Interaction Estimations in High-dimensional Statistical Models

by

Lizhu Tao

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

November 2020
## Contents

List of Tables iv  
List of Figures v  
Acknowledgments vii  
Declarations viii  
Abstract ix  

Chapter 1 Introduction 1  
1.1 Sparse Ising Model for $P \gg n$ with Latent variablEs . . . . . . . . . . 1  
1.1.1 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . 2  
1.1.2 Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . 6  
1.2 Decorrelated Lasso and Interaction Screening . . . . . . . . . . . . . . 6  
1.3 Thesis Organization . . . . . . . . . . . . . . . . . . . . . . . . . . 7  

Chapter 2 Sparse Ising Model for $P \gg n$ with Latent variablEs 8  
2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8  
2.1.1 Graphical Models . . . . . . . . . . . . . . . . . . . . . . . . . 9  
2.1.2 Markov Properties . . . . . . . . . . . . . . . . . . . . . . . . 9  
2.1.3 Pairwise Markov Random Field . . . . . . . . . . . . . . . . . 10  
2.1.4 The Ising Model . . . . . . . . . . . . . . . . . . . . . . . . . 10  
2.1.5 Estimation Methods of the Ising Model . . . . . . . . . . . . . 12  
2.2 The SIMPLE Method . . . . . . . . . . . . . . . . . . . . . . . . . 16  
2.2.1 Model Setup . . . . . . . . . . . . . . . . . . . . . . . . . . . 16  
2.2.2 Estimation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20  
2.3 Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22  
2.3.1 Randomized Block Coordinate Descent . . . . . . . . . . . . . . 23  
2.3.2 Majorization-Minimization Principle . . . . . . . . . . . . . . . 24
2.3.3 Acceleration Method ......................................... 27
2.3.4 Step Size ....................................................... 29
2.3.5 Restart .......................................................... 31
2.3.6 Full Algorithm ............................................... 32
2.4 Convergence Analysis ......................................... 36
  2.4.1 Global Convergence ....................................... 36
  2.4.2 Convergence Rate ......................................... 38
2.5 Theoretical Analysis .......................................... 38
  2.5.1 Global Error ............................................... 39
  2.5.2 Consistent Estimation of Graph Structure ................. 44
2.6 Simulation Study .............................................. 49
  2.6.1 Simulation Setting ......................................... 49
  2.6.2 Simulation Results ........................................ 51
2.7 Real-world Data Applications ............................... 54
  2.7.1 Application 1: the Stock Prices of Oil Companies ........ 54
  2.7.2 Application 2: the Exchange Rates to the US Dollar ....... 61

Chapter 3 Decorrelated Lasso and Interaction Screening 68
  3.1 Introduction .................................................. 68
    3.1.1 Our Setup .............................................. 68
    3.1.2 Related Work .......................................... 71
    3.1.3 Contributions .......................................... 72
  3.2 The DECO-IS Method ........................................ 73
  3.3 Theoretical Analysis ........................................ 75
  3.4 Simulation Study ............................................ 78

Chapter 4 Conclusion ............................................. 82

Appendix A Proofs in Chapter 2 84
  A.1 Proof of Lemma 1 ........................................... 84
  A.2 Proof of Lemma 2 .......................................... 85
  A.3 Proof of Lemma 3 .......................................... 85
  A.4 Proof of Lemma 4 .......................................... 86
  A.5 Proof of Corollary 1 ....................................... 87
  A.6 Proof of Theorem 1 ....................................... 87
  A.7 Proof of Theorem 3 ....................................... 87
  A.8 Proof of Corollary 2 ..................................... 90
  A.9 Proof of Theorem 4 ..................................... 95
List of Tables

3.1 The frequency (%) of choosing the true model in Example 1 . . . . . 80
3.2 The frequency (%) of choosing the true model in Example 2 . . . . . 81
List of Figures

2.1 The effect of the latent variable on the graph structure. In the left panel, there is one latent variable labeled as “h” in red. The black solid lines represent the edges between observed nodes, and the dashed red lines represent the edges between the observed nodes and the latent “h” node. In the right panel, the dashed black lines represent “fake” connections between observed nodes due to ignoring the latent node. ................................................................. 17

2.2 Performance of different estimation methods when there is one latent variable. The red line represents the SIMPLE method, the green line represents the Node-wise-BOTH method, the blue line represents the Node-wise-OR method, and the yellow line represents the Pseudo approach. ................................................................. 52

2.3 Performance of different estimation methods when there are three moderately connected latent variables. The red line represents the SIMPLE method, the green line represents the Node-wise-BOTH method, the blue line represents the Node-wise-OR method, and the yellow line represents the Pseudo approach. ................................................................. 53

2.4 Performance of different estimation methods when there are ten weakly connected latent variables. The red line represents the SIMPLE method, the green line represents the Node-wise-BOTH method, the blue line represents the Node-wise-OR method, and the yellow line represents the Pseudo approach. ................................................................. 54

2.5 Graph (a) and (b) show graph structures of stock prices using the SIMPLE method and the Node-wise-BOTH method. ...................... 58

2.6 Graph (a) and (b) show the circular layouts of graph structures using the SIMPLE method and the Node-wise-BOTH method. .............. 59

2.7 Figure (a) and (b) show the sparsity patterns of stock prices using the SIMPLE method and the Node-wise-BOTH method. .............. 60
2.8 Graph (a) and (b) show the graph structures of exchange rates using the SIMPLE method and the Node-wise-BOTH method. . . . . . . . 65

2.9 Graph (a) and (b) show the maps of graph structures of exchange rates using the SIMPLE method and the Node-wise-BOTH method. 66

2.10 Figure (a) and (b) show the sparsity patterns of exchange rates using the SIMPLE method and the Node-wise-BOTH method. . . . . . . . 67
Acknowledgments

I would first like to say a big thank you to my supervisor Prof Chenlei Leng for his invaluable supervision, support, encouragement and patience. Without him I would not be able to complete my PhD study.

I would also like to deeply thank my collaborator Prof Yiyuan She for his guidance, insights, and suggestions.

My sincere thanks also to Prof Degui Li and Dr Yi Yu for their help and time as committee members.

I am also grateful to our group members and friends in Warwick, and I learn a lot from them. They include Jenny, Xiaohang, Paige, Zhenzhong, Stefan, Mengchu, Sida, Qiang, Haodong, Esther, Mark, Yuan, among many others.

Finally, a very heartfelt thank you to my parents for always believing in me and encouraging me during my PhD period and my life.
Declarations

I hereby declare that this thesis represents my original work which has been done after registration for the degree of PhD at University of Warwick, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.
Abstract

High-dimensional statistics, which focus on datasets with a relatively large number of variables compared to the sample size, have seen a dramatic surge of research interest and activities in the last two decades. In this thesis, we contribute to this area by developing a new approach for understanding the conditional independence structure of multiple variables in a binary graphical model when some variables are unobserved, and proposing a new procedure for estimating interactions in a linear regression model.

Graphical models have attracted much attention in recent years due to their wide applications in science and engineering. However, challenges arise when there are variables that are unobserved, especially concerning binary graphs. For the first part of the thesis, we propose a new method, abbreviated as SIMPLE, to estimate the interactions in the undirected binary graphical model with latent components, when more variables exist than observations. We utilize a pseudo-likelihood function with an $\ell_1$ penalty for delineating the sparsity of the graph and a nuclear norm for estimating the structure of the latent variables. The resulting convex optimization problem can be efficiently solved via a newly proposed randomized block coordinate descent algorithm with convergence properties studied. Theoretical guarantees of global error and consistent graph structure estimation are provided. Simulations and real-world applications demonstrate the efficacy of the proposed approach.

For the second part of the thesis, we propose an efficient two-step procedure, called DECO-IS, to estimate the interaction terms in a two-way interaction model by focusing on its main and interaction terms separately. Computationally, the
proposed method is easy and fast to implement. Simulations under different settings
demonstrate the value of this method. Theoretical results on the consistency of the
estimation are also provided.
Chapter 1

Introduction

In the past two decades, advances in technology have produced many modern datasets with more variables than observations. In statistics, high-dimensional problems have attracted considerable interest due to a phenomenal rise in such large datasets. However, learning these datasets is not an easy task. Traditional methods, limited to a small number of predictors and a large sample size, are inadequate to solve high-dimensional problems. Complex data structures due to their high-dimensionality pose challenges in deriving statistical theories and designing computational tools.

Over time, a class of methodologies based on the concept of regularization has been developed to tackle high-dimensional issues. These methodologies are able to estimate parameters and select variables simultaneously by augmenting the likelihood function with regularizers on the parameters for encouraging sparsity. Based on this general principle, this thesis addresses two important problems in high-dimensional statistics. The first is to develop a new method which can learn interactions in the sparse Ising model with latent variables (Section 1.1). The second is to design a new approach to estimate the linear model with two-way interaction terms (Section 1.2).

1.1 Sparse Ising Model for $P \gg n$ with Latent variables

Graphical models are a canonical tool for understanding the interactions between multiple variables in complex datasets. In these models, each node usually denotes a random variable and an edge represents their interaction. The Ising model, also known as the binary Markov random field, is useful for understanding the interactions between binary variables, and has been widely used in many applications as
binary data becomes increasingly available.

However, progressively there is a common difficulty when handling real-world problems. For various reasons, in practice only a subset of all the variables are accessible to us; That is, more often than not, relevant variables are not always collected. These unobserved variables will be referred to as latent variables hereafter. The presence of these unobserved variables may severely destroy the dependency structure. However, estimating the Ising model with the presence of latent variables is challenging. First, the missing information in these latent variables introduces identifiability and tractability issues. Traditional estimation methods designed for the Ising model are inadequate to incorporate latent components. Second, learning the binary graphical model is itself hard due to the complex normalization constant in the probability function. Finally, for high-dimensional problems, the number of parameters is often greater than the sample size, leading to consistent estimation difficulties.

For the first part of the thesis, we propose a new method called Sparse Ising Model for $P \gg n$ with Latent variablEs, abbreviated as SIMPLE, to overcome the challenges outlined above.

1.1.1 Related Work

Up to now, only a few papers looked at the problem of estimating a graphical model with latent variables under the high-dimensional setting. Remarkably, the binary graphical model with the presence of latent variables has not been investigated. Chandrasekaran et al. (2012), Frot et al. (2017) and Frot et al. (2018) considered Gaussian graphical models when observed and unobserved variables are both continuous, while Chen et al. (2016); Nussbaum and Giesen (2019) considered discrete observed variables and continuous latent variables. A brief literature review is given now.

1. Chandrasekaran et al. (2012), Frot et al. (2017) and Frot et al. (2018)

Chandrasekaran et al. (2012) focused on the latent variable problem in the Gaussian graphical model in which the nodes represent a collection of joint Gaussian random variables. Denote the random variables as $X = (X_1, ..., X_p)^T$, the distribution of the Gaussian graphical model is

$$P(X|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right\},$$

where $\mu$ is the mean and $\Sigma$ is the covariance matrix. It is known that the
precision matrix $\Sigma^{-1}$ directly encodes the graph structure in the Gaussian graphical model. That is, for a node pair $i$ and $j$, there is an edge $e_{ij}$ in the underlying graphical model if and only if the $(i,j)$th entry of $\Sigma^{-1}$ is nonzero.

Chandrasekaran et al. (2012) assumed that the observed and latent variables had the joint Gaussian distribution with a sparse graph structure, and they observed that the marginal precision matrix of observed variables could be decomposed into the sum of a sparse matrix and a low-rank matrix. These two matrices revealed the conditional graph structure of these observed variables given the latent variables, as well as the effect of these latent variables.

Built on this important finding, Chandrasekaran et al. (2012) proposed a method to learn a sparse plus low-rank decomposition where sparsity is achieved by imposing an $\ell_1$ norm and the low-rank matrix is obtained by a nuclear norm. Within their framework, the authors addressed the identifiability of the Gaussian graphical model when latent variables are present. They provided theoreic guarantees regarding consistent estimation of the conditional graph structure among observed variables.

Inspired by Chandrasekaran et al. (2012), Frot et al. (2017) considered a conditional Gaussian graphical model with latent variables. Denoting $O, X$ and $H$ as the variables to condition on, the observed variables and the latent variables respectively, Frot et al. (2017) assumed that variables $(O, X, H)$ formed a conditional Gaussian graphical model, i.e., conditioned on $O$, the joint distribution of variables $(X, H)$ is normal. Similar to Chandrasekaran et al. (2012), they noticed that the parameters of the conditional Gaussian graphical model with latent variables can be decomposed into the sum of a sparse and a low-rank matrix, and then proposed a regularized maximum likelihood estimator.

