_k)} \exp\Big\{ - \frac{n^2\epsilon_n^2 (p_k - \triangle_k)^2}{4\ln\{1.25/\delta\}} \Big\} \\
&= O\Big(\frac{1}{n\epsilon_n} \exp\Big\{ - \frac{n^2\epsilon_n^2 (p_k - \triangle_k)^2}{4\ln\{1.25/\delta\}} \Big\} \Big).
\end{aligned}
$$

Hence, we have

$$
P(Y_k^* \neq Y_{Tk}^*) = O\Big(\exp\Big\{ - 2n\triangle_k \Big\} + \frac{1}{n\epsilon_n} \exp\Big\{ - \frac{n^2\epsilon_n^2 (p_k - \triangle_k)^2}{4\ln\{1.25/\delta\}} \Big\} \Big).
$$

□

Next we give a proposition specifying the rate of convergence of root-n-scaled random errors injected by the optimal mechanism with L_1 loss.

Proposition D.5. *For $0 < \epsilon_n < \infty$ and some fixed $i \in \{0, 1, ..., n\}$, let $err_i = r - i$ for r taking values in $\{0, ..., n\}$ with probability mass p_{ir}^* as defined in Section 5.3 of Chapter V. Then for any $a > 0$,*

$$
P\Big(\Big| err_i / \sqrt{n} \Big| > a \Big) = O\Big(\max\{1 + e^{\epsilon_n}/2, 2\} e^{-\epsilon_n a \sqrt{n}} \Big) \quad as \quad n \to \infty.
$$

Proof. Consider

$$P\Big(\Big|\frac{err_i}{\sqrt{n}}\Big| > a\Big) = P(|err_i| > a\sqrt{n}) = \sum_{r:|r-i|>a\sqrt{n}, r\in\{0,...,n\}} p^*_{ir}$$
$$= \sum_{r:|r-i|>a\sqrt{n}, r\in\{0,...,n\}} \Big(\sum_{r'=0}^{n} g_{ir'} x_{r'r}\Big)$$
$$= \sum_{r'=0}^{n} g_{ir'} \Big(\sum_{r:|r-i|>a\sqrt{n}, r\in\{0,...,n\}} x_{r'r}\Big), \tag{D.10}$$

where $x_{r'r}$ is the (r', r) entry of the optimal remap as defined in Section 5.3 of Chapter V. Note that from the computation of the optimal remap x, for any given $r \in \{0, ..., n\}$, $x_{r,r^*} = 1$ if r^* equals the conditional median of $i \in \{0, ..., n\}$ with probability mass $P(i \mid r')$, and $x_{r,r'} = 0$ for all $r' \neq r^*$. Write $\alpha = e^{-\epsilon_n}$. Note that for any given $r' \in \{0, 1, ..., n\}$,

$$P(i \mid r') = \frac{\alpha^{|i-r'|}}{\sum_{i'=0}^{n} \alpha^{|i'-r'|}}.$$

By Lemma D.2, we know that
$|r' - r^*| = O\Big(\max\big\{ -\log(1 + 1/2\alpha^{-1})/\log(\alpha), \quad \log(1/2)/\log(\alpha)\big\}\Big)$.

$$|r^* - i|$$
$$= |r^* - r' + r' - i| \leq |r^* - r'| + |r' - i|$$
$$\leq |r' - i| + O\Big(\max\big\{ -\log(1 + 1/2\alpha^{-1})/\log(\alpha), \log(1/2)/\log(\alpha)\big\}\Big) \quad \text{as} \quad n \to \infty. \tag{D.11}$$

Note that there exists n' such that for all $n > n'$, $|r^* - r'| \leq C$, where $C = \max\big\{ -\log(1+1/2\alpha^{-1})/\log(\alpha), \log(1/2)/\log(\alpha)\big\}$. Also note that whenever $|r'-i| \leq a\sqrt{n}-C$, we have $|r^* - i| \leq a\sqrt{n}$ for $n > n'$. It follows that for all $n > n'$,

$$\Big(\sum_{r:|r-i|>a\sqrt{n}, r\in\{0,\ldots,n\}} x_{r'r} \Big) = 0.$$

Hence for any $n > n'$,

$$(\text{D.10}) = \sum_{r':|r'-i|>a\sqrt{n}-C} g_{ir'} \leq 2P(y > a\sqrt{n} - C) = O(e^{-a\epsilon_n\sqrt{n}+C\epsilon_n}),$$

where Y has discretized $\text{Lap}(0, 1/\epsilon_n)$ distribution. Hence, it follows that

$$P\Big(\Big|\frac{err_i}{\sqrt{n}}\Big| > a\Big) = O\Big(e^{\epsilon_n(C-a\sqrt{n})}\Big) \quad \text{as} \quad n \to \infty. \tag{D.12}$$

Since $\alpha = e^{-\epsilon_n}$, further note that

$$\begin{aligned} C &= \max\big\{-\log(1+1/2\alpha^{-1})/\log(\alpha), \quad \log(1/2)/\log(\alpha)\big\} \\ &= \max\big\{\frac{\log(1+1/2e^{\epsilon_n})}{\epsilon_n}, \frac{\log(2)}{\epsilon_n}\big\}. \end{aligned}$$

Substitute into (D.12), we have

$$\begin{aligned} P\Big(\Big|\frac{err_i}{\sqrt{n}}\Big| > a\Big) &= O\Big(e^{\epsilon_n\left(\max\left\{\frac{\log(1+1/2e^{\epsilon_n})}{\epsilon_n}, \frac{\log(2)}{\epsilon_n}\right\} - a\sqrt{n}\right)}\Big) \\ &= O\Big(\max\big\{1+\frac{e^{\epsilon_n}}{2}, 2\big\}e^{-\epsilon_n a\sqrt{n}}\Big). \end{aligned}$$

□

Next, we give proofs for the theorems. We start with proving Theorem V.10 first.

Proof of Theorem V.10. Note that

$$T_k = \begin{cases} \frac{1}{\sqrt{np_k}}\big(Y_k^* - np_k\big) = \sqrt{\frac{n}{p_k}}\Big(\frac{Y_k - np_k}{n}\Big) + \frac{err_k}{\sqrt{np_k}}, & \text{Lap,} \quad \text{(D.13)} \\ \frac{1}{\sqrt{np_k}}\Big(Y_k^* - np_k\Big) = \sqrt{\frac{n}{p_k}}\Big(\frac{Y_k - np_k}{n}\Big) + \frac{err_k}{\sqrt{np_k}} + \frac{b_k}{\sqrt{np_k}}, & \text{TLap,} \quad \text{(D.14)} \end{cases}$$

where $err_k \sim \text{Lap}(0, 1/\epsilon_n)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -Y_k, \quad \text{(D.15)} \\ -(Y_k + err_k), & \text{otherwise.} \quad \text{(D.16)} \end{cases}$$

First, we seek to show that $P(\sum_k (\text{D.13})^2 \neq \sum_k (\text{D.14})^2) = o(1)$ as $n \to \infty$, so that we can ignore the bias term b_k in the test statistic in the remaining proofs. Note that for any $0 < \triangle_k < p_k$, by Lemma D.3, we have

$$P(b_k \neq 0) = O\Big(e^{n(\triangle_k - p_k)\epsilon_n} + e^{-2n\triangle_k}\Big). \quad \text{(D.17)}$$

Now take $\triangle_k = \min\{p_1, ..., p_K\}/2$. Further note that

$$
\begin{aligned}
&P(\sum_k (\text{D.13})^2 \neq \sum_k (\text{D.14})^2) \\
&= P(\cup_k \{b_k \neq 0\}) \\
&\leq \sum_{k=1}^{K} P(b_k \neq 0) \\
&= O\Big(\exp\{-\frac{1}{2} n\epsilon_n \min\{p_1, ..., p_K\}\} + \exp\{-n \min\{p_1, ..., p_K\}\} \Big).
\end{aligned}
$$

where the last step follows directly from (D.17) and the fact that $K < \infty$. Since the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$, it follows $P(\sum_k (\text{D.13})^2 \neq \sum_k (\text{D.14})^2) = o(1)$ as $n \to \infty$. Therefore, we can ignore the bias term b_k in the following proofs. Note that marginally, $Y_k \sim \text{Binomial}(n, p_k)$. By the Central Limit Theorem (CLT), we know that $(Y_k - np_k)/n$ will converge to a Gaussian variable as $n \to \infty$ since Y_k can be viewed as a sum of n i.i.d. $\text{Ber}(p_k)$ under H_0. As a direct consequence of Lemma D.1, when the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$, the second term in (D.13), $err_k/\sqrt{np_k} \xrightarrow{p} 0$ as $n \to \infty$. Therefore, overall, T_k will converge to a Gaussian distribution as $n \to \infty$. However, instead of treating $err_k/\sqrt{n}$ as 0, we take its first and second moments into account to have better finite sample approximations while maintaining correct asymptotic distribution. Note $err_k \sim \text{Lap}(0, 1/\epsilon_n)$, we have $\mathbb{E}[T_k] = 0$ and

$$
\begin{aligned}
Var(T_k) = \frac{1}{np_k} Var(Y_k^*) &= \frac{1}{np_k} \big\{ Var(Y_k) + Var(err_k) \big\} \\
&= 1 - p_k + 2/(np_k\epsilon_n^2).
\end{aligned}
$$

Note also that correlations amongst Y_k induce correlations amongst T_k. For some $k \neq j$, consider,

$$\begin{aligned} Cov(T_k, T_j) &= Cov\Big(\sqrt{\frac{n}{p_k}}\Big(\frac{Y_k^*}{n} - p_k\Big), \sqrt{\frac{n}{p_j}}\Big(\frac{Y_j^*}{n} - p_j\Big)\Big) \\ &= \frac{1}{n\sqrt{p_k p_j}} Cov\Big(Y_k + err_k, Y_j + err_j\Big) \\ &= \frac{1}{n\sqrt{p_k p_j}} Cov\Big(Y_k, Y_j\Big) \\ &= -\sqrt{p_k p_j}. \end{aligned}$$

Let Σ be the covariance matrix of $\boldsymbol{T} = (T_1, T_2, ..., T_K)$. Consider a matrix $O = [v_1, ..., v_K] \in \mathbb{R}^{K \times K}$ consisting of the orthonormal eigenvectors of Σ as columns. So we must have $\Sigma O = \Lambda O$, where Λ is a diagonal matrix with diagonal elements Λ_k being the eigenvalues of Σ with respect to v_k. We require $||v_k|| = 1$ for all $k = 1, ..., K$. So we must have $OO^T = O^T O = I_K$. Consider transformed vector $\boldsymbol{T'} = O\boldsymbol{T} = (T_1', ..., T_K')$. First note that each T_k' is asymptotically normal since it is a linear combination of normal distributions. Further, we also have $Cov(\boldsymbol{T'}) = Cov(O\boldsymbol{T}) = \Lambda$. Hence T_k' are independent $\mathrm{N}(0, \Lambda_k)$ for $k = 1, ..., K$. So, $T^* = \sum_{k=1}^K T_k^2 = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{T'}^T\boldsymbol{T'} \to \sum_{k=1}^K \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1.