Furthermore, Frot et al. (2018) focused on estimating the directed acyclic graph (DAG) in the presence of latent variables. In DAGs, nodes represent random variables and edges encode direct causal effects. Under the assumptions that the underlying DAG among the observed variables is sparse, and that only a few latent variables have a direct effect on many of the observed variables, Frot et al. (2018) proposed a two-stage approach to estimate the underlying DAG among the observed variables which were sub-Gaussian or transelliptically distributed conditional on the latent variables. They first employed the method in Chandrasekaran et al. (2012) to remove the effect of the latent variables and then estimated the direct causal effects between observed variables. Theoretical results of the consistent graphical structure recovery
and the causal effect estimation were provided.

2. Chen et al. (2016)

Unlike Chandrasekaran et al. (2012) which considered observed and latent variables to be joint Gaussian, Chen et al. (2016) assumed that the observed variables follow the Ising model, while the latent variables are continuous. The motivation of their method comes from psychology where many observed human behaviors are driven by latent attributes that are often unobserved. Item response theory (IRT) models are one category of latent variable models that aim to describe the relationship between observed responses to items by their latent attributes.

More specifically, Chen et al. (2016) focused on a two-parameter logistic model which is a popular IRT model. It assumes that a random vector $X = (X_1, ..., X_p)^T$ with binary components, and the latent parameter vector $\gamma \in \mathbb{R}^k$ is continuous. Given the latent vector, the conditional distribution of each response in the two-parameter logistic model is

$$P(X_j = 1 | \gamma) = \frac{\exp\{a_j^T \gamma + b_j\}}{1 + \exp\{a_j^T \gamma + b_j\}}.$$  

However, since the original two-parameter logistic IRT model was inadequate to capture all the dependence structure of the responses (as it only considered the interactions between observed responses and the latent vector), Chen et al. (2016) further assumed that the observed responses follow the Ising model to capture common dependence through the latent vector and the graph structure among observed responses. The marginal distribution of observed responses in the new model is

$$P(X|L, S) \propto \exp\left\{\frac{1}{2}X^T(L + S)X\right\},$$

where $L$ characterizes the effect of the latent vector and $S$ represents the graph structure among the observed responses.

Chen et al. (2016) assumed that most dependence among observed responses was driven by a few latent attributes, therefore the matrix $S$ was sparse and the dimension of the latent vector stayed low. This sparse plus low-rank problem can be solved by minimizing a regularized log-likelihood function with an $\ell_1$ penalty and a nuclear norm. The authors showed that their estimators can
consistently recover the conditional graph structure of the observed responses. However, these theoretic guarantees were valid under a setting where only the sample size is allowed to grow and the number of variables is fixed. Moreover, they did not provide sufficient and convincing simulation results since there was no comparison with other methods.


A graphical model with binary observed variables and continuous latent variables, similar to the model in Chen et al. (2016), was proposed in Nussbaum and Giesen (2019) which assumed that the observed variables $X$ followed the Ising model, and the latent variables $Z$ were conditional Gaussian variables. The joint density of the mixed model in Nussbaum and Giesen (2019) is

$$P(X, Z) \propto \exp(X^T S X + Z^T R X - \frac{1}{2} Z^T A Z),$$

and the marginal distribution of the observed variables is

$$P(X) \propto \exp \left( \langle S + \frac{1}{2} R^T A^{-1} R, X X^T \rangle \right)$$

$$= \exp \left( \langle S + L, \Phi(X) \rangle \right),$$

where $S$, $R$, and $A$ represent the interactions between observed binary variables, the interactions between the observed binary variables and the latent conditional Gaussian variables, and the interactions between the latent conditional Gaussian variables respectively, and $L = R^T A^{-1} R$ denotes the effect of the latent variables.

Nussbaum and Giesen (2019) assumed that there were only a few connections among the observed variables and that the number of latent variables was small. They proposed a similar sparse plus low-rank learning method by minimizing a regularized log-likelihood function. Following the proof in Chandrasekaran et al. (2012), Nussbaum and Giesen (2019) provided theoretical results on consistent estimation of the graph structure among observed variables. However, one drawback of this work is that they only provide theoretical results. No simulations or applications are presented to demonstrate the feasibility of their method.
1.1.2 Contributions

We consider the graph structure estimation problem of the sparse Ising model with latent variables in high-dimensional data. To the best of our knowledge, this is the first work that looks at the binary latent variables. As we stated before, all previous studies focused on continuous latent variables.

Our first contribution is to address the problem of graph structure estimation in the sparse Ising model given only observed variables. Compared with the Gaussian case, the existence of latent variables in the Ising model is more difficult to handle due to the specific form of the probability function. We tackle this problem by simply adding a latent component to the conditional probability function of each node, and optimizing a regularized pseudo log-likelihood function with an $\ell_1$ penalty to ensure sparsity and a nuclear norm to obtain the low-rank matrix.

Second, we develop an efficient iterative optimization algorithm to solve the resulting sparse plus low-rank learning problem. The randomized block coordinate descent algorithm with multiple acceleration techniques is proposed that can be easily implemented with algorithmic convergence guarantees.

Finally, we provide a new strategy to solve the theoretical difficulties when latent variables exist. Theoretical results show that the global error of our estimators measured by the Bregman divergence achieves satisfactory accuracy, and the graph structure can be consistently recovered with only observed data available. We also provide extensive simulation studies under several settings and two real-world applications which are able to demonstrate the advantages of our new method.

1.2 Decorrelated Lasso and Interaction Screening

In many statistical applications, it is often noticed that the linear model which only considers main effects is inadequate to capture all important information between the response and predictors. Adding some higher-order terms to the model, such as interactions, is helpful to improve the prediction ability and the model interpretability when predictors work together. In this part, we focus on a high-dimensional linear model with two-way interaction terms which is widely used in practice due to its simplicity and interpretable structure.

However, identifying the main effects and interactions in such a model is a nontrivial task, mainly due to the increased dimension of parameters in order to incorporate interactions. For a dataset with $p$ main effects, the first challenge is the number of two-way interaction terms becomes $p(p-1)/2$, considerably larger than the corresponding model with only main effects, even if $p$ is moderate. It
is well recognized that this increment in the dimension poses difficulties in both computational feasibility and theoretical analysis. The second challenge associated with interaction models is the need to specify a model such that it will satisfy so-called “strong” (or “weak”) hierarchical constraints to ensure the parsimony of a model, i.e., the interaction term is usually assumed to be nonzero unless both (or at least one of) the corresponding main effects are nonzero. Such constraints introduce extra difficulties in computational implementation.

For this part, we propose a method called DECO-IS for estimating the parameters in a linear model with two-way interaction terms. A two-stage procedure is developed that is not only computationally convenient but also able to maintain the hierarchy structure efficiently. At stage one, only the main effects are considered and selected via the decorrelated Lasso (DECO-Lasso) method. At stage two, interaction terms are selected by a similar procedure. Theoretical guarantees on the consistent estimation of the main effects are provided. Simulation studies in different settings demonstrate the efficacy of our new method.

1.3 Thesis Organization

The rest of the thesis is organized as follows.

In Chapter 2, we first provide the SIMPLE method. The details of the algorithm to solve the optimization problem and its convergence results are provided next. Theoretical guarantees on the accuracy of the global error and the consistency of the recovery of the graph structure are discussed next. Extensive simulation results are then presented, which enable us to investigate the performance of our new method in different settings. Finally, two real-world applications, one on the stock prices of oil companies in multiple markets, and the other on the exchange rates of several currencies to the US dollar, demonstrate the usefulness of the SIMPLE method.

In Chapter 3, first we provide the details of the DECO-IS method. Theoretical results on the consistent estimation of the main effects are discussed next. We then present simulation studies using synthetic data under multiple settings to demonstrate the value of our new method.

In the last Chapter, an overall conclusion as well as avenues for future work are provided. All the technical details are relegated to the Appendix.
Chapter 2

Sparse Ising Model for $P \gg n$ with Latent variables

2.1 Introduction

In this chapter, we propose a method termed SIMPLE to estimate the graph structure of the sparse Ising model with latent variables under the high-dimensional setting. In section 2.1, we provide a brief introduction of graphical models, and discuss in more detail the pairwise Markov random field, the Ising model, with related estimation methods. In section 2.2, we show the details of the SIMPLE method. In section 2.3, we present an efficient algorithm which is developed to optimize the resulting minimization problem with convergence properties studied in Section 2.4. Section 2.5 provides theoretical analysis about the global error of our estimators measured by the Bregman divergence, and the consistent estimation of the graph structure. Section 2.6 and 2.7 contain extensive simulation studies and real-world data applications which demonstrate the efficacy of our new approach.

For the convenience of the reader, here we summarize the notations to be used throughout this chapter. For a vector $\theta$, we use $\|\theta\|_1$ to denote the sum of the absolute values of the entries of $\theta$. Given a matrix $M$, $\|M\|_2$ denotes the spectral norm, which is the largest singular value of $M$; $\|M\|_F$ denotes the Frobenius norm, which is the square root of the sum of the squares of the entries of $M$; $\|M\|_*$ denotes the nuclear norm, which is the sum of the singular values of $M$. We use $\|M\|_\infty$ to denote the largest entry in magnitude of $M$, and $\|M\|_1$ to denote the sum of the absolute values of the entries of $M$. Note that $\|M\|_\infty$ and $\|M\|_1$ are not matrix norms, but rather norms on the vectorized form of the matrix. Finally, we use $\text{vec}(M)$ to denote the vectorization of $M$ by stacking its columns.
2.1.1 Graphical Models

Graphical models represent a class of very attractive models, often seen as a marriage between probability theory and graph theory (Jordan, 1998). A graphical model is a probabilistic model which uses a graph to represent the conditional dependence relationship between random variables. Such a graph, is often represented as a pair $G = (V, E)$, where $V$ is a finite set of vertices and $E$ is the set of edges. Edges $(\alpha, \beta) \in E$ with both $(\alpha, \beta)$ and $(\beta, \alpha)$ in $E$ are called an undirected graph. On the other hand, an edge $(\alpha, \beta) \in E$ with its opposite not necessary satisfying $(\beta, \alpha) \in E$ is called a directed graph.

An advantage of graphical models is that the conditional dependence relationships between random variables can be directly read by a graph, therefore they are widely employed in many areas to analyze the relationships between objects, including the handwriting recognition, the telecommunication network diagnosis, the financial information analysis and so on (Perttunen, 2006; Plötz and Fink, 2009; Bashar, 2015; Abbruzzo et al., 2016).

2.1.2 Markov Properties

The undirected graphical model, often called the Markov random field, is widely used in many research areas, for example, image analysis in computer science, social network analysis in sociology, and spatial problems in statistics (Li, 2009; Everitt, 2012; Sorbye and Rue, 2014; Jia et al., 2017). It involves a set of random variables having Markov properties that can be described by an undirected graph. Specifically, given an undirected graph $G = (V, E)$, the vertex $i \in V$ represents a random variable $X_i$ and the edge $(i, j) \in E$ represents the conditional dependency between the pair of variables $(X_i, X_j)$. Together they form a Markov random field with respect to $G$ if the following Markov properties are satisfied:

1. **Pairwise Markov property**: any two non-adjacent variables $X_i$, $X_j$ are conditionally independent given all other variables $X_{V \setminus \{i,j\}}$

   \[ X_i \perp X_j \mid X_{V \setminus \{i,j\}}. \]

2. **Local Markov property**: a variable $X_i$ is conditionally independent of all other variables $X_{V \setminus N[i]}$ given its neighbors $X_{N[i]}$

   \[ X_i \perp X_{V \setminus N[i]} \mid X_{N[i]}. \]
3. **Global Markov property**: any two subsets of variables $X_A, X_B$ are conditionally independent given a separating subset $X_S$ in the sense that

$$X_A \perp X_B \mid X_S.$$  

These Markov properties can be represented graphically. The vertices in a graph represent random variables, and the edges describe the conditional dependency between these random variables. Two variables are conditional independent if there are separating sets disconnecting them. That is, there is no path between them if the separation sets are removed.

### 2.1.3 Pairwise Markov Random Field

The pairwise Markov random field only considers pairwise interactions among variables. It is important and widely used due to its simple expression. The basic idea behind it is to model the edges as parameters. If there is an edge between two vertices $X_i$ and $X_j$, then there is a non-zero parameter $\theta_{ij}$ for this node pair $(X_i, X_j)$. In other words, the matrix of parameters encodes the graph structure.

**Definition 1.** A pairwise Markov random field is a probability distribution $P$ over random variables $X = (X_1, X_2, ..., X_p)^T$ with respect to an undirected graph $G$ where the vertices in $G$ correspond to random variables $X_i$. The probability distribution $P$ has the following form

$$P(X) = \frac{1}{Z} \exp \left\{ \sum_{(i,j) \in E} \Phi_{ij}(X_i, X_j) \right\},$$

where for each edge $(i,j) \in E$, $\Phi_{ij}$ is a mapping from pairs $(X_i, X_j) \in \mathcal{X}_i \times \mathcal{X}_j$ to the real line, and the partition function,

$$Z = \sum_X \exp \left\{ \sum_{(i,j) \in E} \Phi_{ij}(X_i, X_j) \right\},$$

also called normalization constant, makes sure the probability sums to one.

### 2.1.4 The Ising Model

Two types of undirected graphical models that focus on different types of random variables have been studied extensively in the past years: the multivariate Gaussian graphical model for continuous data, and the Ising model for binary data. In the
Gaussian graphical model, random variables are assumed to follow a multivariate Gaussian distribution, and the graph structure is fully represented by the inverse of the covariance matrix, also called the precision matrix. Therefore, the estimation of a Gaussian graphical model is equivalent to identifying the non-zero, off-diagonal elements in the precision matrix. A great deal of work has been conducted on the estimation of the Gaussian graphical model in the high-dimensional setting, such as Meinshausen et al. (2006), Yuan and Lin (2007), Friedman et al. (2008), Lam and Fan (2009), Peng et al. (2009), Zhou et al. (2011) and Cai et al. (2011). Most of these studies employed penalized likelihood functions, and provided theoretical guarantees on the consistent estimation of the precision matrix.

The Ising model, also known as the binary Markov random field, depicts the dependence structure of random variables that take binary values. Without loss of generality, we assume that these two values are 0 and 1. This model has been widely used in many applications since binary data commonly appear in many areas such as politics (Banerjee et al., 2008; Geng et al., 2017) and psychology (Reckase, 2009; Hyde et al., 2019). Consider a set of random variables \[X = (X_1, ..., X_p)^T \in \{0, 1\}^p.\] The probability function of the Ising model takes the form in the exponential family distribution as

\[
P(X; \Theta) = \frac{1}{Z(\Theta)} \exp \left\{ \sum_i \theta_i X_i + \sum_{i \neq j} \theta_{ij} X_i X_j \right\},
\]

where \(\theta_i\) is the parameter associated with the node \(i\), \(\theta_{ij}\) is the parameter associated with the edge between node \(i\) and node \(j\), and \(Z(\Theta)\) is the normalization constant, also called the partition function.

The graph structure of the Ising model can be obtained from the coefficients of the interaction term \(\theta_{ij}\) in the above probability function (2.1.1). More specifically, an edge exists between node \(i\) and \(j\) if and only if \(\theta_{ij} \neq 0\). However, compared with the Gaussian case, the estimation of the parameters in the Ising model is more challenging since the partition function involves a summation of \(2^p\) terms, and thus is not computationally tractable for large \(p\). To overcome this, the estimation of the Ising model has drawn considerable attention in recent years. Loosely speaking, there are two main approaches to estimate the graph structure of the Ising model. The first is the node-wise approach (Ravikumar et al., 2010) which estimates the whole graph structure by solving single logistic regression on each node and then combining all results. The second is the pseudo-likelihood approach (Höfling and Tibshirani, 2009) which utilizes a pseudo-likelihood function to simultaneously estimate all parameters and identify the graph structure. Both approaches minimize
some negative log-likelihood functions with an \( \ell_1 \) penalty term to induce sparsity.

### 2.1.5 Estimation Methods of the Ising Model

In the Ising model (2.1.1), the log-likelihood function of a single observation \( x = (x_1, x_2, ..., x_p)\) takes the following form

\[
\ell(x, \Theta) = \log P(x, \Theta) = \sum_{i,j=1}^{p} \theta_{ij} x_i x_j - \Psi(\Theta),
\]

where

\[
\Psi(\Theta) = \log \sum_{x} \exp \left\{ \sum_{i,j=1}^{p} \theta_{ij} x_i x_j \right\}.
\]

Since it is an undirected graph, both edges \((i,j)\) and \((j,i)\) are in \(E\). Assume that the parameters \(\theta_{ij}\) and \(\theta_{ji}\) are the same, hence the parameter matrix \(\Theta\) is a \(p \times p\) symmetric matrix. Given \(n\) observations \(X = (x^{(1)}, x^{(2)}, ..., x^{(n)})\), where \(x^{(k)} = (x^{(k)}_1, x^{(k)}_2, ..., x^{(k)}_p)^T\), the penalized log-likelihood is

\[
\sum_{k=1}^{n} \sum_{i,j=1}^{p} \theta_{ij} x^{(k)}_i x^{(k)}_j - n\Psi(\Theta) - n\|R \ast \Theta\|_1,
\]

where \(R\) is a \(p \times p\) matrix of penalty parameters and \(\ast\) denotes component-wise multiplication between matrices.

The maximum likelihood method is to maximize the above log-likelihood function

\[
\arg \max_{\Theta \in \mathbb{R}^{p \times p}} \left\{ \sum_{k=1}^{n} \sum_{i,j=1}^{p} \theta_{ij} x^{(k)}_i x^{(k)}_j - n\Psi(\Theta) - n\|R \ast \Theta\|_1 \right\},
\]

where \(\|R \ast \Theta\|_1 = \sum_{i,j} |R_{ij} \cdot \theta_{ij}|\) is the \(\ell_1\) norm of the matrix \(R \ast \Theta\), or equivalently minimize the negative log-likelihood function which is more commonly used in practice

\[
\arg \min_{\Theta \in \mathbb{R}^{p \times p}} \left\{ -\sum_{k=1}^{n} \sum_{i,j=1}^{p} \theta_{ij} x^{(k)}_i x^{(k)}_j + n\Psi(\Theta) + n\|R \ast \Theta\|_1 \right\}.
\]

However, the partition function \(\Psi(\Theta)\) poses a problem since it consists of \(2^p\) possible configurations for \(x \in \{0,1\}^p\) and thus is intractable. The methods introduced in the following aim to avoid calculating the partition function directly to estimate the structure of the Ising model.
(1) Node-wise Estimation Method

Ravikumar et al. (2010) proposed a node-wise method to estimate the graph structure of the Ising model. It is based on the following fact: since the random variable $X_i$ only takes binary value 0 and 1, given the rest of the variables $X_{-i}$, the conditional probability is given as follows

$$P(X_i = 1 \mid X_{-i}) = \frac{P(X_i = 1, X_{-i})}{P(X_i = 1, X_{-i}) + P(X_i = 0, X_{-i})}, \quad (2.1.2)$$

which has the advantage of not including the partition function $\Psi(\Theta)$.

Furthermore, consider the odds ratio of the above conditional probability, we have the following equation which simplifies model (2.1.2):

$$\frac{P(X_i = 1 \mid X_{-i})}{P(X_i = 0 \mid X_{-i})} = \frac{P(X_i = 1, X_{-i})}{P(X_i = 0, X_{-i})}. \quad (2.1.3)$$

The joint distribution of the Ising model takes the exponential form (2.1.1), then the odds ratio (2.1.3) can be written as

$$\frac{P(X_i = 1 \mid X_{-i})}{P(X_i = 0 \mid X_{-i})} = \exp \left\{ \theta_i + \sum_{i \neq j} \theta_{ij} X_j \right\},$$

and the log-odds ratio is

$$\log \frac{P(X_i = 1 \mid X_{-i})}{P(X_i = 0 \mid X_{-i})} = \theta_i + \sum_{i \neq j} \theta_{ij} X_j, \quad (2.1.4)$$

which has the same form as the logistic regression. It follows that in the Ising model, the conditional probability of $X_i$ given the rest of the variables $X_{-i}$ is:

$$P(X_i \mid X_{-i}) = \frac{\exp \left\{ X_i \left( \theta_i + \sum_{i \neq j} \theta_{ij} X_j \right) \right\}}{1 + \exp \left\{ \theta_i + \sum_{i \neq j} \theta_{ij} X_j \right\}}.$$  

Given $n$ observations $X = (x^{(1)}, ..., x^{(n)})^T$, where $x^{(k)} = (x^{(k)}_1, ..., x^{(k)}_p)^T$, the associated negative log-likelihood function takes the following form:

$$\ell(\theta) = \sum_{k=1}^n \left\{ -x^{(k)}_i \theta_i + \sum_{i \neq j} \theta_{ij} x^{(k)}_j + \log(1 + \exp \left( \theta_i + \sum_{i \neq j} \theta_{ij} x^{(k)}_j \right)) \right\}.$$  

Built on this fact, Ravikumar et al. (2010) proposed the node-wise estimation method of the Ising model. For the random variable $X_i$, condition on the remaining
variables $X_{-i}$, it solves the following separated $\ell_1$-penalized logistic regression:

$$\hat{\theta} = \arg \min_{\theta} \left\{ -\ell(\theta) + \lambda \| \theta \|_1 \right\}, \quad (2.1.5)$$

where $\theta \in \mathbb{R}^p$ is the parameter vector, the $\ell_1$ norm $\| \theta \|_1 = \sum_{j \in V \setminus i} |\theta_{ij}|$, and $\lambda$ is the penalty parameter.

In the Ising model, the estimator $\hat{\theta}_{ij} \neq 0$ in (2.1.5) indicates that there is an edge between node $i$ and $j$. Repeating this procedure for each node and combining all the results, Ravikumar et al. (2010) can obtain the whole graph structure. However, the drawback of this method is that the results may be contradictory, since parameter estimators are obtained from each node separately, they are not necessarily to be nonzero or zero at the same time. Therefore extra actions are needed to be taken on the results. The following two procedures were proposed in Ravikumar et al. (2010) to symmetrize the parameter $\hat{\Theta}$. The first one, called “OR”, is defined as

$$\theta_{ij} = \theta_{ji} = \left\{ \begin{array}{ll} \hat{\theta}_{ij}, & \text{if } |\hat{\theta}_{ij}| > |\hat{\theta}_{ji}| \\ \hat{\theta}_{ji}, & \text{if } |\hat{\theta}_{ij}| \leq |\hat{\theta}_{ji}| \end{array} \right.,$$

which indicates that there is an edge between node $i$ and node $j$ if any one of $\hat{\theta}_{ij}$ and $\hat{\theta}_{ji}$ is nonzero. The second method, called “BOTH”, is defined as

$$\theta_{ij} = \theta_{ji} = \left\{ \begin{array}{ll} \hat{\theta}_{ij}, & \text{if } |\hat{\theta}_{ij}| < |\hat{\theta}_{ji}| \\ \hat{\theta}_{ji}, & \text{if } |\hat{\theta}_{ij}| \geq |\hat{\theta}_{ji}| \end{array} \right.,$$

which indicates that there is an edge between node $i$ and node $j$ if both of $\hat{\theta}_{ij}$ and $\hat{\theta}_{ji}$ are nonzero.

(2) Pseudo-likelihood Approach

As shown before, the partition function $\Psi(\Theta)$ is complex and therefore it is computational expensive to design an algorithm to estimate the Ising model. One possible way to solve this problem is to find an approximate version of the likelihood function instead. Pseudo-likelihood function (Besag, 1975) is useful in this case.