Now for part (b) of the theorem, again we can express

$$T_k = \begin{cases} \dfrac{1}{\sqrt{np_k}}\big(Y_k^* - np_k\big) = \sqrt{\dfrac{n}{p_k}}\big(\dfrac{Y_k - np_k}{n}\big) + \dfrac{err_k}{\sqrt{np_k}}, & \text{GDP}, \quad \text{(D.18)} \\ \dfrac{1}{\sqrt{np_k}}\big(Y_k^* - np_k\big) = \sqrt{\dfrac{n}{p_k}}\big(\dfrac{Y_k - np_k}{n}\big) + \dfrac{err_k}{\sqrt{np_k}} + \dfrac{b_k}{\sqrt{np_k}}, & \text{TGDP}, \quad \text{(D.19)} \end{cases}$$

where $err_k \sim \mathrm{N}\big(0, 2\ln\{1.25/\delta\}/\epsilon_n^2\big)$ and b_k is the bias term which can be expressed as follows,

$$b_k = \begin{cases} 0, & \text{if } err_k \geq -Y_k, \quad \text{(D.20)} \\ -(Y_k + err_k), & \text{otherwise.} \quad \text{(D.21)} \end{cases}$$

Similarly, we seek to show that $P(\sum_k (\text{D.18})^2 \neq \sum_k (\text{D.19})^2) = o(1)$ as $n \to \infty$, so that we can ignore the bias term b_k in the test statistic in the remaining proofs. Note that for any $0 < \triangle_k < p_k$, by Lemma D.4, we have

$$P(b_k \neq 0) = O\Big(\exp\Big\{ -2n\triangle_k \Big\} + \frac{1}{n\epsilon_n} \exp\Big\{ -\frac{n^2\epsilon_n^2 (p_k - \triangle_k)^2}{4\ln\{1.25/\delta\}} \Big\} \Big). \quad \text{(D.22)}$$

Again we can take $\triangle_k = \min\{p_1, ..., p_K\}/2$. Further note that

$$\begin{aligned} & P(\sum_k (\text{D.18})^2 \neq \sum_k (\text{D.19})^2) \\ & = P(\cup_k \{b_k \neq 0\}) \\ & \leq \sum_k P(b_k \neq 0) \\ & = O\Big(\exp\Big\{ -n \min\{p_1, ..., p_K\} \Big\} + \frac{1}{n\epsilon_n} \exp\Big\{ -\frac{n^2\epsilon_n^2 \min\{p_1, ..., p_K\}^2}{16\ln\{1.25/\delta\}} \Big\} \Big), \end{aligned}$$

where the last step follows directly from (D.22) and $K < \infty$. Again, since the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$, it follows $P(\sum_k (\text{D.18})^2 \neq \sum_k (\text{D.19})^2) = o(1)$ as $n \to \infty$. Therefore, we can ignore the bias term b_k in the test statistic. The remaining proof is similar to the Laplace case, except for the difference

in the covariance Σ matrix. Note that in the case of (ϵ_n, δ)-Gaussian mechanism,

$$\begin{aligned} Var(T_k) &= \frac{1}{np_k}\Big(Var(Y_k) + Var(err_k)\Big) \\ &= \frac{1}{np_k}\Big(np_k(1-p_k) + (2\log(1.25/\delta) - 1)/\epsilon_n^2\Big) \\ &= 1 - p_k + \{2\log(1.25/\delta) - 1\}/\{np_k\epsilon_n^2\}, \end{aligned}$$

and $\text{Cov}(T_k, T_j) = -\sqrt{p_k p_j}$. Again, by considering the orthonormal eigenvectors and corresponding eigenvalues Λ_k of Σ, we can arrive at the same conclusion that as $n \to \infty$, $T^* \to \sum_{k=1}^K \Lambda_k Z_k$. Therefore, the results of the theorem follow. □

Proof of Theorem V.8. The proof is similar as the proof of Theorem V.10 except that there is an additional de-bias term in the test statistic and err_k has probability mass $p^*_{k,\cdot}$ in this case.

We focus on T'_k first,

$$T'_k = \frac{1}{\sqrt{np_k}}\big(Y^*_k - np_k - b(y^*_k)\big) = \sqrt{\frac{n}{p_k}}\big(\frac{Y_k - np_k}{n}\big) + \frac{err_k - b(y^*_k)}{\sqrt{np_k}}. \tag{D.23}$$

Note that by CLT, we know that $(Y_k - np_k)/n$ will converge to a Gaussian variable as $n \to \infty$. As a direct consequence of Proposition D.5 and the fact that $b(y^*_k) = O(1)$, the second term in (D.23) converges in probability to 0. Therefore, T_k converges in distribution to a normal random variable as $n \to \infty$. We just need to derive its asymptotic mean and variance to pinpoint its distribution. Note that $\mathbb{E}[T'_k] \to 0$. Before giving an estimate for the variance, we first derive the order of $Var(err_k)$. Let $\alpha_n = e^{-\epsilon_n}$. Note that $0 < \alpha_n < 1$. First consider $\tilde{e}_k = j - k$ for $j = 0, 1, ..., n$

with probability mass $g_{kj} = \frac{1-\alpha_n}{1+\alpha_n}\alpha_n^{|j-k|}$ for $j = 1, ..., (n-1)$, $g_{kj} = \alpha_n^{|j-k|}/(1+\alpha_n)$ for $j = 0, n$, with some fixed $k \in \{0, ..., n\}$. Note that

$$\tilde{\mu}_k := \mathbb{E}[\tilde{e}_k] = \sum_{j=0}^{n} g_{kj}(j-k) = \frac{1-\alpha_n}{1+\alpha_n}\sum_{j=1}^{n-1}\alpha_n^{|j-k|}(j-k) + \frac{\alpha_n^{|n-k|}}{1+\alpha_n}(n-k) - \frac{k\alpha_n^k}{1+\alpha_n}$$
$$= O(1) \quad \text{as } n \to \infty.$$

Consider variance of $\tilde{e}_k$,

$$\begin{aligned} Var(\tilde{e}_k) =& \sum_{j=0}^{n} g_{kj}\big(j-k-\tilde{\mu}_k\big)^2 \\ =& \frac{1-\alpha_n}{1+\alpha_n}\sum_{j=1}^{n-1}\alpha_n^{|j-k|}(j-k)^2 - 2\tilde{\mu}_k^2 + \frac{1-\alpha_n}{1+\alpha_n}\sum_{j=1}^{n-1}\alpha_n^{|j-k|}\tilde{\mu}_k^2 \\ &+ \frac{k^2\alpha_n^k}{1+\alpha_n} + \frac{\alpha_n^{|n-k|}}{1+\alpha_n}(n-k)^2 + \frac{\tilde{\mu}_k^2\alpha_n^k}{1+\alpha_n} + \frac{\tilde{\mu}_k^2\alpha_n^{n-k}}{1+\alpha_n} \\ =& O(1) \quad \text{as } n \to \infty. \end{aligned} \tag{D.24}$$

Now denote $\mu_k := \mathbb{E}[err_k]$. As a direct consequence of Lemma D.2, there exists a constant $C_1 < \infty$ such that $|\tilde{\mu}_k - \mu_k| < C_1$ and $C_2 < \infty$ such that the input index r'

and optimally remapped index r^* satisfying $|r' - r^*| < C_2$. Then,

$$\begin{aligned}
Var(err_k) =& \sum_{j=0}^{n} p^*_{kj}\big(j-k-\mu_k\big)^2 \\
\leq& \frac{1-\alpha_n}{1+\alpha_n}\sum_{j=1}^{n-1}\alpha_n^{C_2+|j-k|}(C_1+C_2+j-k-\tilde{\mu}_k)^2 \\
&+ \frac{\alpha_n^{C_2+k}}{1+\alpha_n}(C_1+C_2+j-k-\tilde{\mu}_k)^2 + \frac{\alpha_n^{C_2+n-k}}{1+\alpha_n}(C_1+C_2+j-k-\tilde{\mu}_k)^2 \\
=& \alpha_n^{C_2}\sum_{j=0}^{n} g_{kj}\Big\{(C_1+C_2)^2 + 2(C_1+C_2)(j-k-\tilde{\mu}_k) + (j-k-\tilde{\mu}_k)^2\Big\} \\
=& \alpha_n^{C_2}\Big\{(C_1+C_2)^2 + 2(C_1+C_2)(\tilde{\mu}_k-\tilde{\mu}_k) + Var(\tilde{e}_k)\Big\} \\
=& O(1) \quad \text{as } n\to\infty,
\end{aligned} \tag{D.25}$$

where the last step follows from (D.24). Now we can evaluate the variance term.

$$\begin{aligned}
Var(T'_k) &= \frac{1}{np_k}Var(Y^*_k) = \frac{1}{np_k}\big(Var(Y_k) + Var(err_k)\big) \\
&= 1 - p_k + Var(err_k)/(np_k) \\
&= 1 - p_k + O\Big(\frac{1}{n}\Big) \quad \text{as } n\to\infty.
\end{aligned} \tag{D.26}$$

The last step follows directly from (D.25). Note also that

$$1 - p_k + v(y^*_k)/(np_k) = 1 - p_k + O\Big(\frac{1}{n}\Big) \quad \text{as } n\to\infty. \tag{D.27}$$

We can see this from the construction of the variance estimate $v(y^*_k)$ in Algorithm 10 of Chapter V. First note that in the step 2 of Algorithm 10, $v_i = Var(err_i) = O(1)$.