**Definition 2.** Given a set of random variables $X = \{X_1, X_2, ..., X_p\}^T$. The pseudo-likelihood of a single observation $x = \{x_1, x_2, ..., x_p\}^T$ takes the following form

$$P(X = x) = \prod_i P(X_i = x_i | X_{-i} = x_{-i}),$$
and the pseudo log-likelihood function is

$$\log P(X = x) = \sum_i \log P(X_i = x_i | X_{-i} = x_{-i}).$$

As shown in the last part of the node-wise estimation method, the conditional probability of $X_i$ given the rest of the variables $X_{-i}$ in the Ising model is

$$P(X_i | X_{-i}) = \frac{\exp \left\{ X_i \left( \theta_i + \sum_{i \neq j} \theta_{ij} X_j \right) \right\}}{1 + \exp \left\{ \theta_i + \sum_{i \neq j} \theta_{ij} X_j \right\}}.$$

Based on this result, Höfling and Tibshirani (2009) proposed the pseudo likelihood approach to estimate the graph structure of the Ising model. This approach is similar to the node-wise estimation method, but it is able to estimate all parameters simultaneously, and the resulting estimate parameter matrix $\hat{\Theta}$ is naturally symmetric without additional operations like the ‘OR’ or ‘BOTH’ rules described before. Although the key issue of utilizing the pseudo likelihood function is the misspecification problem, Höfling and Tibshirani (2009) showed that in the Ising model, the accuracy of the estimator (measured by the false positive rate and the true positive rate) using the pseudo likelihood approach was nearly the same as using the real likelihood function, which indicated that the pseudo likelihood function approximated the real likelihood reasonably well, however using the real likelihood was computationally much more expensive.

For a single observation $x = \{x_1, x_2, ..., x_p\}^T$, the pseudo log-likelihood function of the Ising model takes the following form

$$\tilde{\ell}(\Theta | x) = \sum_{i=1}^{p} \log P(X_i = x_i | X_{-i} = x_{-i})$$

$$= \sum_{i=1}^{p} \left\{ x_i \left( \theta_i + \sum_{i \neq j} \theta_{ij} x_j \right) - \log \left( 1 + \exp \left\{ \theta_i + \sum_{i \neq j} \theta_{ij} x_j \right\} \right) \right\}.$$

To ensure the symmetry of parameter matrix $\Theta$, it assumes that $\theta_{ij} = \theta_{ji}$ for $i \neq j$. Given $n$ observations $X = (x^{(1)}, ..., x^{(n)})^T$, where $x^{(k)} = (x_1^{(k)}, ..., x_p^{(k)})^T$, the
pseudo log-likelihood function is given as follows

\[
\ell(\Theta | X) = \sum_{k=1}^{n} \tilde{\ell}(\Theta | x^{(k)}) \\
= \sum_{k=1}^{n} \sum_{i=1}^{p} \left\{ x_i^{(k)} \left( \theta_i + \sum_{i \neq j} \theta_{ij}x_j^{(k)} \right) - \log \left( 1 + \exp \left( \theta_i + \sum_{i \neq j} \theta_{ij}x_j^{(k)} \right) \right) \right\}.
\]

(2.1.6)

Höfling and Tibshirani (2009) employed above pseudo log-likelihood function (2.1.6) to estimate the graph structure of the Ising model by maximizing the following \( \ell_1 \) penalized pseudo log-likelihood function

\[
\max_{\Theta \in \mathbb{R}^{p \times p}} \left\{ \ell(\Theta; X) - n \| R * \Theta \|_1 \right\},
\]

(2.1.7)
or equivalently, minimizing the \( \ell_1 \) penalized negative pseudo log-likelihood function

\[
\min_{\Theta \in \mathbb{R}^{p \times p}} \left\{ - \ell(\Theta; X) + n \| R * \Theta \|_1 \right\},
\]

(2.1.8)

where \( R \) is the penalty parameter matrix, and \( * \) means the component-wise multiplication.

### 2.2 The SIMPLE Method

In this section, we provide details of the new SIMPLE method.

#### 2.2.1 Model Setup

As stated before, the Ising model is widely applicable. However, when handling real-world problems, a common challenge emerges when some relevant variables are unobserved or latent. Naturally, these unobserved variables may heavily destroy the structure of the whole graph. In this situation, traditional estimation methods designed for the Ising model with all variables available to the data analysts are inadequate.

Figure 2.1 illustrates such a scenario where the existence of a latent variable can confound a sparse graphical model with a dense one. In this figure, there are \( p \) nodes in total, but only \( p - 1 \) of them are observed. That is, one of them is latent. The left panel shows the true graph between the observed and unobserved nodes. The black solid lines represent the edges between these observed nodes which are represented as solid black circles. There is one latent node labeled as “h” in red,
and the dashed red lines represent the edges between the observed nodes and this latent node. The right panel shows the resulting graph between observed nodes by naively ignoring the latent one if we insist on fitting an Ising model to the observed. Comparisons between these two graphs indicate that neglecting the latent variable can induce a dense graph with “fake” interactions between the observed variables which are not connected in the true graph on the left. In Figure 2.1, these “fake” connections are represented by dashed black lines in the right panel.

Figure 2.1: The effect of the latent variable on the graph structure. In the left panel, there is one latent variable labeled as “h” in red. The black solid lines represent the edges between observed nodes, and the dashed red lines represent the edges between the observed nodes and the latent “h” node. In the right panel, the dashed black lines represent “fake” connections between observed nodes due to ignoring the latent node.

What is more, due to the specific form of the probability function of the Ising model, when the observed and latent variables are assumed jointly following the Ising model, the marginal distribution of observed variables no longer follows an Ising model. A simple demonstration is given as follows. Assume that there are in total $p$ observed and unobserved variables, denoted as $X = (X_1, X_2, ..., X_p)^T \in \{0, 1\}^p$, then their joint distribution takes the form of (2.1.1). If we assume that the $p$th variable is unobserved, the marginal distribution of the rest $p - 1$ variables, denoted
as $P(X_{-p}; \Theta)$, is seen as

$$P(X_{-p}; \Theta) = P(X_{-p}, X_p = 1; \Theta) + P(X_{-p}, X_p = 0; \Theta)$$

$$= \frac{1}{Z(\Theta)} \exp \left\{ \sum_{i=1}^{p-1} \theta_i X_i + \sum_{j \neq i}^{p-1} \theta_{ij} X_i X_j + \sum_{i=1}^{p-1} \theta_{ip} X_i + \theta_p + \sum_{k=1}^{p-1} \theta_{pk} X_k \right\}$$

$$+ \frac{1}{Z(\Theta)} \exp \left\{ \sum_{i=1}^{p-1} \theta_i X_i + \sum_{j \neq i}^{p-1} \theta_{ij} X_i X_j \right\}.$$  (2.2.1)

Consequently, the above marginal distribution of the $p-1$ observed variables (2.2.1) cannot be written in a form of the Ising model (2.1.1). Moreover, if there is more than one unobserved variable, the marginal distribution becomes even more complex. This simple example indicates that the estimation methods that we introduced before for the Ising model not accounting for latent variables are unreliable when some variables are unobserved.

However, learning the graph structure of the Ising model with the presence of latent variables is a challenging task, especially for high-dimensional data. First, the latent variables are hard to deal with due to identifiability and tractability issues. Second, estimating the graph structure of the Ising model is itself difficult because of the complex form of the partition function.

In this study, we propose a new method called Sparse Ising Model for $P \gg n$ with Latent variablEs, abbreviated as SIMPLE, to overcome the challenges outlined above. We show that SIMPLE is able to incorporate latent variables and correctly estimate the graph structure with high probability. In deriving SIMPLE, we use the fact that in the Ising model, the conditional distribution of each variable $X_i$ given the rest of the variables denoted as $X_{-i}$ takes the form of logistic regression:

$$P(X_i \mid X_{-i}; \Theta) = \frac{\exp \left\{ X_i \left( \theta_i + \sum_{j \neq i} \theta_{ij} X_j \right) \right\}}{1 + \exp \left\{ \theta_i + \sum_{j \neq i} \theta_{ij} X_j \right\}}.$$  (2.2.2)

In such a single logistic regression model, the variable $X_i$ is the response, the rest of the variables $X_{-i}$ play the role of covariates, and the non-zero coefficients encode conditional dependence relationships between the variable $X_i$ and others (i.e., the edges between node $i$ and other nodes). Therefore, the estimation of the graph structure in the Ising model is to identify all the non-zero coefficients in the logistic regression for all the nodes. This is the key motivation for the node-wise
approach (Ravikumar et al., 2010) and the pseudo-likelihood approach (Höfling and Tibshirani, 2009) which are the most popular methods to estimate the Ising model.

In our case, the observed and latent variables are assumed to jointly follow the Ising model. When one observed variable plays the role of the response, the covariates in the logistic regression associated with that variable should include all of the rest observed variables and latent variables. However, because no information is available on the latent variables, the presence of these latent variables poses a challenge. Directly using the methods designed for the Ising model that do not model latent variables will produce inaccurate and unreliable results, especially when the latent variables play an important role in the graph.

From the discussion above, our SIMPLE method includes an additional explanatory component in the single logistic regression to describe the dependence relationship between the observed variable $X_i$ and the latent variables. Denoting these latent variables as $Z$ for the moment, if we observed $Z$, we would have

$$P(X_s | X_{-s}, Z; \Theta) = \frac{\exp \left\{ X_s (\theta_s + \sum_{t \neq s} X_t \theta_{st} + \sum_{j \in H} Z_j \gamma_{sj}) \right\}}{1 + \exp \left\{ \theta_s + \sum_{t \neq s} X_t \theta_{st} + \sum_{j \in H} Z_j \gamma_{sj} \right\}},$$

(2.2.3)

where $\theta_s$ is the parameter associated with the variable $X_s$, $\theta_{st}$ is the parameter associated with the interaction of the observed variables $X_s$ and $X_t$. Assume that there are $k$ latent variables, $\sum_{j \in H} Z_j \gamma_{sj}$ for $s = 1, 2, ..., p$, $j = 1, 2, ..., k$ is a linear combination of $k$ latent variables.

The log-likelihood function of the conditional distribution (2.2.3) takes the following form

$$\log P(X_s, \Theta | X_{-s}, Z) = X_s \left( \theta_s + \sum_{t \neq s} X_t \theta_{st} + \sum_{j \in H} Z_j \gamma_{sj} \right) - \Psi_s(\Theta),$$

where

$$\Psi_s(\Theta) = \log \left\{ 1 + \exp \left( \theta_s + \sum_{t \neq s} X_t \theta_{st} + \sum_{j \in H} Z_j \gamma_{sj} \right) \right\}.$$

As defined before, the pseudo log-likelihood of a single observation $x =$
\[ \{x_1, x_2, \ldots, x_p \}^T \text{ and } z = \{z_1, z_2, \ldots, z_k \}^T \text{ has the following form} \]

\[
\tilde{\ell}(\Theta | x, z) = \sum_{s=1}^{p} \log P(x_s, \Theta | x_{-s}, z) \\
= \sum_{s=1}^{p} \left\{ x_s \left( \theta_s + \sum_{t \neq s} x_t \theta_{st} + \sum_{j \in \mathcal{H}} z_j \gamma_{sj} \right) \\
- \log \left( 1 + \exp \left( \theta_s + \sum_{t \neq s} x_t \theta_{st} + \sum_{j \in \mathcal{H}} z_j \gamma_{sj} \right) \right) \right\}. \tag{2.2.4}
\]

To ensure the symmetry of the parameter matrix \( \Theta \), we assume \( \theta_{st} = \theta_{ts} \) for \( s \neq t \), then the pseudo log-likelihood function for a single observation (2.2.4) can be written in a simpler way

\[
\tilde{\ell}(\beta, \Theta | x, z) = \langle x, x \rangle + \Theta^T X + \Gamma^T z, \quad - \langle 1, b(\beta + x^T \beta + \Gamma^T z) \rangle,
\]

where the parameter vector \( \beta = (\theta_1, \theta_2, \ldots, \theta_p)^T \), the parameter matrix \( \Theta = \Theta^T \in \mathbb{R}^{p \times p} \) and \( \Theta_{jj} = 0 \) for \( j = 1, 2, \ldots, p \), and \( 1_p \) is a \( p \) dimensional vector of one. \( \langle a, b \rangle \) denotes the inner product, and function \( b(a) = \log(1 + \exp(a)) \).

Given \( n \) observations \( X = (x^{(1)}, \ldots, x^{(n)})^T \in \mathbb{R}^{n \times p} \) and \( Z = (z^{(1)}, \ldots, z^{(n)})^T \in \mathbb{R}^{n \times k} \), where \( x^{(l)} = \{x_1^{(l)}, \ldots, x_p^{(l)}\}^T \) and \( z^{(l)} = \{z_1^{(l)}, \ldots, z_k^{(l)}\}^T \), we have the following pseudo log-likelihood function

\[
f(\beta, \Theta | X, Z) = \sum_{l=1}^{n} \tilde{\ell}(\beta, \Theta | x^{(l)}, z^{(l)}) \\
= \langle X, 1_n \beta^T + X \Theta + Z \Gamma \rangle - \langle 1_n \beta^T + X \Theta + Z \Gamma \rangle, \tag{2.2.5}
\]

where the parameter vector \( \beta \) and matrix \( \Theta \) are the same as defined above, \( \Gamma \in \mathbb{R}^{k \times p} \) is the parameter matrix for \( k \) latent variables, and \( 1_n \) is an \( n \times p \) matrix of one. \( \langle A, B \rangle = tr(A^T B) \) is the matrix inner product, and \( b(A) = \log(1 + \exp(A)) \) is an element-wise operator of \( b(\cdot) \) for the matrix \( A \).

### 2.2.2 Estimation

In this section, we address the estimation issue of the SIMPLE method described in the last section. First, since the variables \( Z \) is unobserved, the pseudo log-likelihood function (2.2.5) needs to be reparametrized via \( L = \mathcal{Z} \Gamma \), that is

\[
f(\beta, \Theta, L) = \langle X, 1_n \beta^T + X \Theta + L \rangle - \langle 1_n \beta^T + X \Theta + L \rangle. \tag{2.2.6}
\]
where $\beta \in \mathbb{R}^p$, $\Theta \in \mathbb{R}^{p \times p}$ are parameters for the observed variables, and $L$ is the matrix which shows the effect of the latent variables. Note that since $Z$ are not observed, the best we know about $L$ is that its rank is $k$. Of course, $k$ is unknown in practice.

An advantage of utilizing the pseudo log-likelihood function (2.2.6) is to avoid the computationally demanding task of calculating the partition function. To estimate the graph structure of the Ising model with latent variables, i.e., the nonzero elements in parameter matrix $\Theta$, based on the pseudo likelihood function (2.2.6), there are two assumptions imposed on the parameters $\Theta$ and $L$: first, we make the common and important sparsity assumption for graphical model estimation under a high-dimensional setting (Meinshausen et al., 2006; Yuan and Lin, 2007; Friedman et al., 2008; Ravikumar et al., 2010). In our case, this assumption means that the parameter matrix $\Theta$ for the observed variables is sparse. Second, we assume that the number of latent variables is small compared with the observed variables.

The first sparsity assumption can be achieved by a commonly used $\ell_1$ norm. The second assumption which indicates the matrix $L$ is of low-rank, i.e., $\text{rank}(L) \ll \min\{n, p\}$, can be achieved by the nuclear norm, since it is suitable to measure the “size” of a matrix. The nuclear norm is a convex relaxation of a rank constraint, which is analogous to the $\ell_1$ norm as a relaxation of the cardinality of a vector.

After the discussion above, we are ready to present the SIMPLE method which estimates the graph structure of the Ising model with latent variables by optimizing the following regularized pseudo log-likelihood function

$$
(\hat{\beta}, \hat{\Theta}, \hat{L}) = \arg \min_{\beta, \Theta, L} \left\{ -f(\beta, \Theta, L) + \lambda_1 \|\Theta\|_1 + \lambda_2 \|L\|_* \right\} 
$$

$$
= \arg \min_{\beta, \Theta, L} \left\{ -\langle X, 1_n \beta^T + X \Theta + L \rangle + \langle 1_{n \times p}, b(1_n \beta^T + X \Theta + L) \rangle 
+ \lambda_1 \|\Theta\|_1 + \lambda_2 \|L\|_* \right\},
$$

(2.2.7)

where the $\ell_1$ penalty $\|\Theta\|_1 = \sum_{i \neq j} |\theta_{ij}|$ ensures the sparsity and the nuclear norm $\|L\|_* = \sum_i \sigma_i$, where $\sigma_i$ is the $i$th largest singular value of $L$, ensures the matrix $L$ is of low-rank. It is easily seen that (2.2.7) is a convex optimization problem.

We develop a block coordinate descent algorithm to solve this optimization problem with details given in Section 2.3.

We provide further explanations about these two penalty terms $\|\Theta\|_1$ and $\|L\|_*$ below. For the first term, $\Theta$ is a $p \times p$ symmetric matrix with zero diagonal...
elements, i.e., $\theta_{ij} = \theta_{ji}$ for $i \neq j$ and $\theta_{ii} = 0$. Then the $\ell_1$ penalty is

$$\| \Theta \|_1 = \sum_{i \neq j} |\theta_{ij}|,$$

which penalizes the number of nonzero parameter $\theta_{ij}$’s. Since the nonzero parameter $\theta_{ij}$ indicates an edge between node $i$ and $j$, increasing the regularization parameter $\lambda_1$ will produce less numbers of nonzero $\theta_{ij}$, thus there are less edges in total in the estimated graph, i.e., the graph is more sparse. The $\ell_1$ penalty was originally proposed by Tibshirani (1996) in LASSO model, and later was introduced to graphical models (Yuan and Lin, 2007; Friedman et al., 2008; Ravikumar et al., 2011).

The second penalty term $\| L \|_*$ is the nuclear norm. Note that the matrix $L \in \mathbb{R}^{n \times p}$ has the following singular value decomposition (SVD)

$$L = U \Sigma V^T,$$

where $U$ is an $n \times n$ matrix, $V$ is a $p \times p$ matrix, and $\Sigma$ is an $n \times p$ rectangular diagonal matrix where the diagonal entries $\sigma_i$ is the $i$th largest singular value of $L$. Then the nuclear norm is defined as the sum of singular values

$$\| L \|_* = \sum_{i=1}^I \sigma_i, \text{ where } I = \min\{n, p\}.$$

The nuclear norm is suitable to measure the “size” of a matrix, since it penalizes the number of nonzero singular values of $L$, which is the same as the rank of $L$. This nuclear norm was first proposed in Fazel (2003), and has wide applications in statistics (Koltchinskii et al., 2011; Xue et al., 2012; Zhou and Li, 2014).

### 2.3 Algorithm

In this section, we provide details of the randomized block coordinate descent algorithm developed to solve the regularized pseudo log-likelihood function minimization problem (2.2.7). This optimization problem has three parameters. We employ an iterative algorithm in which at each iteration, only one parameter is optimized while the other two parameters are fixed.
2.3.1 Randomized Block Coordinate Descent

To introduce the algorithm, we consider the optimization problem

$$
\min_{x \in \mathcal{X}} F(x) \triangleq f(x) + \sum_{i=1}^{J} r_i(x_i),
$$

where, with some abuse of notation, we use $x$ to denote the argument that we want to optimize. Here variable $x$ can be decomposed into $J$ blocks $x_1, \ldots, x_J$, $\mathcal{X}$ is the set of feasible points, $f$ is convex and differentiable on $\mathcal{X}$, having a block coordinate Lipschitz continuous gradient, and $r_i$ is a closed convex real valued function and we define $r_i(x_i) = \infty$ if $x_i \notin \text{dom}(r_i)$, $i = 1, 2, \ldots, J$.

The block coordinate descent (BCD) method is to minimize $F(x)$ over only one set $x_i$ at one time while fixing the remaining blocks at their last updated values. Let $x_i^{[k]}$ denote the value of $x_i$ at the $k$th update, the update scheme is

$$
x_i^{[k]} = \arg \min_{x_i \in \mathcal{X}} f(x_i; x_{-i}^{[k-1]}) + r_i(x_i),
$$

where $f(x_i; x_{-i}^{[k-1]})$ is the function of $x_i$ which keeps the rest groups at their last updated values.

The efficiency of a BCD method depends on the choice of which block is to be updated in the next iteration. The simplest strategy is a cyclic coordinate update, i.e., coordinates are updated one by one. Another possibility is to move along the direction with the largest descent. However, it is computationally expensive since it requires to compute all partial derivatives. Another approach is the randomized block coordinate descent (RBCD) which can avoid the worst-case scenario on the number of iterations and is highly efficient with less memory demand (Nesterov, 2012; Richtárik and Takáč, 2014).

The implementation of a RBCD is straight-forward. Specifically, the index $i$ of the group of variables $x_i$ to be updated at the next iteration is selected with probability

$$
P_{\alpha}^{(i)} = \frac{M_i^\alpha}{\sum_{j=1}^J M_j^\alpha},
$$

where $\alpha \in \mathbb{R}$, $M_i$ is the Lipschitz constant of $\nabla f(x_i)$ and $\nabla f(x_i)$ is the partial derivative of $f(x)$ with respect to the $i$th group $x_i$. In practice, if the Lipschitz constant $M_i$ is easy to calculate, the value of $\alpha$ can be set as 1, which means the probability of selecting the $i$th group is proportional to $M_i$. Otherwise, the value of $\alpha$ can be set as 0, which indicates the index is sampled from a uniform distribution.
Recall our optimization problem

$$\min_{\beta, \Theta, L} f(\beta, \Theta, L) + \lambda_1 \|\Theta\|_1 + \lambda_2 \|L\|_\star,$$

the RBCD algorithm solves it by randomly optimizing the following three minimization problems:

- For $\Theta$, optimize the following function (2.3.3) of $\Theta$ with fixed $\beta$ and $L$:
  $$\arg\min_{\Theta} f(\beta, \Theta, L) + \lambda_1 \|\Theta\|_1. \quad (2.3.3)$$

- For $L$, optimize the following function (2.3.4) of $L$ with fixed $\beta$ and $\Theta$:
  $$\arg\min_{L} f(\beta, \Theta, L) + \lambda_2 \|L\|_\star. \quad (2.3.4)$$

- For $\beta$, optimize the following function (2.3.5) of $\beta$ with fixed $\Theta$ and $L$:
  $$\arg\min_{\beta} f(\beta, \Theta, L). \quad (2.3.5)$$

2.3.2 Majorization-Minimization Principle

For each block, the RBCD algorithm needs to solve the following minimization problem:

$$x_i^{[k]} = \arg\min_{x_i \in \mathcal{X}} f(x_i; x_i^{[k-1]}_{-i}) + r_i(x_i), \quad (2.3.6)$$

where $x_i^{[k]}$ is the $k$th update of $x_i$, and $f(x_i; x_i^{[k-1]}_{-i})$ is the function of $x_i$ which keeps the rest groups at their last updated values.

Directly calculating the minimizer is usually difficult and time consuming. Therefore we want to design a good approximation function to replace the target function $f(x_i)$. The majorization-minimization (MM) principle can be employed to achieve this goal by making the minimization problem easier to solve. To proceed, we construct a surrogate function $g(x_i, x_i^{[k-1]})$ to replace $f(x_i)$, which satisfies

$$g(x_i, x_i^{[k-1]}) \geq f(x_i) \text{ and } g(x_i, x_i) = f(x_i), \quad (2.3.7)$$

where $x_i^{[k-1]}$ is the $k-1$th updated value of block $x_i$. Now the following updating scheme (2.3.8) which replaces the function $f(x_i; x_i^{[k-1]}_{-i})$ by the surrogate function $g(x_i, x_i^{[k-1]})$

$$x_i^{[k]} = \arg\min_{x_i \in \mathcal{X}} g(x_i, x_i^{[k-1]}), \quad (2.3.8)$$
leads to a convergent algorithm (in terms of functional value), that is

\[ f(x_i^{[k]}) \leq g(x_i^{[k]}, x_i^{[k-1]}) \leq g(x_i^{[k-1]}, x_i^{[k-1]}) = f(x_i^{[k-1]}), \]

This means the process of minimizing the surrogate function \( g \) will guarantee to minimize the original target function \( f \), since we have \( f(x_i^{[k]}) \leq f(x_i^{[k-1]}) \) by the construction. A useful surrogate function of \( f(x_i; x_{-i}^{[k-1]}) \) is linearizing it by

\[ g(x_i, x_i^{[k-1]}) = f(x_i^{[k-1]}) + \langle \nabla f(x_i^{[k-1]}), x_i - x_i^{[k-1]} \rangle + \frac{M_i}{2} \| x_i - x_i^{[k-1]} \|_2^2, \]

where \( \| \cdot \|_2 \) is the \( \ell_2 \) norm, \( M_i \) is the Lipschitz constant of \( \nabla f(x_i) \) and \( \nabla f(x_i) \) is the partial derivative of \( f(x) \) with respect to the \( i \)th group.

It is easy to see that \( g(x_i^{[k-1]}, x_i^{[k-1]]) = f(x_i^{[k-1]}) \). To ensure the first inequality \( f(x_i) \leq g(x_i, x_i^{[k-1]}) \) in (2.3.7), it suffices to have

\[ f(x_i) - f(x_i^{[k-1]}) - \langle \nabla f(x_i^{[k-1]}), x_i - x_i^{[k-1]} \rangle \leq \frac{M_i}{2} \| x_i - x_i^{[k-1]} \|_2^2. \quad (2.3.9) \]

The above inequality (2.3.9) can be obtained from the following lemma 1, and the proof is given in Appendix A.1.