$\{f_{iy_k^*} : i = 1, ..., n\}$ is a probability distribution such that $\sum_{i=1}^n f_{iy_k^*} = 1$. Hence, $v(y_k^*) = \sum_{i=1}^n f_{iy_k^*} v_i = O(1)$ as $n \to \infty$. From Equations (D.26) and (D.27), we have

$$\Big|Var(T_k') - \big(1 - p_k + v(y_k^*)/(np_k)\big)\Big| = O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty.$$

Lastly, since Y_k are correlated, T_k' are correlated. For any $k \neq j$, consider,

$$\begin{aligned} Cov(T_k', T_j') &= Cov\Big(\sqrt{\frac{n}{p_k}}\Big(\frac{Y_k^*}{n} - p_k\Big), \sqrt{\frac{n}{p_j}}\Big(\frac{Y_j^*}{n} - p_j\Big)\Big) \\ &= \frac{1}{n\sqrt{p_k p_j}} Cov\Big(Y_k + err_k, Y_j + err_j\Big) \\ &= \frac{1}{n\sqrt{p_k p_j}} Cov\Big(Y_k, Y_j\Big) \\ &= -\sqrt{p_k p_j}. \end{aligned}$$

Let $\Sigma \in \mathbb{R}^{K \times K}$ be a matrix with diagonal entries $\Sigma_{kk} = 1 - p_k + v(y_k^*)/(np_k)$ and off-diagonal entries $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $k \neq j$. Let $\tilde{\Sigma}$ be the covariance matrix of $\boldsymbol{T}' = (T_1', T_2', ..., T_K')$. We have $\|\Sigma - \tilde{\Sigma}\|_\infty = O(n^{-1})$, as $n \to \infty$. Consider a matrix $O = [v_1, ..., v_k] \in \mathbb{R}^{K \times K}$ consisting of the orthonormal eigenvectors of Σ as columns. So we must have $\Sigma O = \Lambda O$, where Λ is a diagonal matrix with diagonal elements Λ_k being the eigenvalues of Σ with respect to v_k. We require $\|v_k\| = 1$ for all $k = 1, ..., K$. Following a similar argument as in the proof of Theorem V.10, we can derive that $T_{opt}^* \to \sum_{k=1}^K \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1. $\square$

Proof of Theorem V.11. We start with proving part (a). Consider T_{mk},

$$\begin{aligned} T_{mk} &= \frac{1}{\sqrt{np_k}}\Big(\sum_{j=1}^{C} Y_{jk}^* - np_k - b_M(\{y_{jk}^*\}_{j=1}^C)\Big) \\ &= \sqrt{\frac{n}{p_k}}\Big(\frac{\sum_{j=1}^C Y_{jk} - np_k}{n}\Big) + \frac{\sum_{j=1}^C \big(err_{jk} - b(y_{jk}^*)\big)}{\sqrt{np_k}}. \end{aligned} \tag{D.28}$$

Assume H_0 is true, by CLT, we know that $\Big(\sum_{j=1}^C Y_{jk} - np_k\Big)/n$ will converge to a Gaussian distribution with mean 0 as $n \to \infty$ since $\sum_{j=1}^C Y_{jk}$ can be viewed as a sum of $n = n_1 + n_2 + ... + n_C$ i.i.d. Ber(p_k) random variables. As a direct consequence of Proposition D.5 and the facts that $b(y_{jk}^*), C < \infty$, when the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$, the second term of (D.28) will converge in probability to 0 as $n \to \infty$. Therefore, overall, T_{mk} will converge to a Gaussian random variable. Denote $Y_k = \sum_{j=1}^C Y_{jk}$ and $Y_k^* = \sum_{j=1}^C Y_{jk}^*$. Also

$$\begin{aligned} Var(T_{mk}) &= \frac{1}{np_k} Var(Y_k^*) = \frac{1}{np_k}\Big(Var(Y_k) + Var(\sum_{j=1}^{C} err_{jk})\Big) \\ &= \frac{1}{np_k}\Big(Var(Y_k) + \sum_{j=1}^{C} Var(err_{jk})\Big) \\ &= 1 - p_k + \frac{\sum_{j=1}^C Var(err_{jk})}{np_k} \\ &= 1 - p_k + O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty. \end{aligned} \tag{D.29}$$

The last step follows from (D.25) where $Var(err_{jk}) = O(1)$ and the fact that $C < \infty$.

Note also that

$$1 - p_k + \frac{\sum_{j=1}^{C} v(y_{jk}^*)}{np_k} = 1 - p_k + O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty. \tag{D.30}$$

Equation (D.30) follows from the facts that $v(y_{jk}^*) = O(1)$ as $n \to \infty$ (for details see proof of Theorem V.8) and $C < \infty$. From (D.29) and (D.30), we must have

$$\Big| Var(T_{mk}) - \Big(1 - p_k + \frac{\sum_{j=1}^{C} v(y_{jk}^*)}{np_k}\Big)\Big| = O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty.$$

Lastly, since Y_k are correlated, T_{mk} are correlated. For some $k \neq j$, consider,

$$\begin{aligned}
&Cov(T_{mk}, T_{mj}) \\
&= Cov\Big(\sqrt{\frac{n}{p_k}}\Big(\frac{Y_k + \sum_{i=1}^{C} err_{ik}}{n} - p_k\Big), \sqrt{\frac{n}{p_j}}\Big(\frac{Y_j + \sum_{i=1}^{C} err_{ij}}{n} - p_j\Big)\Big) \\
&= \frac{1}{n\sqrt{p_k p_j}} Cov\Big(Y_k + \sum_{i=1}^{C} err_{ik}, Y_j + \sum_{i=1}^{C} err_{ij}\Big) \\
&= \frac{1}{n\sqrt{p_k p_j}} Cov\Big(Y_k, Y_j\Big) \\
&= -\sqrt{p_k p_j}.
\end{aligned}$$

Similarly, let $\Sigma \in \mathbb{R}^{K \times K}$ be a matrix with diagonal $\Sigma_{kk} = 1 - p_k + (\sum_{j=1}^{C} v(y_{jk}^*))/(np_k)$ and off-diagonal $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $k \neq j$. Let $\tilde{\Sigma}$ be the covariance matrix of $\boldsymbol{T}_M = \{T_{m1}, ..., T_{mK}\}$. We have $\|\Sigma - \tilde{\Sigma}\|_\infty = O(n^{-1})$, as $n \to \infty$. By a similar argument as in the proof of Theorem V.8 we can derive that $T_M^* \to \sum_{k=1}^{K} \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1 and Λ_k are the eigenvalues

of Σ corresponding to a set of orthonormal eigenvectors of Σ. Hence, we have the result of part (a) follows.

For part (b), note that

$$T_{mk} = \begin{cases} \sqrt{\frac{n}{p_k}}\Big(\frac{\sum_{j=1}^{C} Y_{jk} - np_k}{n}\Big) + \frac{\sum_{j=1}^{C}\big(err_{jk}\big)}{\sqrt{np_k}}, & \text{Lap,} \quad \text{(D.31)} \\ \sqrt{\frac{n}{p_k}}\Big(\frac{\sum_{j=1}^{C} Y_{jk} - np_k}{n}\Big) + \frac{\sum_{j=1}^{C}\big(err_{jk} + b_{jk}\big)}{\sqrt{np_k}}, & \text{TLap,} \quad \text{(D.32)} \end{cases}$$

where $err_{jk} \sim \text{Lap}(0, 1/\epsilon_n)$ and b_{jk} is the bias term which can be expressed as follows,

$$b_{jk} = \begin{cases} 0, & \text{if } err_{jk} \geq -Y_{jk}, \quad \text{(D.33)} \\ -(Y_{jk} + err_{jk}), & \text{otherwise.} \quad \text{(D.34)} \end{cases}$$

It can be shown that $P(\sum_k (\text{D.31})^2 \neq \sum_k (\text{D.32})^2) = o(1)$ as $n \to \infty$. To see this, note for any $0 < \triangle_{jk} < p_k$, by Lemma D.3, it follows

$$P(b_{jk} \neq 0) = O\Big(e^{n(\triangle_{jk} - p_k)\epsilon_n} + e^{-2n\triangle_{jk}}\Big). \quad \text{(D.35)}$$

Take $\triangle_{jk} = \min\{p_1, ..., p_K\}/2$ for all $j = 1, ..., C$. Further note that as $n \to \infty$,

$$\begin{aligned} &P(\sum_k (\text{D.31})^2 \neq \sum_k (\text{D.32})^2) \\ &= P(\cup_k \cup_j \{b_{jk} \neq 0\}) \\ &\leq \sum_{k=1}^{K}\sum_{j=1}^{C} P(b_{jk} \neq 0) \\ &= O\Big(\exp\Big\{-\frac{1}{2}n\epsilon_n \min\{p_1, ..., p_K\}\Big\} + \exp\Big\{-n\min\{p_1, ..., p_K\}\Big\}\Big). \end{aligned}$$

where the last step follows directly from (D.35) and the fact that $K, C < \infty$. Since the privacy regime ϵ_n satisfies $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$, it follows $P(\sum_k (\text{D.31})^2 \neq \sum_k (\text{D.32})^2) = o(1)$ as $n \to \infty$. Therefore, we can ignore all the bias terms b_{jk} and work only with (D.31) in the following proofs. The remaining proofs for part (b) are similar as in part (a), except for the difference in Σ. In the case of ϵ_n-Laplace mechanism in part (b), $\text{Cov}(T_{mk}, T_{mj})$ remains the same for $j \neq k$ but

$$\begin{aligned} Var(T_{mk}) &= \frac{1}{np_k} Var(Y_k^*) \\ &= \frac{1}{np_k}\Big(Var(Y_k) + Var(\sum_{j=1}^{C} err_{jk})\Big) \\ &= \frac{1}{np_k}\Big(Var(Y_k) + \sum_{j=1}^{C} Var(err_{jk})\Big) \\ &= \frac{1}{np_k}\big(np_k(1-p_k) + 2C/\epsilon_n^2\big) \\ &= 1 - p_k + 2C/(\epsilon_n^2 np_k). \end{aligned}$$

Now for part (c), again we can express

$$T_{mk} = \begin{cases} \sqrt{\dfrac{n}{p_k}}\Big(\dfrac{\sum_{j=1}^{C} Y_{jk} - np_k}{n}\Big) + \dfrac{\sum_{j=1}^{C}\big(err_{jk}\big)}{\sqrt{np_k}}, & \text{GDP}, \qquad \text{(D.36)} \\ \sqrt{\dfrac{n}{p_k}}\Big(\dfrac{\sum_{j=1}^{C} Y_{jk} - np_k}{n}\Big) + \dfrac{\sum_{j=1}^{C}\big(err_{jk} + b_{jk}\big)}{\sqrt{np_k}}, & \text{TGDP}, \qquad \text{(D.37)} \end{cases}$$

where $err_{jk} \sim \text{N}\big(0, 2\ln\{1.25/\delta\}/\epsilon_n^2\big)$ and b_{jk} is the bias term remains the same as

before,

$$b_{jk} = \begin{cases} 0, & \text{if } err_{jk} \geq -Y_{jk}, \quad \text{(D.38)} \\ -(Y_{jk} + err_{jk}), & \text{otherwise.} \quad \text{(D.39)} \end{cases}$$

Similarly, it can be shown that $P(\sum_k (\text{D.36})^2 \neq \sum_k (\text{D.37})^2) = o(1)$ as $n \to \infty$. To see this, note for any $0 < \triangle_{jk} < p_k$, by Lemma D.4, we have

$$P(b_{jk} \neq 0) = O\Big(\exp\Big\{ -2n\triangle_{jk} \Big\} + \frac{1}{n\epsilon_n} \exp\Big\{ -\frac{n^2\epsilon_n^2 (p_k - \triangle_{jk})^2}{4\ln\{1.25/\delta\}} \Big\} \Big). \quad \text{(D.40)}$$

Again, take $\triangle_{jk} = \min\{p_1, ..., p_K\}/2$ for all $j = 1, ..., C$. Further note that as $n \to \infty$,

$$\begin{aligned} & P(\sum_k (\text{D.36})^2 \neq \sum_k (\text{D.37})^2) \\ & = P(\cup_k \cup_j \{b_{jk} \neq 0\}) \\ & \leq \sum_{k=1}^{K} \sum_{j=1}^{C} P(b_{jk} \neq 0) \\ & = O\Big(\exp\Big\{ -n \min\{p_1, ..., p_K\} \Big\} + \frac{1}{n\epsilon_n} \exp\Big\{ -\frac{n^2\epsilon_n^2 \min\{p_1, ..., p_K\}^2}{16\ln\{1.25/\delta\}} \Big\} \Big). \end{aligned}$$

where the last step follows directly from (D.40) and the fact that $K, C < \infty$. Since the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$, it follows $P(\sum_k (\text{D.36})^2 \neq \sum_k (\text{D.37})^2) = o(1)$ as $n \to \infty$. Therefore, we can ignore all the bias terms b_{jk} and work only with (D.36). The remaining proofs for part (c) are similar to that in part (a), except for the difference in Σ. In the case of (ϵ_n, δ)-Gaussian Mechanism in part

(c), $\text{Cov}(T_{mk}, T_{mj})$ remains the same for $j \neq k$ but

$$\begin{aligned} Var(T_{mk}) &= \frac{1}{np_k} Var(Y_k^*) \\ &= \frac{1}{np_k}\Big(Var(Y_k) + Var(\sum_{j=1}^{C} err_{jk})\Big) \\ &= \frac{1}{np_k}\Big(Var(Y_k) + \sum_{j=1}^{C} Var(err_{jk})\Big) \\ &= \frac{1}{np_k}\big(np_k(1-p_k) + (2C\log(1.25/\delta) - 1)/\epsilon_n^2\big) \\ &= 1 - p_k + C(2\log(1.25/\delta) - 1)/(np_k\epsilon_n^2). \end{aligned}$$

We have both parts (b) and (c) of the theorem follow. □

Proof of Theorem V.12. We start with part (a). Consider T_{m1},

$$\begin{aligned} T_{m1} &= \frac{1}{\sqrt{np_1}}\Big(\sum_{k=1}^{M} Y_k^* - np_1 - b_M(\{y_j^*\}_{j=1}^M)\Big) \\ &= \sqrt{\frac{n}{p_1}}\Big(\frac{\sum_{k=1}^M Y_k - np_1}{n}\Big) + \frac{\sum_{k=1}^M \big(err_k - b(y_k^*)\big)}{\sqrt{np_1}}. \end{aligned} \tag{D.41}$$

Assume H_0 is true, by CLT, we know that $\Big(\sum_{k=1}^M Y_k - np_1\Big)/n$ is Gaussian asymptotically with mean 0 as $n \to \infty$. Following from Proposition D.5 and the facts that $M, b(y_k^*) < \infty$, when the privacy regime ϵ_n satisfying $n^{-1/2}\epsilon_n^{-1} \to 0$ as $n \to \infty$, the second term of (D.41) will converge in probability to 0 as $n \to \infty$. Therefore, overall, T_{m1} will converge to a Gaussian distribution. Following a similar argument as the proof of Theorem V.10, we know that T_{mk} for $k = 2, ..., K$ will also converge to Gaussian random variables with mean 0 asymptotically. Denote

$Y_{m1} = \sum_{k=1}^{M} Y_k$ and $Y^*_{m1} = \sum_{k=1}^{M} Y^*_k$. Note

$$\begin{aligned} Var(T_{m1}) &= \frac{1}{np_1} Var(Y^*_{m1}) = \frac{1}{np_1}\Big(Var(Y_{m1}) + Var(\sum_{k=1}^{M} err_k)\Big) \\ &= \frac{1}{np_1}\Big(Var(Y_{m1}) + \sum_{k=1}^{M} Var(err_k)\Big) \\ &= 1 - p_1 + O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty. \end{aligned} \tag{D.42}$$

For $k = 2, ..., K$,

$$\begin{aligned} Var(T_{mk}) &= \frac{1}{np_k} Var(Y^*_{mk}) = \frac{1}{np_k}\Big(Var(Y_{mk}) + Var(err_k)\Big) \\ &= 1 - p_k + O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty. \end{aligned} \tag{D.43}$$

Equations (D.42) and (D.43) follow directly from (D.25) and the fact that $M < K < \infty$. Note also that

$$1 - p_1 + \frac{\sum_{j=1}^{M} v(y^*_j)}{np_1} = 1 - p_1 + O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty. \tag{D.44}$$

$$1 - p_k + \frac{v(y^*_k)}{np_k} = 1 - p_k + O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty. \tag{D.45}$$

Equations (D.44) and (D.45) follow from $v(y^*_j) = O(1)$ as $n \to \infty$ (for more details see proofs of Theorem V.8 and the fact that $M < K < \infty$. From (D.42) and (D.44), we have

$$\Big|Var(T_{m1}) - \Big(1 - p_1 + \frac{\sum_{j=1}^{M} v(y^*_j)}{np_1}\Big)\Big| = O\Big(\frac{1}{n}\Big) \quad \text{as } n \to \infty.$$

Similarly, from Equations (D.43) and (D.45), we can derive for $k = 2, ..., K$,

$$\left|Var(T_{mk}) - \left(1 - p_k + \frac{v(y_k^*)}{np_k}\right)\right| = O\left(\frac{1}{n}\right) \quad \text{as } n \to \infty.$$

Lastly, we derive covariance amongst T_{mk}, $Cov(T_{mk}, T_{mj}) = -\sqrt{p_k p_j}$, for any $k \neq j$. Similarly, let $\Sigma \in \mathbb{R}^{(K-M+1)\times(K-M+1)}$ be a matrix with diagonal entries $\Sigma_{11} = 1 - p_1 + (\sum_{j=1}^{M} v(y_j^*))/(np_1)$ and $\Sigma_{kk} = 1 - p_k + v(y_k^*)/(np_k)$ for $k = 2, ..., K - M + 1$, and off-diagonal entries $\Sigma_{kj} = -\sqrt{p_k p_j}$ for $k \neq j$. Let $\tilde{\Sigma}$ be the covariance matrix of $\boldsymbol{T}_M = (T_{m1}, ..., T_{m(K-M+1)})$. We have $\|\Sigma - \tilde{\Sigma}\|_\infty = O(n^{-1})$, as $n \to \infty$. Then with a similar argument as in the proof of Theorem V.10, $T_M^* \to \sum_{k=1}^{K} \Lambda_k Z_k$, where Z_k are i.i.d. Chi-square distribution with degree of freedom of 1 and Λ_k are eigenvalues of Σ corresponding to a set of orthonormal eigenvectors. Hence, the result of part (a) follows.