**Lemma 1.** If the function \( f \) is convex and the gradient \( \nabla f \) is Lipschitz continuous, i.e.,

\[ \| \nabla f(x) - \nabla f(y) \|_2 \leq M \| x - y \|_2, \]

then for any \( x, y \in \text{dom } f \), we have

\[ f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{M}{2} \| y - x \|_2^2. \]

where \( M \) is the Lipschitz constant of the gradient function \( \nabla f \).

It then follows that the following function

\[ g(x_i, x_i^{[k-1]}) + r_i(x_i) = f(x_i^{[k-1]}) + \langle \nabla f(x_i^{[k-1]}), x_i - x_i^{[k-1]} \rangle + \frac{M_i}{2} \| x_i - x_i^{[k-1]} \|_2^2 + r_i(x_i) \]

is similarly a surrogate of the target function in (2.3.6).

Now the updating scheme using the MM principle becomes

\[ x_i^{[k]} = \arg \min_{x_i \in \mathcal{X}} f(x_i^{[k-1]}) + \langle \nabla f(x_i^{[k-1]}), x_i - x_i^{[k-1]} \rangle + \frac{M_i}{2} \| x_i - x_i^{[k-1]} \|_2^2 + r_i(x_i) \]

\[ = \arg \min_{x_i \in \mathcal{X}} \frac{M_i}{2} \| x_i - (x_i^{[k-1]} - \frac{1}{M_i} \nabla f(x_i^{[k-1]})) \|_2^2 + r_i(x_i). \]

(2.3.10)
Define a proximal mapping

\[ \text{prox}_t(x) = \arg \min_z \frac{1}{2t} \|x - z\|_2^2 + h(z), \]

then the \( k \)th update of block \( x_i \) in (2.3.10) can be written as

\[ x_i^{[k]} = \text{prox}_{\frac{1}{m_i}} \left( x_i^{[k-1]} - \frac{1}{M_i} \nabla f(x_i^{[k-1]}) \right), \]

where the proximal operator can be computed for many important functions, such as the \( \ell_1 \) norm and the nuclear norm. The following Lemmas 2 and 3 show the analytical forms of the proximal operator associated with the \( \ell_1 \) penalty and the nuclear norm respectively. The proofs are given in Appendix A.2 and A.3.

**Lemma 2.** The proximal mapping of the \( \ell_1 \) norm

\[ \text{prox}_t(x) = \arg \min_z \frac{1}{2t} \|x - z\|_2^2 + \lambda_1 \|z\|_1 \quad (2.3.11) \]

\[ = T_{\lambda_1 t}(x), \]

where \( x, z \in \mathbb{R}^p \), and \( T_{\lambda_1 t}(x) \) is the soft-thresholding operator. That is

\[ [T_{\lambda_1 t}(x)]_i = \begin{cases} 
  x_i - \lambda_1 t, & \text{if } x_i > \lambda_1 t, \\
  0, & \text{if } -\lambda_1 t \leq x_i \leq \lambda_1, \\
  x_i + \lambda_1 t, & \text{if } x_i < -\lambda_1 t.
\end{cases} \]

**Lemma 3.** The proximal mapping of the nuclear norm

\[ \text{prox}_t(X) = \arg \min_Z \frac{1}{2t} \|X - Z\|_F^2 + \lambda_2 \|Z\|_* \quad (2.3.12) \]

\[ = D_{\lambda_2 t}(X), \]

where \( D_{\lambda_2 t}(X) \) is the singular value soft-thresholding operator. It is defined as follows: consider a matrix \( X \in \mathbb{R}^{n \times p} \) of rank \( r \), the singular value decomposition (SVD) of \( X \) takes the form of (2.2.8). Then the singular value soft-thresholding operator \( D_{\lambda_2 t}(X) \) is

\[ D_{\lambda_2 t}(X) := U \tilde{D} V^T, \quad \tilde{D} = \text{diag} \left( \left\{ (\sigma_i - \lambda_2 t)_+ \right\}_{1 \leq i \leq r} \right), \]

where the function \( (\sigma_i - \lambda_2 t)_+ = \max(\sigma_i - \lambda_2 t, 0) \).

In our problem, the RBCD algorithm with the majorization-minimization principle solves our minimization problem via the following three sub-problems:
• For Θ, the kth update with MM principle is

\[
\Theta[k] = \arg \min_{\Theta} \langle \nabla_{\Theta} f(\Theta[k-1]), \Theta - \Theta[k-1] \rangle + \frac{M_{\Theta}}{2} \| \Theta - \Theta[k-1] \|^2 + \lambda_1 \| \Theta \|_1
\]

\[
= \arg \min_{\Theta} \frac{M_{\Theta}}{2} \left\| \Theta - \left( \Theta[k-1] - \frac{1}{M_{\Theta}} \nabla_{\Theta} f(\Theta[k-1]) \right) \right\|^2_F + \lambda_1 \| \Theta \|_1
\]

\[
= T_{\lambda_1/M_{\Theta}} \left( \Theta[k-1] - \frac{1}{M_{\Theta}} \nabla_{\Theta} f(\Theta[k-1]) \right),
\]

where \( \nabla_{\Theta} f \) is the partial derivative of function \( f \) with respect to \( \Theta \), and \( M_{\Theta} \) is the Lipschitz constant of \( \nabla_{\Theta} f \).

• For L, the kth update with MM principle is

\[
L[k] = \arg \min_{L} \langle \nabla_{L} f(L[k-1]), L - L[k-1] \rangle + \frac{M_{L}}{2} \| L - L[k-1] \|^2 + \lambda_2 \| L \|_*
\]

\[
= \arg \min_{L} \frac{M_{L}}{2} \left\| L - \left( L[k-1] - \frac{1}{M_{L}} \nabla_{L} f(L[k-1]) \right) \right\|^2_F + \lambda_2 \| L \|_*
\]

\[
= D_{\lambda_2/M_{L}} \left( L[k-1] - \frac{1}{M_{L}} \nabla_{L} f(L[k-1]) \right),
\]

where \( \nabla_{L} f \) is the partial derivative of function \( f \) with respect to \( L \), and \( M_{L} \) is the Lipschitz constant of \( \nabla_{L} f \).

• For β, the kth update with MM principle is

\[
\beta[k] = \arg \min_{\beta} \langle \nabla_{\beta} f(\beta[k-1]), \beta - \beta[k-1] \rangle + \frac{M_{\beta}}{2} \| \beta - \beta[k-1] \|^2
\]

\[
= \arg \min_{\beta} \frac{M_{\beta}}{2} \left\| \beta - \left( \beta[k-1] - \frac{1}{M_{\beta}} \nabla_{\beta} f(\beta[k-1]) \right) \right\|^2_2
\]

\[
= \beta[k-1] - \frac{1}{M_{\beta}} \nabla_{\beta} f(\beta[k-1]),
\]

where \( \nabla_{\beta} f \) is the partial derivative of function \( f \) with respect to \( \beta \), and \( M_{\beta} \) is the Lipschitz constant of \( \nabla_{\beta} f \).

### 2.3.3 Acceleration Method

In the last section, the majorization-minimization principle replaces the target function \( f(x_i) \) by its quadratic approximation \( g(x_i, x_i[k-1]) \). This replacement enables us to solve the minimization problem in each block via the proximal gradient descent method. Moreover, in our case, the proximal mappings of the \( \ell_1 \) norm and nuclear
norm have analytical forms, i.e., the soft-thresholding operator and the singular value soft-thresholding operator respectively.

The advantage of proximal gradient descent is in its simplicity. However, for the regularized problem with the $\ell_1$ and nuclear penalty in our problem, proximal gradient descent is known to be slow with a convergence rate of $O(1/k)$, where $k$ is the number of iterations. Therefore, acceleration techniques (Tseng, 2008; Beck and Teboulle, 2009; Nesterov, 2013; Xu and Yin, 2013) have been designed to achieve the improved rate $O(1/k^2)$ while keeping its simple form.

In this study, we employ the FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) type acceleration method. Recall that the $k$th update of block $x_i$ using the proximal gradient descent formula in the last section is

$$x_i^{[k]} = \text{prox} \frac{1}{\lambda_i} \left( x_i^{[k-1]} - \frac{1}{M_i} \nabla f(x_i^{[k-1]}) \right).$$

The FISTA type acceleration method replaces $x_i^{[k-1]}$ by an extrapolated point $z_i^{[k]}$ which is a specific linear combination of previous points $(x_i^{[k-1]}, x_i^{[k-2]})$. That is,

$$z_i^{[k]} = x_i^{[k-1]} + \frac{t_i^{[k-1]}}{t_i^{[k]}} - 1 (x_i^{[k-1]} - x_i^{[k-2]}), \quad (2.3.13)$$

$$x_i^{[k]} = \text{prox} \frac{1}{\lambda_i} \left( z_i^{[k]} - \frac{1}{M_i} \nabla f(z_i^{[k]}) \right),$$

where $t_i^{[k]} = 1 + \frac{\sqrt{1 + 4(t_i^{[k-1]} - 1)^2}}{2}$. The second term $\frac{t_i^{[k-1]}}{t_i^{[k]}} - 1 (x_i^{[k-1]} - x_i^{[k-2]})$ in (2.3.13) can be seen as a momentum of the sequence $\{x_i^{[k]}\}_{k=1}^\infty$, and the term $\frac{t_i^{[k-1]}}{t_i^{[k]}} - 1$ is the extrapolation weight. It enlarges the moving distance of the sequences which leads them closer to the optimum more quickly.

In our case, combined with above acceleration method, the update scheme for each block becomes the following.

- For $\Theta$, the $k$th update with MM and acceleration is

$$z_{\Theta}^{[k]} = \Theta^{[k-1]} + \frac{t_{\Theta}^{[k-1]}}{t_{\Theta}^{[k]}} - 1 (\Theta^{[k-1]} - \Theta^{[k-2]}),$$

$$\Theta^{[k]} = T_{\lambda_1/M_\Theta} \left( z_{\Theta}^{[k-1]} - \frac{1}{M_\Theta} \nabla f(z_{\Theta}^{[k-1]}) \right).$$
• For $L$, the $k$th update with MM and acceleration is

$$z_L^{[k]} = L^{[k-1]} - \frac{1}{t_L^{[k-1]}}(L^{[k-1]} - L^{[k-2]}),$$

$$L^{[k]} = D_{\lambda_2/M_L}\left(z_L^{[k-1]} - \frac{1}{M_L}\nabla_L f(z_L^{[k-1]})\right).$$

• For $\beta$, the $k$th update with MM and acceleration is

$$z_\beta^{[k]} = \beta^{[k-1]} - \frac{1}{t_\beta^{[k-1]}}(\beta^{[k-1]} - \beta^{[k-2]}),$$

$$\beta^{[k]} = z_\beta^{[k-1]} - \frac{1}{M_\beta}\nabla_\beta f(z_\beta^{[k-1]}),$$

where $M_\Theta, M_L, M_\beta, \nabla_\Theta f, \nabla_L f$ and $\nabla_\beta f$ are the same as defined in the last section.

### 2.3.4 Step Size

At the $k$th iteration of the block $x_i$, the accelerated proximal gradient descent is given by

$$x_i^{[k]} = \text{prox}_{\frac{1}{M_i}}\left(z_i^{[k]} - \frac{1}{M_i}\nabla f(z_i^{[k]})\right).$$

The quantity of the step size $\frac{1}{M_i}$ plays an important role since it controls how substantial the reduction of the objective function is. In previous sections, we use the Lipschitz constant $M_i$ for the simplicity of analysis. It is safe and convenient to use if the Lipschitz constant is easy to calculate in some cases. However, the drawback is that it is usually too conservative, and does not make enough decrease of the function value. To illustrate this, we present the following Lemma 4 which is key to the acceleration method. The proof is given in Appendix A.4.

**Lemma 4.** Let $f(x)$ and $g(x)$ be two convex functions defined on the convex set $X$, and $f(x)$ is differentiable. Let $F(x) = f(x) + g(x)$, $\mu > 0$ and

$$x^* = \arg\min_{x \in X} (\nabla f(y), x - y) + \frac{\mu}{2}\|x - y\|^2 + g(x).$$

If

$$F(x^*) \triangleq f(x^*) + g(x^*) \leq Q(x^*, y) \triangleq f(y) + (x^* - y, \nabla f(y)) + \frac{\mu}{2}\|x^* - y\|^2 + g(x^*),$$

(2.3.14)
then we have

$$F(x) - F(x^*) \geq \frac{\mu}{2} \|x^* - y\|^2 + \mu \langle y - x, x^* - y \rangle, \quad \forall \ x \in \mathcal{X}. $$

Here the key condition (2.3.14) which can be written in the following simplified form

$$f(x^*) \leq f(y) + \langle x^* - y, \nabla f(y) \rangle + \frac{\mu}{2} \|x^* - y\|^2, \quad (2.3.15)$$

is always satisfied if $\mu = M_f$, where $M_f$ is the Lipschitz constant of the function $f$. However, the universal choice of $M_f$ is usually not good enough, since it might be too large to be an efficient step size. The suitable quantity of $\mu$ can be evaluated at each iteration via so called line search methods, which ensures the condition (2.3.15) is satisfied.

Beck and Teboulle (2009) provided a back-tracking line search method of the FISTA, which starts from a relative large estimate of the step size and iteratively shrinks it until a desired decrease of the objective function is observed. The following Algorithm 1 shows the detail.

**Algorithm 1** FISTA with nonincreasing backtracking line search

Choose $\beta > 1$, initialize $M^{[0]}, t^{[0]}, x^{[0]} = x^{[-1]}$

for $k=1,2,...$ till converge do

1. $M^{[k]} \leftarrow M^{[k-1]}$
2. $t^{[k]} \leftarrow \frac{1 + \sqrt{1 + 4(t^{[k-1]})^2}}{2}$
3. $y^{[k]} \leftarrow x^{[k-1]} + \frac{t^{[k-1]} - 1}{t^{[k-1]}}(x^{[k-1]} - x^{[k-2]})$
4. $x^{[k]} \leftarrow \text{prox}_{\frac{1}{M^{[k]}}} (y^{[k]} - \frac{1}{M^{[k]}} \nabla f(y^{[k]}))$

while $f(x^{[k]}) > f(y^{[k]}) + \nabla f(y^{[k]})^T(x^{[k]} - y^{[k]}) + \frac{M^{[k]}}{2} \|x^{[k]} - y^{[k]}\|^2$ do

1. $M^{[k]} \leftarrow \beta M^{[k]}$
2. $x^{[k]} \leftarrow \text{prox}_{\frac{1}{M^{[k]}}} (y^{[k]} - \frac{1}{M^{[k]}} \nabla f(y^{[k]}))$

end while

end for

However, the main drawback of Algorithm 1 is that the step size at each iteration is non-increasing (i.e., the estimate of the Lipschitz constant is non-decreasing). This can substantially limit the performance of the algorithm, since if a small step size is encountered too early, all the subsequent step sizes will be very small.

To overcome this problem, Scheinberg et al. (2014) proposed a full backtracking strategy which allows the step size to increase (i.e., the estimate of the Lipschitz constant to decrease) from one iteration to the next. The following Algorithm 2
Algorithm 2 FISTA with full backtracking line search

Choose $\eta_u > 1$, $\eta_d > 1$, initialize $M^{[0]}$, $t^{[0]}$, $x^{[0]} = x^{-1}$

for $k=1,2,...$ till converge do

\[ M^{[k]} \leftarrow M^{[k-1]} / \eta_d \]

\[ t^{[k]} \leftarrow 1 + \sqrt{1 + 4(M^{[k]} / M^{[k-1]})(t^{[k-1]})^2} \]

\[ y^{[k]} \leftarrow x^{[k-1]} + \frac{t^{[k-1]-1}}{t^{[k]}} (x^{[k-1]} - x^{[k-2]}) \]

\[ x^{[k]} \leftarrow \text{prox}_{\frac{1}{M^{[k]}}} (y^{[k]} - \frac{1}{M^{[k]}} \nabla f(y^{[k]})) \]

while $f(x^{[k]}) > f(y^{[k]}) + \nabla f(y^{[k]})^T (x^{[k]} - y^{[k]}) + \frac{M^{[k]}}{2} \| x^{[k]} - y^{[k]} \|^2$ do

\[ M^{[k]} \leftarrow \eta_u M^{[k]} \]

\[ t^{[k]} \leftarrow 1 + \sqrt{1 + 4(M^{[k]} / M^{[k-1]})(t^{[k-1]})^2} \]

\[ y^{[k]} \leftarrow x^{[k-1]} + \frac{t^{[k-1]-1}}{t^{[k]}} (x^{[k-1]} - x^{[k-2]}) \]

\[ x^{[k]} \leftarrow \text{prox}_{\frac{1}{M^{[k]}}} (y^{[k]} - \frac{1}{M^{[k]}} \nabla f(y^{[k]})) \]

end while

end for

2.3.5 Restart

One key characteristic of accelerated gradient methods is the non-monotonicity. That is, the function values may increase for a couple of iterates, before decreasing again. This undesirable property may hurt the performance of fast gradient algorithms (Beck and Teboulle, 2009; O’donoghue and Candes, 2015).

In addition, the amount of momentum $\frac{t^{[k-1]-1}}{t^{[k]}} (x^{[k-1]} - x^{[k-2]})$ decides the behavior of acceleration algorithm as well. When the value of the extrapolation weight $\frac{t^{[k-1]-1}}{t^{[k]}}$ is close to 1, i.e., the momentum is high, the generated sequences can overshoot and oscillate around the optimal solution. In order to solve these non-monotonicity and overshooting problems, Giselsson and Boyd (2014) proposed an adaptive restart strategy for fast gradient methods which improves the convergence performance of the algorithm. The restart strategy is that when a certain restart condition holds, the extrapolated point $z^{[k]}_i$ is set to $x^{[k-1]}_i$. They introduced three restart conditions: the exact non-monotonicity condition, the gradient-mapping based condition and the non-monotonicity implying condition. In this study, we employ the exact non-monotonicity condition, i.e., $F(x^{[k]}_i) > F(x^{[k-1]}_i)$. 

shows the detail of the FISTA with full line search strategy.
2.3.6 Full Algorithm

Based on the discussion in previous sections, now we present the full algorithm which is employed to solve the regularized pseudo log-likelihood function minimization problem (2.2.7). It is a randomized block coordinate descent algorithm with majorization-minimization principle, acceleration technique and full line search strategy. The following Algorithm 3 shows the detail. The initial values of the parameters $\beta \in \mathbb{R}^p$, $\Theta \in \mathbb{R}^{p \times p}$ and $L \in \mathbb{R}^{n \times p}$, i.e., the values of $\beta^{[0]}$, $\Theta^{[0]}$ and $L^{[0]}$, are set as $\beta^{[0]}_j = \log \left( \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right)$, $\Theta^{[0]}_{ij} = 1$ with $\Theta^{[0]}_{ii} = 0$, and $L^{[0]}_{kj}$ is randomly selected from $\{-1, 0, 1\}$, where $i,j = 1, \ldots, p$ and $k = 1, \ldots, n$. 
Algorithm 3 RBCD-MM-AC-FLS (Part 1)

Initialize $(t^{[0]}_\beta, t^{[0]}_d, t^{[0]}_L)$, $(M^{[0]}_\beta, M^{[0]}_d, M^{[0]}_L)$, $(\beta^{[0]}, \Theta^{[0]}, L^{[0]}) = (\beta^{-1}, \Theta^{-1}, L^{-1})$, $\eta_u > 1$, $\eta_d > 1$

for $k=1,2,...$ do

1. randomly choose index $i \in \{1, 2, 3\}$

2. update

   if $i = 1$ then

   $M^{[k]}_\beta \leftarrow M^{[k-1]}_\beta / \eta_d$

   $t^{[k]}_\beta \leftarrow \frac{1 + \sqrt{1 + 4(M^{[k]}_\beta / M^{[k-1]}_\beta)(t^{[k-1]}_\beta)^2}}{2}$

   $z^{[k]}_\beta \leftarrow \beta^{[k-1]} + \frac{t^{[k-1]}_\beta - 1}{t^{[k]}_\beta} (\beta^{[k-1]} - \beta^{[k-2]})$

   $\beta^{[k]} \leftarrow z^{[k]} - \frac{1}{M^{[k]}_\beta} \nabla f(z^{[k]}_\beta)$

   while $f(\beta^{[k]}) > f(z^{[k]}_\beta) + \nabla f(z^{[k]}_\beta)^T (\beta^{[k]} - z^{[k]}_\beta) + \frac{M^{[k]}_\beta}{2} \| \beta^{[k]} - z^{[k]}_\beta \|^2$ do

   $M^{[k]}_\beta \leftarrow \eta_u M^{[k]}_\beta$

   $t^{[k]}_\beta \leftarrow \frac{1 + \sqrt{1 + 4(M^{[k]}_\beta / M^{[k-1]}_\beta)(t^{[k-1]}_\beta)^2}}{2}$

   $z^{[k]}_\beta \leftarrow \beta^{[k-1]} + \frac{t^{[k-1]}_\beta - 1}{t^{[k]}_\beta} (\beta^{[k-1]} - \beta^{[k-2]})$

   $\beta^{[k]} \leftarrow z^{[k]}_\beta - \frac{1}{M^{[k]}_\beta} \nabla f(z^{[k]}_\beta)$

end while

if $F(\beta^{[k]}) > F(\beta^{[k-1]})$ then set $z^{[k]}_\beta = \beta^{[k-1]}$ and redo this step

end if
Algorithm 3 RBCD-MM-AC-FLS (Part 2)

else if $i = 2$ then

\[
M_{\Theta}^{[k]} \leftarrow M_{\Theta}^{[k-1]} / \eta_d
\]

\[
t_{\Theta}^{[k]} \leftarrow \frac{1 + \sqrt{1 + 4(M_{\Theta}^{[k]} / M_{\Theta}^{[k-1]})(t_{\Theta}^{[k-1]})^2}}{2}
\]

\[
z_{\Theta}^{[k]} \leftarrow \Theta_{[k-1]} + \frac{t_{\Theta}^{[k-1]} - 1}{t_{\Theta}^{[k]}} (\Theta_{[k-1]} - \Theta_{[k-2]})
\]

\[
\Theta_{[k]} \leftarrow T_{\lambda_1 / M_{\Theta}^{[k]}} (z_{\Theta}^{[k]} - \frac{1}{M_{\Theta}^{[k]}} \nabla_{\Theta} f(z_{\Theta}^{[k]}))
\]

while $f(\Theta_{[k]}) > f(z_{\Theta}^{[k]}) + \nabla_{\Theta} f(z_{\Theta}^{[k]})^T (\Theta_{[k]} - z_{\Theta}^{[k]}) + \frac{M_{[k]}}{2} \| \Theta_{[k]} - z_{\Theta}^{[k]} \|^2$ do

\[
M_{\Theta}^{[k]} \leftarrow \eta_u M_{\Theta}^{[k]}
\]

\[
t_{\Theta}^{[k]} \leftarrow \frac{1 + \sqrt{1 + 4(M_{\Theta}^{[k]} / M_{\Theta}^{[k-1]})(t_{\Theta}^{[k-1]})^2}}{2}
\]

\[
z_{\Theta}^{[k]} \leftarrow \Theta_{[k-1]} + \frac{t_{\Theta}^{[k-1]} - 1}{t_{\Theta}^{[k]}} (\Theta_{[k-1]} - \Theta_{[k-2]})
\]

\[
\Theta_{[k]} \leftarrow T_{\lambda_1 / M_{\Theta}^{[k]}} (z_{\Theta}^{[k]} - \frac{1}{M_{\Theta}^{[k]}} \nabla_{\Theta} f(z_{\Theta}^{[k]}))
\]

end while

if $F(\Theta_{[k]}) > F(\Theta_{[k-1]})$ then set $z_{\Theta}^{[k]} = \Theta_{[k-1]}$ and redo this step

end if

34
Algorithm 3 RBCD-MM-AC-FLS (Part 3)

else if \( i = 3 \) then

\[
M_L^{[k]} \leftarrow M_L^{[k-1]} / \eta_d
\]

\[
t_L^{[k]} \leftarrow \frac{1 + \sqrt{1 + 4(M_L^{[k]} / M_L^{[k-1]})(t_L^{[k-1]})^2}}{2}
\]

\[
z_L^{[k]} \leftarrow L^{[k-1]} + \frac{t_L^{[k-1]}}{t_L^{[k]}} (L^{[k-1]} - L^{[k-2]})
\]

\[
L^{[k]} \leftarrow \mathcal{D}_{\lambda_2 / M_L^{[k]}} (z_L^{[k]} - \frac{1}{M_L^{[k]}} \nabla_L f(z_L^{[k]}))
\]

while \( f(L^{[k]}) > f(z_L^{[k]}) + \nabla_L f(z_L^{[k]})^T (L^{[k]} - z_L^{[k]}) + \frac{M_L^{[k]}}{2} \| L^{[k]} - z_L^{[k]} \|^2 \) do

\[
M_L^{[k]} \leftarrow \eta_u M_L^{[k]}
\]

\[
t_L^{[k]} \leftarrow \frac{1 + \sqrt{1 + 4(M_L^{[k]} / M_L^{[k-1]})(t_L^{[k-1]})^2}}{2}
\]

\[
z_L^{[k]} \leftarrow L^{[k-1]} + \frac{t_L^{[k-1]}}{t_L^{[k]}} (L^{[k-1]} - L^{[k-2]})
\]

\[
L^{[k]} \leftarrow \mathcal{D}_{\lambda_2 / M_L^{[k]}} (z_L^{[k]} - \frac{1}{M_L^{[k]}} \nabla_L f(z_L^{[k]}))
\]

end while

if \( F(L^{[k]}) > F(L^{[k-1]}) \) then set \( z_L^{[k]} = L^{[k-1]} \) and redo this step

end if

end if

if stopping criterion is satisfied then

return the result

end if

end for
2.4 Convergence Analysis

In this section, we first show the global convergence of the algorithm developed to solve the regularized pseudo log-likelihood function (2.2.7), and second establish its asymptotic rate of convergence.