For part (b) and (c), first note that using similar arguments as in the proof for Theorem V.10(a) and V.10(b), it can be shown that the probability of test statistics obtained from the Laplace/Gaussian mechanisms not equal to their counterparts obtained from the truncated Laplace/Gaussian mechanisms tends to 0 exponentially fast as n goes to infinity. Therefore, we can in fact ignore the truncation effect. The proofs for the remaining are similar to the proof for part (a), except for the difference in Σ. In the case of ϵ_n-Laplace mechanism in part (b), $Cov(T_{mk}, T_{mj}) = -\sqrt{p_k p_j}$ for

$k \neq j$ remains the same but

$$\begin{aligned} Var(T_{m1}) &= \frac{1}{np_1} Var(Y^*_{m1}) = \frac{1}{np_1}\Big(Var(Y_{m1}) + Var(\sum_{k=1}^{M} err_k)\Big) \\ &= \frac{1}{np_1}\Big(Var(Y_{m1}) + \sum_{k=1}^{M} Var(err_k)\Big) \\ &= 1 - p_1 + 2M/(\epsilon_n^2 np_1). \end{aligned}$$

For $k = 2, ..., K$,

$$Var(T_{mk}) = \frac{1}{np_k} Var(Y^*_{mk}) = \frac{1}{np_k}\Big(Var(Y_{mk}) + Var(err_k)\Big) = (1 - p_k) + 2/(np_k\epsilon_n^2).$$

In the case of (ϵ_n, δ)-Gaussian Mechanism in part (c), $Cov(T_{mk}, T_{mj}) = -\sqrt{p_k p_j}$ for $k \neq j$ remains the same but

$$\begin{aligned} Var(T_{m1}) &= \frac{1}{np_1} Var(Y^*_{m1}) = \frac{1}{np_1}\Big(Var(Y_{m1}) + Var(\sum_{k=1}^{M} err_k)\Big) \\ &= \frac{1}{np_1}\Big(Var(Y_{m1}) + \sum_{k=1}^{M} Var(err_k)\Big) \\ &= (1 - p_1) + M(2\log(1.25/\delta) - 1)/(np_1\epsilon_n^2). \end{aligned}$$

For $k = 2, ..., K$,

$$\begin{aligned} Var(T_{mk}) &= \frac{1}{np_k} Var(Y^*_{mk}) = \frac{1}{np_k}\Big(Var(Y_{mk}) + Var(err_k)\Big) \\ &= (1 - p_k) + (2\log(1.25/\delta) - 1)/np_k\epsilon_n^2. \end{aligned}$$

The results of both part (b) and (c) follow from a similar argument as in the proof of Theorem V.10. □

BIBLIOGRAPHY

BIBLIOGRAPHY

Abowd, J. M., and L. Vilhuber (2008), How protective are synthetic data?, in *International Conference on Privacy in Statistical Databases*, pp. 239–246, Springer, New York, U.S.

Andersen, E. B. (1970), Asymptotic properties of conditional maximum-likelihood estimators, *Journal of the Royal Statistical Society: Series B (Methodological)*, *32*(2), 283–301.

Araújo, L., P. Costa, and N. Crato (2021), Assessment background: What PISA measures and how, *Improving a Country's Education*, pp. 249–263.

Avella-Medina, M. (2021), Privacy-preserving parametric inference: A case for robust statistics, *Journal of the American Statistical Association*, *116*(534), 969–983, doi: 10.1080/01621459.2019.1700130.

Awan, J., and A. Slavković (2018), Differentially private uniformly most powerful tests for binomial data, in *Advances in Neural Information Processing Systems*, vol. 31, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Curran Associates, Inc., New York, U.S.

Bahadur, R. R. (1971), *Some limit theorems in statistics*, SIAM, 3600 Market Street, 6th Floor Philadelphia, PA 19104 USA.

Barrientos, A. F., J. P. Reiter, A. Machanavajjhala, and Y. Chen (2019), Differentially private significance tests for regression coefficients, *Journal of Computational and Graphical Statistics*, *28*(2), 440–453.

Bauer, D. J., W. C. Belzak, and V. T. Cole (2020), Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning, *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55, doi: 10.1080/10705511.2019.1642754.

Belzak, W., and D. J. Bauer (2020), Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning., *Psychological Methods*, *25*(6), 673–690, doi:10.1037/met0000253.

Bengio, Y., and O. Delalleau (2009), Justifying and generalizing contrastive divergence, *Neural computation*, *21*(6), 1601–1621.

Benjamini, Y., and Y. Hochberg (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300, doi:10.1111/j.2517-6161.1995.tb02031.x.

Berk, R. H. (1972), Consistency and asymptotic normality of MLE's for exponential models, *The Annals of Mathematical Statistics*, *43*, 193–204.

Bhaskar, S. A. (2016), Probabilistic low-rank matrix completion from quantized measurements, *The Journal of Machine Learning Research*, *17*(1), 2131–2164.

Bhaskar, S. A., and A. Javanmard (2015), 1-bit matrix completion under exact low-rank constraint, in *2015 49th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE.

Birnbaum, A. L. (1968), Some latent trait models and their use in inferring an examinee's ability, *Statistical theories of mental test scores*.

Biswas, P., T. Lian, T. Wang, and Y. Ye (2006), Semidefinite programming based algorithms for sensor network localization, *ACM Transactions on Sensor Networks (TOSN)*, *2*(2), 188–220.

Bowen, C. M., and F. Liu (2020), Comparative study of differentially private data synthesis methods, *Statistical Science*, *35*(2), 280–307.

Boyd, S., and L. Vandenberghe (2004), *Convex optimization*, Cambridge University Press, Cambridge, UK.

Bradley, R. A., and M. E. Terry (1952), Rank analysis of incomplete block designs: I. the method of paired comparisons, *Biometrika*, *39*(3/4), 324–345.

Cai, L., S. Du Toit, and D. Thissen (2011), Irtpro: Flexible, multidimensional, multiple categorical irt modeling [computer software], *Chicago, IL: Scientific Software International*.

Cai, T., and W. Zhou (2013), A max-norm constrained minimization approach to 1-bit matrix completion, *The Journal of Machine Learning Research*, *14*, 3619–3647.

Campbell, Z., A. Bray, A. Ritz, and A. Groce (2018), Differentially private ANOVA testing, in *2018 1st International Conference on Data Intelligence and Security (ICDIS)*, pp. 281–285, doi:10.1109/ICDIS.2018.00052.

Candell, G. L., and F. Drasgow (1988), An iterative procedure for linking metrics and assessing item bias in item response theory, *Applied Psychological Measurement*, *12*(3), 253–260, doi:10.1177/014662168801200304.

Candès, E. J., and B. Recht (2009), Exact matrix completion via convex optimization, *Foundations of Computational Mathematics*, *9*(6), 717–772.

Candès, E. J., and T. Tao (2010), The power of convex relaxation: Near-optimal matrix completion, *IEEE Transactions on Information Theory*, *56*(5), 2053–2080.

Canonne, C. L., G. Kamath, and T. Steinke (2020), The discrete Gaussian for differential privacy, *Advances in Neural Information Processing Systems*, *33*, 15,676–15,688.

Cao, M., L. Tay, and Y. Liu (2017), A Monte Carlo study of an iterative Wald test procedure for DIF analysis, *Educational and Psychological Measurement*, *77*(1), 104–118, doi:10.1177/0013164416637104.

Carreira-Perpinan, M. A., and G. E. Hinton (2005), On contrastive divergence learning., in *Aistats*, vol. 10, pp. 33–40, Citeseer.

Charest, A. S. (2011), How can we analyze differentially private synthetic datasets?, *Journal of Privacy and Confidentiality*, *2*(2), doi:10.29012/jpc.v2i2.589.

Chatterjee, S., P. Diaconis, A. Sly, et al. (2011), Random graphs with a given degree sequence, *Annals of Applied Probability*, *21*(4), 1400–1435.

Chaudhuri, K., A. Sarwate, and K. Sinha (2012), Near-optimal differentially private principal components, in *Advances in Neural Information Processing Systems*, pp. 989–997.

Chen, Y., J. Liu, G. Xu, and Z. Ying (2015), Statistical analysis of Q-matrix based diagnostic classification models, *Journal of the American Statistical Association*, *110*(510), 850–866, doi:10.1080/01621459.2014.934827.

Chen, Y., S. A. Culpepper, Y. Chen, and J. Douglas (2018), Bayesian estimation of the DINA Q matrix, *Psychometrika*, *83*(1), 89–108, doi:10.1007/s11336-017-9579-4.

Chen, Y., J. Fan, C. Ma, and Y. Yan (2019a), Inference and uncertainty quantification for noisy matrix completion, *Proceedings of the National Academy of Sciences*, *116*(46), 22,931–22,937.

Chen, Y., X. Li, and S. Zhang (2019b), Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis, *Psychometrika*, *84*(1), 124–146.

Chen, Y., Y. Chi, J. Fan, C. Ma, and Y. Yan (2020a), Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization, *SIAM Journal on Optimization*, *30*(4), 3098–3121.

Chen, Y., X. Li, and S. Zhang (2020b), Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications, *Journal of the American Statistical Association*, *115*, 1756–1770.

Chen, Y., C. Li, and G. Xu (2021a), DIF statistical inference and detection without knowing anchoring items, *arXiv preprint arXiv:2110.11112*.

Chen, Y., C. Li, and G. Xu (2021b), A note on statistical inference for noisy incomplete 1-bit matrix, *arXiv preprint arXiv:2105.01769*.

Chiu, C. Y. (2013), Statistical refinement of the Q-matrix in cognitive diagnosis, *Applied Psychological Measurement*, *37*(8), 598–618, doi:10.1177/0146621613488436.

Choi, K., Y. S. Lee, and Y. S. Park (2015), What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics, *Eurasia Journal of Mathematics, Science and Technology Education*, *11*, 1563–1577, doi:10.12973/eurasia.2015.1421a.

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

Clauser, B., K. Mazor, and R. K. Hambleton (1993), The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure, *Applied Measurement in Education*, *6*(4), 269–279, doi:10.1207/s15324818ame0604_2.

Collins, M., A. Globerson, T. K. Koo, X. Carreras, and P. L. Bartlett (2008), Exponentiated gradient algorithms for conditional random fields and max-margin markov networks, *Journal of Machine Learning Research*, *9*, 1775–1822.

Cordero, J. M., C. Polo, D. Santín, and R. Simancas (2018), Efficiency measurement and cross-country differences among schools: A robust conditional nonparametric analysis, *Economic Modelling*, *74*, 45–60.

Cordero, J. M., M. Gil-Izquierdo, and F. Pedraja-Chaparro (2022), Financial education and student financial literacy: A cross-country analysis using PISA 2012 data, *The Social Science Journal*, *59*(1), 15–33.

Couch, S., Z. Kazan, K. Shi, A. Bray, and A. Groce (2019), Differentially private nonparametric hypothesis testing, in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 737–751.

Culpepper, S. (2019), Estimating the cognitive diagnosis Q-matrix with expert knowledge: Application to the fraction-subtraction dataset, *Psychometrika*, *84*(2), 333–357, doi:10.1007/s11336-018-9643-8.

D'Agostino, A. (2016), Does private school competition improve country-level student achievement? A cross-country analysis using PISA 2012., Ph.D. thesis, Georgetown University.

Davenport, M. A., Y. Plan, E. Van Den Berg, and M. Wootters (2014), 1-bit matrix completion, *Information and Inference: A Journal of the IMA*, *3*, 189–223.

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

de la Torre, and C. Y. Chiu (2016), A general method of empirical Q-matrix validation., *Psychometrika.*, *81*(2), 253–73, doi:0.1007/s11336-015-9467-8.

de la Torre, J. (2008), An empirically based method of Q-matrix validation for the DINA model: Development and applications, *Journal of Educational Measurement*, *45*(4), 343–362.

de la Torre, J., L. A. van der Ark, and G. Rossi (2018), Analysis of clinical data from a cognitive diagnosis modeling framework, *Measurement and Evaluation in Counseling and Development*, *51*(4), 281–296, doi:10.1080/07481756.2017.1327286.

DeCarlo, L. T. (2012), Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model., *Applied Psychological Measurement*, *36(6)*, 447–468.

Degue, K. H., and J. Le Ny (2018), On differentially private Gaussian hypothesis testing, in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 842–847, IEEE.

Ding, B., H. Nori, P. Li, and J. Allen (2018), Comparing population means under local differential privacy: with significance and power, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

Dorans, N. J., and E. Kulick (1986), Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test, *Journal of Educational Measurement*, *23*(4), 355–368, doi:10.1111/j.1745-3984.1986.tb00255.x.

Drechsler, J. (2011), *Synthetic datasets for statistical disclosure control: theory and implementation*, vol. 201, Springer Science & Business Media, Berlin, Germany.

Dwork, C., and A. Roth (2014), The algorithmic foundations of differential privacy., *Foundations and Trends in Theoretical Computer Science*, *9*(3-4), 211–407.

Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor (2006a), Our data, ourselves: privacy via distributed noise generation, in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006b), Calibrating noise to sensitivity in private data analysis, in *Theory of Cryptography Conference*, pp. 265–284, Springer.

Dwork, C., M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin (2010), Pan-private streaming algorithms., in *ICS*, pp. 66–80.

Escorial, S., and M. J. Navas (2007), Analysis of the gender variable in the Eysenck Personality Questionnaire–revised scales using differential item functioning techniques, *Educational and Psychological Measurement*, *67*(6), 990–1001, doi:10.1177/0013164406299108.

Eysenck, S. B., H. J. Eysenck, and P. Barrett (1985), A revised version of the psychoticism scale, *Personality and Individual Differences*, *6*(1), 21–29, doi:10.1016/0191-8869(85)90026-1.

Ferrando, C., S. Wang, and D. Sheldon (2022), Parametric bootstrap for differentially private confidence intervals, in *International Conference on Artificial Intelligence and Statistics*, pp. 1598–1618, PMLR.

Fetvadjiev, V. H., and F. J. van de Vijver (2015), Measures of personality across cultures, in *Measures of Personality and Social Psychological Constructs*, edited by G. Boyle, D. H. Saklofske, and G. Matthews, pp. 752–776, Academic Press, London, UK, doi:10.1016/B978-0-12-386915-9.00026-7.

Fidalgo, A., G. J. Mellenbergh, and J. Muñiz (2000), Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures, *Methods of Psychological Research Online*, *5*(3), 43–53, doi:10.1.1.586.7639.

Frick, H., C. Strobl, and A. Zeileis (2015), Rasch mixture models for DIF detection: A comparison of old and new score specifications, *Educational and Psychological Measurement*, *75*(2), 208–234, doi:10.1177/0013164414536183.

Friedman, A., S. Berkovsky, and M. A. Kaafar (2016), A differential privacy framework for matrix factorization recommender systems, *User Modeling and User-Adapted Interaction*, *26*(5), 425–458.

Gaboardi, M., H. Lim, R. Rogers, and S. Vadhan (2016), Differentially private Chi-squared hypothesis testing: Goodness of fit and independence testing, in *International Conference on Machine Learning*, pp. 2111–2120, PMLR.

García, P., J. Olea, and J. de la Torre (2014), Application of cognitive diagnosis models to competency-based situational judgment tests., *Psicothema.*, *26(3)*, 372–7, doi:10.7334/psicothema2013.322.

Geng, Q., and P. Viswanath (2014), The optimal mechanism in differential privacy, in *2014 IEEE international symposium on information theory*, pp. 2371–2375, IEEE.

Geng, Q., and P. Viswanath (2015), The optimal noise-adding mechanism in differential privacy, *IEEE Transactions on Information Theory*, *62*(2), 925–951.

Ghosh, A., T. Roughgarden, and M. Sundararajan (2012), Universally utility-maximizing privacy mechanisms, *SIAM Journal on Computing*, *41*(6), 1673–1693.

Goldberg, D., D. Nichols, B. M. Oki, and D. Terry (1992), Using collaborative filtering to weave an information tapestry, *Communications of the ACM*, *35*(12), 61–70.

Golle, P., and K. Partridge (2009), On the anonymity of home/work location pairs, in *Pervasive Computing*, pp. 390–397, Springer, Berlin, Heidelberg.

González, J., and M. Wiberg (2017), *Applying test equating methods*, Springer, New York.

Gu, Y., and G. Xu (2019), The sufficient and necessary condition for the identifiability and estimability of the DINA model., *Psychometrika*, *84*(2), 468–483, doi:10.1007/s11336-018-9619-8.

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

Gu, Y., and G. Xu (2020b), Sufficient and necessary conditions for the identifiability of the Q-matrix, *Statistica Sinica, to appear*.

Gutman, R., G. DeDe, D. Caplan, and J. S. Liu (2011), Rasch model and its extensions for analysis of aphasic deficits in syntactic comprehension, *Journal of the American Statistical Association*, *106*(496), 1304–1316.

Haberman, S. J. (1977), Maximum likelihood estimates in exponential response models, *The Annals of Statistics*, *5*, 815–841.

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

Hagquist, C., and D. Andrich (2017), Recent advances in analysis of differential item functioning in health research using the Rasch model, *Health and Quality of Life Outcomes*, *15*(1), 1–8.

Han, R., R. Ye, C. Tan, K. Chen, et al. (2020), Asymptotic theory of sparse Bradley–Terry model, *Annals of Applied Probability*, *30*, 2491–2515.

Hartz, S. (2002), A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality., *Unpublished doctoral dissertation.*

Hay, M., A. Machanavajjhala, G. Miklau, Y. Chen, and D. Zhang (2016), Principled evaluation of differentially private algorithms using dpbench, in *Proceedings of the 2016 International Conference on Management of Data*, pp. 139–154.

He, X., and Q. Shao (2000), On parameters of increasing dimensions, *Journal of Multivariate Analysis*, *73*(1), 120–135.

Henson, R. A., J. L. Templin, and J. T. Willse (2008), Defining a family of cognitive diagnosis models using log-linear models with latent variables, *Psychometrika*, *74*(2), 191, doi:10.1007/s11336-008-9089-5.

Hinton, G. E. (2002), Training products of experts by minimizing contrastive divergence, *Neural Computation*, *14*(8), 1771–1800.

Hinton, G. E., and R. R. Salakhutdinov (2006), Reducing the dimensionality of data with neural networks, *Science*, *313*(5786), 504–507, doi:10.1126/science.1127647.

Hovdhaugen, E., F. Karlsson, C. C. Henriksen, and B. Sigurd (2000), *The history of linguistics in the Nordic countries*, Societas Scientiarum Fennica Helsinki, Finland.

Huang, P. H. (2018), A penalized likelihood method for multi-group structural equation modelling, *British Journal of Mathematical and Statistical Psychology*, *71*(3), 499–522, doi:10.1111/bmsp.12130.

Jiang, B., T. Y. Wu, Y. Jin, W. H. Wong, et al. (2018), Convergence of contrastive divergence algorithm in exponential family, *The Annals of Statistics*, *46*(6A), 3067–3098.

Johnson, A., and V. Shmatikov (2013), Privacy-preserving data exploration in genome-wide association studies, in *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1079–1087.

Joo, S. H., L. Khorramdel, K. Yamamoto, H. J. Shin, and F. Robin (2021), Evaluating item fit statistic thresholds in PISA: Analysis of cross-country comparability of cognitive items, *Educational Measurement: Issues and Practice*, *40*(2), 37–48.

Junker, B. W., and K. Sijtsma (2001), Cognitive assessment models with few assumptions, and connections with nonparametric item response theory, *Applied Psychological Measurement*, *25*(3), 258–272, doi:10.1177/01466210122032064.

Kairouz, P., K. Bonawitz, and D. Ramage (2016), Discrete distribution estimation under local privacy, in *International Conference on Machine Learning*, pp. 2436–2444, PMLR.

Kantorovich, L., and G. Akilov (1964), *Functional analysis in normed spaces (translated by D.G. Brown)*, Pergamon Presss, Oxford.

Karwa, V., and A. Slavković (2016), Inference using noisy degrees: differentially private model and synthetic graphs, *The Annals of Statistics*, *44*(1), 87–112.

Karwa, V., P. N. Krivitsky, and A. B. Slavković (2017), Sharing social network data: differentially private estimation of exponential family random-graph models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(3), 481–500.

Keshavan, R. H., A. Montanari, and S. Oh (2010), Matrix completion from noisy entries, *The Journal of Machine Learning Research*, *11*, 2057–2078.

Kidwell, P., G. Lebanon, and K. Collins Thompson (2011), Statistical estimation of word acquisition with application to readability prediction, *Journal of the American Statistical Association*, *106*(493), 21–30.

Kim, S. H., A. S. Cohen, and T. H. Park (1995), Detection of differential item functioning in multiple groups, *Journal of Educational Measurement*, *32*(3), 261–276, doi:10.1111/j.1745-3984.1995.tb00466.x.

Klopp, O., J. Lafond, E. Moulines, and J. Salmon (2015), Adaptive multinomial matrix completion, *Electronic Journal of Statistics*, *9*(2), 2950–2975.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press, Cambridge, UK.

Koenker, R., S. Portnoy, P. T. Ng, A. Zeileis, P. Grosjean, and B. D. Ripley (2018), Package 'quantreg', *Cran R-project. org*.

Kolen, M. J., and R. L. Brennan (2014), *Test equating, scaling, and linking: Methods and practices*, Springer Science & Business Media, New York.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011), Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, *The Annals of Statistics*, *39*(5), 2302–2329.

Kopf, J., A. Zeileis, and C. Strobl (2015a), Anchor selection strategies for DIF analysis: Review, assessment, and new approaches, *Educational and Psychological Measurement*, *75*(1), 22–56, doi:10.1177/0013164414529792.

Kopf, J., A. Zeileis, and C. Strobl (2015b), A framework for anchor methods and an iterative forward approach for DIF detection, *Applied Psychological Measurement*, *39*(2), 83–103, doi:10.1177/0146621614544195.

Kornacki, S. (2011), Why healthcare may not doom Mitt Romney after all, `https://www.salon.com/2011/05/12/romney_healthcare/`.

Kuhn, H. W. (1955), The Hungarian method for the assignment problem, *Naval Research Logistics Quarterly*, *2*(1-2), 83–97.

languages Slovene, O., and I. Italians (2002), Republic of Slovenia Republika Slovenija (Slovene), *Ethnic Groups*, *4*(83), 83.

Larochelle, H., and Y. Bengio (2008), Classification using discriminative restricted Boltzmann machines, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 536–543, ACM, New York, USA, doi:10.1145/1390156.1390224.

Lee, Y. S., Y. S. Park, and D. Taylan (2011), A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007, *International Journal of Testing*, *11*, 144–177, doi:10.1080/15305058.2010.534571.

Lehti, M., and D. J. Smith (2003), *Post-Cold War identity politics: northern and Baltic experiences*, Psychology Press, Hove, East Sussex, United Kingdom.

Li, C., C. Ma, and G. Xu (2022a), Learning large Q-matrix by restricted Boltzmann machines, *Psychometrika, to appear*, doi:10.1007/s11336-021-09828-4.

Li, C., N. Wang, and G. Xu (2022b), Inference for optimal differential privacy procedures for frequency tables., *Journal of Data Science*, *20*(2).

Lindsay, B., C. C. Clogg, and J. Grego (1991), Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis, *Journal of the American Statistical Association*, *86*(413), 96–107.

Lingefjärd, T. (2018), The history of mathematics education in the Nordic countries, including sweden, *University of Gothenburg*.

Little, R. J. (1993), Statistical analysis of masked data, *Journal of Official Statistics*, *9*(2), 407–426.

Liu, C., X. He, T. Chanyaswad, S. Wang, and P. Mittal (2019), Investigating statistical privacy frameworks from the perspective of hypothesis testing, *Proceedings on Privacy Enhancing Technologies*, *3*, 233–254.

Liu, J., G. Xu, and Z. Ying (2012), Data-driven learning of Q-matrix, *Applied Psychological Measurement*, *36*(7), 548–564, doi:10.1177/0146621612456591.

Liu, Z., and L. Vandenberghe (2010), Interior-point method for nuclear norm approximation with application to system identification, *SIAM Journal on Matrix Analysis and Applications*, *31*(3), 1235–1256.

Long, P. M., and R. A. Servedio (2010), Restricted Boltzmann machines are hard to approximately evaluate or simulate, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, p. 703–710, Omnipress, Madison, WI, USA.

Lord, F. M. (1980), *Applications of item response theory to practical testing problems*, Routledge, New York.

Louis, T. A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 226–233, doi:10.1111/j.2517-6161.1982.tb01203.x.

M. Clifford, R., D. Bryant, M. Burchinal, O. Barbarin, D. Early, C. Howes, R. Pianta, and P. Winton (2017), *National Center for Early Development and Learning Multistate Study of Pre-Kindergarten*, Inter-university Consortium for Political and Social Research [distributor], doi:10.3886/ICPSR04283.v4.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008), Privacy: Theory meets practice on the map, in *2008 IEEE 24th international conference on data engineering*, pp. 277–286, IEEE.

MacKay, D. (2001), Failures of the one-step learning algorithm, in *Available electronically at http://www. inference. phy. cam. ac. uk/mackay/abstracts/gbm. html*.

Magis, D., F. Tuerlinckx, and P. De Boeck (2015), Detection of differential item functioning using the lasso approach, *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135, doi:10.3102/1076998614559747.

Mantel, N., and W. Haenszel (1959), Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute*, *22*(4), 719–748, doi:10.1093/jnci/22.4.719.

Marshall, P. J. (2001), *The Cambridge illustrated history of the British Empire*, Cambridge University Press, Shaftesbury Road, Cambridge, CB2 8BS, United Kingdom.

May, H. (2006), A multilevel Bayesian item response theory method for scaling socioeconomic status in international studies of education, *Journal of Educational and Behavioral Statistics*, *31*(1), 63–79, doi:10.3102/10769986031001063.

Mazurek, J., C. F. García, and C. P. Rico (2021), Inequality and students' PISA 2018 performance: a cross-country study, *Comparative Economic Research. Central and Eastern Europe*, *24*(3), 163–183.

McClure, D., and J. P. Reiter (2012), Differential privacy and statistical disclosure risk measures: an investigation with binary synthetic data., *Trans. Data Priv.*, *5*(3), 535–552.

McConney, A., M. C. Oliver, A. Woods Mcconney, R. Schibeci, and D. Maor (2014), Inquiry, engagement, and literacy in science: A retrospective, cross-national analysis using PISA 2006, *Science Education*, *98*(6), 963–980.

Meinshausen, N., and B. Yu (2009), Lasso-type recovery of sparse representations for high-dimensional data, *The Annals of Statistics*, *37*, 246–270, doi:10.1214/07-AOS582.

Millsap, R. E. (2012), *Statistical approaches to measurement invariance*, Routledge, New York, NY.

Mohammed, N., R. Chen, B. C. Fung, and P. S. Yu (2011), Differentially private data release for data mining, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–501.

Narayanan, A., and V. Shmatikov (2008), Robust de-anonymization of large sparse datasets, in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125.

Negahban, S., and M. J. Wainwright (2012), Restricted strong convexity and weighted matrix completion: optimal bounds with noise, *The Journal of Machine Learning Research*, *13*(1), 1665–1697.

Oort, F. J. (1998), Simulation study of item bias detection with restricted factor analysis, *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(2), 107–124, doi:10.1080/10705519809540095.

Parlapiano, A., W. Andrews, J. Lee, and R. Shorey (2017), How each senator voted on Obamacare Repeal Proposals, `https://www.nytimes.com/interactive/2017/07/25/us/politics/senate-votes-repeal-obamacare.html`.

PISA (2020), Organisation for Economic Co-operation and Development., & Programme for International Student Assessment, *PISA 2018 technical report. Paris: OECD.*

Poole, K. T., and H. Rosenthal (1991), Patterns of congressional voting, *American Journal of Political Science*, *35*, 228–278.

Poole, K. T., and H. Rosenthal (2001), Dnominate after 10 years: A comparative update to congress: a political-economic history of roll-call voting, *Legislative Studies Quarterly*, *26*, 5–29.

Poole, K. T., H. Rosenthal, and K. Koford (1991), On dimensionalizing roll call votes in the us congress, *The American Political Science Review*, *85*, 955–976.

Portnoy, S. (1988), Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, *The Annals of Statistics*, *16*, 356–366.

Quick, H. (2021), Generating Poisson-distributed differentially private synthetic data, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(3), 1093–1108.

Raab, G. M., B. Nowok, and C. Dibben (2016), Practical data synthesis for large samples, *Journal of Privacy and Confidentiality*, *7*(3), 67–97.

Raghunathan, T. E., J. P. Reiter, and D. B. Rubin (2003), Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics*, *19*(1), 1–16.

Raju, N. S. (1988), The area between two item characteristic curves, *Psychometrika*, *53*(4), 495–502.

Raju, N. S. (1990), Determining the significance of estimated signed and unsigned areas between two item response functions, *Applied Psychological Measurement*, *14*(2), 197–207, doi:10.1177/014662169001400208.

Rao, C. R. (1973), *Linear statistical inference and its applications*, Wiley New York.

Rasch, G. (1960), *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.*, Nielsen & Lydiche, Oxford, England.

Reiter, J. P. (2005), Using CART to generate partially synthetic public use microdata, *Journal of Official Statistics*, *21*(3), 441–462.

Renfrew, C. (1989), The origins of Indo-European languages, *Scientific American*, *261*(4), 106–115.

Rice, K. M. (2004), Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications, *Journal of the American Statistical Association*, *99*(466), 510–522.

Rinaldo, A., S. Petrović, and S. E. Fienberg (2013), Maximum lilkelihood estimation in the β-model, *The Annals of Statistics*, *41*, 1085–1110.

Rinott, Y., C. M. O'Keefe, N. Shlomo, and C. Skinner (2018), Confidentiality and differential privacy in the dissemination of frequency tables, *Statistical Science*, *33*(3), 358–385, doi:10.1214/17-STS641.

Robitzsch, A., T. Kiefer, A. C. George, A. Uenlue, and M. A. Robitzsch (2020), Package 'cdm', *Handbook of diagnostic classification models. New York: Springer.*

Rogers, R., A. Roth, A. Smith, and O. Thakkar (2016), Max-information, differential privacy, and post-selection hypothesis testing, in *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 487–494, IEEE.

Rosasco, L. (2009), Sparsity based regularization., *MIT class notes.*

Rubin, D. B. (1993), Statistical disclosure limitation, *Journal of Official Statistics*, *9*(2), 461–468.

Salakhutdinov, R., A. Mnih, and G. Hinton (2007), Restricted Boltzmann machines for collaborative filtering, in *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 791–798, ACM, New York, NY, USA, doi:10.1145/1273496.1273596.

Schauberger, G., and P. Mair (2020), A regularization approach for the detection of differential item functioning in generalized partial credit models, *Behavior Research Methods*, *52*(1), 279–294, doi:10.3758/s13428-019-01224-2.

Schlueter, J. (2014), Restricted Boltzmann machine derivations, *Notes.*

Schwarz, G. (1978), The Bayesian information criterion, *Annals of Statistics*, *6*, 461–464.

Shealy, R., and W. Stout (1993), A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF, *Psychometrika*, *58*(2), 159–194, doi:10.1007/BF02294572.

Sheffet, O. (2017), Differentially private ordinary least squares, in *International Conference on Machine Learning*, pp. 3105–3114, PMLR.

Shor, N. Z. (2012), *Minimization methods for non-differentiable functions*, Springer, New York, NY.

Simons, G., and Y. Yao (1999), Asymptotics when the number of parameters tends to infinity in the bradley-terry model for paired comparisons, *The Annals of Statistics*, *27*, 1041–1060.

Smolensky, P. (1986), Information processing in dynamical systems: Foundations of harmony theory, *Tech. rep.*, Colorado University at Boulder Department of Computer Science.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic (2018), General and specific utility measures for synthetic data, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(3), 663–688.

Soares, T. M., F. B. Gonçalves, and D. Gamerman (2009), An integrated Bayesian model for DIF analysis, *Journal of Educational and Behavioral Statistics*, *34*(3), 348–377, doi:10.3102/1076998609332752.

Steenkamp, J.-B. E., and H. Baumgartner (1998), Assessing measurement invariance in cross-national consumer research, *Journal of Consumer Research*, *25*(1), 78–90, doi:10.1086/209528.

Su, Y. L., K. Choi, W. Lee, T. Choi, and M. McAninch (2013), Hierarchical cognitive diagnostic analysis for timss 2003 mathematics, *Centre for Advanced Studies in Measurement and Assessment*, *35*, 1–71.

Subačius, G., and D. Tekorienė (2002), *The Lithuanian language: Traditions and trends*, The Lithuanian Institute, Tilto g. 17, Vilnius 01101, Lithuania.

Sutskever, I., and T. Tieleman (2010), On the convergence properties of contrastive divergence, in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 789–795.

Swaminathan, H., and H. J. Rogers (1990), Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement*, *27*(4), 361–370, doi:10.1111/j.1745-3984.1990.tb00754.x.

Sweeney, L. (2013), Matching known patients to health records in Washington state data, *arXiv preprint arXiv:1307.1370*.

Task, C., and C. Clifton (2016), Differentially private significance testing on paired-sample data., in *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)*, pp. 153–161, doi:10.1137/1.9781611974348.18.

Tay, L., A. W. Meade, and M. Cao (2015), An overview and practical guide to IRT measurement equivalence analysis, *Organizational Research Methods*, *18*(1), 3–46, doi:10.1177/1094428114553062.

Tay, L., Q. Huang, and J. K. Vermunt (2016), Item response theory with covariates (IRT-C) assessing item recovery and differential item functioning for the three-parameter logistic model, *Educational and Psychological Measurement*, *76*(1), 22–42, doi:10.1177/0013164415579488.

Templin, J., and R. Henson (2006), Measurement of psychological disorders using cognitive diagnosis models., *Psychological methods*, *11(3)*, 287–305, doi:10.1037/1082-989X.11.3.287.

Thissen, D. (1988), Use of item response theory in the study of group differences in trace lines, in *Test validity*, edited by H. E. Wainer and H. I. Braun, pp. 147–172, Lawrence Erlbaum Associates, Inc, Mahwah, NJ.

Thissen, D. (2001), Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning, *Chapel Hill: University of North Carolina at Chapel Hill*.

Thissen, D., L. Steinberg, and M. Gerrard (1986), Beyond group-mean differences: The concept of item bias, *Psychological Bulletin*, *99(1)*, 118–128, doi:10.1037/0033-2909.99.1.118.

Thissen, D., L. Steinberg, and H. Wainer (1993), Detection of differential item functioning using the parameters of item response models, in *Differential item functioning*, edited by P. W. Holland and H. Wainer, pp. 67–113, Lawrence Erlbaum Associates, Inc, Mahwah, NJ.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 267–288, doi:10.1111/j.2517-6161.1996.tb02080.x.

Tsuruoka, Y., J. Tsujii, and S. Ananiadou (2009), Stochastic gradient descent training for L1-regularized log-linear models with cumulative penalty, in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 477–485, Association for Computational Linguistics.

Tutz, G., and G. Schauberger (2015), A penalty approach to differential item functioning in Rasch models, *Psychometrika*, *80*(1), 21–43, doi:10.1007/s11336-013-9377-6.

Udell, M., C. Horn, R. Zadeh, and S. Boyd (2016), Generalized low rank models, *Foundations and Trends® in Machine Learning*, *9*, 1–118.

van der Linden, W. J., and R. K. Hambleton (2013), *Handbook of modern item response theory*, Springer, New York, NY.

van der Vaart, A. W. (2000), *Asymptotic statistics*, Cambridge University Press, Cambridge, UK.

von Davier, M. (2005), A general diagnostic model applied to language testing data (ETS research report RR-05-16)., *Princeton: Educational Testing Service.*

von Davier, M. (2008), A general diagnostic model applied to language testing data, *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307, doi:10.1348/000711007X193957.

Vu, D., and A. Slavkovic (2009), Differential privacy for clinical trial data: Preliminary evaluations, in *2009 IEEE International Conference on Data Mining Workshops*, pp. 138–143, doi:10.1109/ICDMW.2009.52.

Wang, L. (2011), GEE analysis of clustered binary data with diverging number of covariates, *The Annals of Statistics*, *39*(1), 389–417.

Wang, R., Y. F. Li, X. Wang, H. Tang, and X. Zhou (2009a), Learning your identity and disease from research papers: information leaks in genome wide association study, in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, pp. 534–544.

Wang, W. C., and Y. H. Su (2004), Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method, *Applied Measurement in Education*, *17*(2), 113–144, doi:10.1207/s15324818ame1702_2.

Wang, W. C., and Y. L. Yeh (2003), Effects of anchor item methods on differential item functioning detection with the likelihood ratio test, *Applied Psychological Measurement*, *27*(6), 479–498, doi:10.1177/0146621603259902.

Wang, W. C., C. L. Shih, and C. C. Yang (2009b), The MIMIC method with scale purification for detecting differential item functioning, *Educational and Psychological Measurement*, *69*(5), 713–731, doi:10.1177/0013164409332228.

Wang, Y., J. Lee, and D. Kifer (2015a), Revisiting differentially private hypothesis tests for categorical data, *arXiv preprint arXiv:1511.03376*.

Wang, Y. X., S. Fienberg, and A. Smola (2015b), Privacy for free: posterior sampling and stochastic gradient Monte Carlo, in *International Conference on Machine Learning*, pp. 2493–2502.

Wasserman, L., and S. Zhou (2010), A statistical framework for differential privacy, *Journal of the American Statistical Association*, *105*(489), 375–389.

Wilson, P. H. (2011), *The Holy Roman Empire 1495-1806*, Macmillan International Higher Education, 4 Crinan Street. London, England N1 9SQ, United Kingdom.

Woods, C. M., L. Cai, and M. Wang (2013), The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT, *Educational and Psychological Measurement*, *73*(3), 532–547, doi:10.1177/0013164412464875.

Wu, Z., M. Deloria-Knoll, and S. L. Zeger (2016), Nested partially latent class models for dependent binary data; estimating disease etiology, *Biostatistics*, *18*(2), 200–213, doi:10.1093/biostatistics/kxw037.

Xia, D., and M. Yuan (2021), Statistical inferences of linear forms for noisy matrix completion, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*(1), 58–77.

Xu, G. (2017), Identifiability of restricted latent class models with binary responses, *Annals of Statistics*, *45*(2), 675–707, doi:10.1214/16-AOS1464.

Xu, G., and Z. Shang (2018), Identifying latent structures in restricted latent class models, *Journal of the American Statistical Association*, *113*(523), 1284–1295, doi:10.1080/01621459.2017.1340889.

Yu, F., S. E. Fienberg, A. B. Slavković, and C. Uhler (2014), Scalable privacy-preserving data sharing methodology for genome-wide association studies, *Journal of Biomedical Informatics*, *50*, 133–141.

Yuan, K., H. Liu, and Y. Han (2021), Differential item functioning analysis without a priori information on anchor items: QQ plots and graphical test, *Psychometrika*, *86*, 345–377, doi:10.1007/s11336-021-09746-5.

Yuille, A. L. (2004), The convergence of contrastive divergences, *Advances in neural information processing systems*, *17*, 1593–1600.

Zwick, R., and D. T. Thayer (2002), Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test, *Applied Psychological Measurement*, *26*(1), 57–76, doi: 10.1177/0146621602026001004.

Zwick, R., D. T. Thayer, and C. Lewis (2000), Using loss functions for DIF detection: An empirical Bayes approach, *Journal of Educational and Behavioral Statistics*, *25*(2), 225–247, doi:10.3102/10769986025002225.