2.4.1 Global Convergence

In this part, we provide the convergence results of Algorithm 3 which is a randomized block coordinate descent algorithm with the majorization-minimization principle, the acceleration technique, and the full line search strategy. Before proceeding with the analysis of convergence, let us first introduce the Kurdyka-Lojasiewicz (KL) inequality.

**Definition 3.** A function \( \psi(x) \) satisfies the Kurdyka-Lojasiewicz inequality at point \( \bar{x} \in \text{dom}(\partial \psi) \) if in a certain neighborhood \( \mathcal{U} \) of \( \bar{x} \), there exists \( \phi(s) = cs^{1-\theta} \) for some \( c > 0 \) and \( \theta \in (0, 1) \), such that

\[
\phi'(\|\psi(x) - \psi(\bar{x})\|) \text{dist}(0, \partial \psi(x)) \geq 1 \quad \text{for any } x \in \mathcal{U} \cap \text{dom}(\partial \psi) \text{ and } \psi(x) \neq \psi(\bar{x}),
\]

where \( \partial \psi(x) \) is the subdifferential of function \( \psi \) at \( x \), \( \text{dom}(\partial \psi) \triangleq \{ x : \partial \psi(x) \neq \emptyset \} \) and \( \text{dist}(0, \partial \psi(x)) \triangleq \min\{\|y\| : y \in \partial \psi(x)\} \).

The Lojasiewicz inequality was introduced by Lojasiewicz (1993) on real analytic functions, which provided an upper bound for the distance of a critical point of a given real analytic function. Kurdyka (1998) extended it to functions on the o-minimal structure, and the corresponding generalized inequality is called the Kurdyka-Lojasiewicz inequality. There are many commonly used functions that satisfy the KL inequality; For instance, the polynomial function, the logistic loss function, the \( \ell_1 \)-norm, the Euclidean norm, the nuclear norm, and the \( \ell_\infty \)-norm.

Based on the KL inequality, we are ready to present the following results which show the global convergence of Algorithm 3.

**Lemma 5.** (Lemma 2.6 in Xu and Yin (2013)) Consider the optimization problem (2.3.1) where \( F(x) \) is continuous in \( \text{dom}(F) \) and \( \inf_{x \in \text{dom}(F)} F(x) > -\infty \), let \( \{x[k]\} \) be the sequence of Algorithm 3 with bounded extrapolation weight, we assume that

- 1. \( F(x) \) satisfies the KL inequality at \( \bar{x} \);
- 2. \( \nabla f(x) \) is Lipschitz continuous;
- 3. The initial point \( x[0] \) is sufficiently close to \( \bar{x} \), and \( F(x[k]) > F(\bar{x}) \) for \( k \geq 0 \).
Then there is some $B \subset U \cap \text{dom}(\partial F)$, such that $\{x[k]\} \subset B$ and $\{x[k]\}$ converges to a point in $B$.

The above Lemma 5 is key to establish the global convergence result. Its assumptions are however simple and not restrictive. The first assumption of the KL inequality is the most important one, and there are many classes of functions that satisfy it, such as the real analytic functions, the strongly convex functions, the semialgebraic functions, and so on (Xu and Yin, 2013; Attouch et al., 2010). The second Lipschitz continuous gradient condition is common in the analysis of gradient-based algorithms. For the third assumption, the required closeness of the initial point to $\bar{x}$ depends on the target function $F$, the choice of function $\phi$, and the neighborhood $U$. Based on the above Lemma 5, we have the following Corollary 1 which states the convergence of Algorithm 3 to the global minimum. The proof of Corollary 1 is given in Appendix A.5.

**Corollary 1.** Under the assumptions of Lemma 5, the sequence $\{x[k]\}$ generated by Algorithm 3 converges to a global minimizer of (2.3.1) if the initial point is sufficiently close to any global minimizer $\bar{x}$.

Corollary 1 is important since it guarantees that Algorithm 3 converges to the global minimum of the target function under the condition that the initializer is close enough to the global minimizer. This condition is based on Lemma 5 which requires the initial point is close to the critical point. However, in practice, it is not easy to choose a proper initializer which satisfies the condition, since we usually have no information about the value of the critical point. Hence, the following Theorem 1 which shows the global convergence of Algorithm 3 without any condition of the initial point is more general and provide more guidance in practice. The proof of Theorem 1 is given in Appendix A.6.

**Theorem 1.** Under the assumptions of Lemma 5 and the fact that $\{x[k]\}$ has a finite limit point $\bar{x}$ where $F$ satisfies the KL inequality, then the sequence $\{x[k]\}$ converges to $\bar{x}$, which is a critical point of (2.3.1).

The above Theorem 1 shows that Algorithm 3 is globally convergent, i.e., for any chosen initial points, the sequence generated by it converges to a point for which an optimality condition holds. It is worth mentioning that the global convergence of an algorithm, which is contrary to the literal meaning, does not imply that it will converge to a global optimum. It expresses the certainty that the algorithm works for all initial points. However, for our convex target function (2.2.7), the problem becomes easier since for convex functions, any point that satisfies the optimality condition is a global minimum.
2.4.2 Convergence Rate

In the last section, we show the global convergence of Algorithm 3. The following Theorem 2 provides its convergence rate, which is first given in Xu and Yin (2013).

Theorem 2. (Theorem 2.9 in Xu and Yin (2013)) Under the assumptions in Lemma 5, and suppose that the sequence \( \{x^{[k]}\} \) converges to a critical point \( \bar{x} \), at which \( F \) satisfies the KL inequality with \( \phi(s) = cs^{1-\theta} \) for \( c > 0 \) and \( \theta \in [0, 1) \). Then we have the following results:

1. If \( \theta = 0 \), \( x^{[k]} \) converges to \( \bar{x} \) in finitely many iterations,

2. If \( \theta \in (0, \frac{1}{2}] \), \( \|x^{[k]} - \bar{x}\| \leq C\tau^k \) for all \( k \geq k_0 \), for certain \( k_0 > 0 \), \( C > 0 \), \( \tau \in [0, 1) \),

3. If \( \theta \in (\frac{1}{2}, 1) \), \( \|x^{[k]} - \bar{x}\| \leq Ck^{-(1-\theta)/(2\theta-1)} \) for all \( k \geq k_0 \), for certain \( k_0 > 0 \), \( C > 0 \),

where results 1, 2, 3 correspond to finite convergence, linear convergence and sub-linear convergence respectively.

The above Theorem 2 provides the results of convergence rate of Algorithm 3. It is important to note that for different target function \( F \), the parameter \( \theta \) of function \( \phi \) may vary which leads to different convergence rates. If \( \theta = 0 \), it has the slowest finite convergence rate which means the sequence converges to the limit in finite number of iterations. If \( \theta \in (\frac{1}{2}, 1) \), it has the sublinear convergence rate which is a bit slower than the linear convergence rate. If \( \theta \in (0, \frac{1}{2}] \), it has the linear convergence rate which implies the sequence converges to the limit linearly. More precisely, for the general case where the target function is convex, Algorithm 3 achieves a global sublinear iteration complexity of \( O(1/r) \), where \( r \) is the iteration index. For our problem (2.2.7) where the target function is strongly convex, it converges to the global minimum linearly.

2.5 Theoretical Analysis

In this section, we show that our estimators \( (\hat{\beta}, \hat{\Theta}, \hat{L}) \) which minimize the regularized negative pseudo log-likelihood function (2.2.7), achieve satisfactory accuracy with the error measure with respect to the Bregman divergence, and are consistent in identifying the graph structure.
2.5.1 Global Error

First we introduce the Bregman divergence as a discrepancy measure to facilitate the analysis. Denote \((\beta^*, \Theta^*, L^*)\) as the truth, and \((\hat{\beta}, \hat{\Theta}, \hat{L})\) as our estimators which minimize (2.2.7). The Bregman divergence, abbreviated as \(D_f\), is defined as follows, based on the first-order Taylor expansion of \(f(\beta, \Theta, L)\):

\[
D_f(\hat{\beta}, \hat{\Theta}, \hat{L}; \beta^*, \Theta^*, L^*) = f(\hat{\beta}, \hat{\Theta}, \hat{L}) - f(\beta^*, \Theta^*, L^*) - \langle \nabla_\beta f(\beta^*, \Theta^*, L^*), \hat{\beta} - \beta^* \rangle \\
- \langle \nabla_\Theta f(\beta^*, \Theta^*, L^*), \hat{\Theta} - \Theta^* \rangle - \langle \nabla_L f(\beta^*, \Theta^*, L^*), \hat{L} - L^* \rangle,
\]

where \(\nabla_\beta f(\beta^*, \Theta^*, L^*), \nabla_\Theta f(\beta^*, \Theta^*, L^*)\) and \(\nabla_L f(\beta^*, \Theta^*, L^*)\) are partial derivatives of function \(f\) with respect to \(\beta, \Theta\) and \(L\) at the statistical truth \((\beta^*, \Theta^*, L^*)\) respectively. In the following theorem, we define the effective noises \(G(\beta), G(\Theta)\) and \(G(L)\) as

\[
G(\beta) = \nabla_\beta f(\beta^*, \Theta^*, L^*), \\
G(\Theta) = \nabla_\Theta f(\beta^*, \Theta^*, L^*), \\
G(L) = \nabla_L f(\beta^*, \Theta^*, L^*).
\]

In the following Theorem 3 and Corollary 2, we use \(\|G\|_\infty\) to denote \(\max_{j,k} |G_{jk}|\), and \(\|\Theta\|_1 = \sum_{i \neq j} |\Theta_{ij}|\). Define the support set of \(\Theta^*\) by \(S = \{(i,j): \Theta^*_{ij} \neq 0\}\), and its complement set by \(S^c\). Define the decomposable subspaces for the nuclear norm \(\|L^*\|_\ast\) by \(M\) and \(M^\perp\) whose details are given in the proof. Then we have the following results.

**Theorem 3.** Define the cardinality of the support set of \(\Theta^*\) by \(s\), the rank of the matrix \(L^*\) by \(r\). Assume that there exists large enough \(K \geq 0\), \(\vartheta_1, \vartheta_2, \vartheta_3 > 0\) such that

\[
K\left\{\left(\vartheta_1^2 + \vartheta_2^2 + \vartheta_3^2\right)(s + r)D_f\right\}^{1/2} \geq (1 + \vartheta_1)\|(\Theta^* - \Theta)_S\|_1 - \|\Theta_{SC}\|_1 \\
+ (1 + \vartheta_2)\|(L^* - L)_M\|_\ast - \|(L^* - L)_{M^\perp}\|_\ast \\
+ \vartheta_3\|\beta^* - \beta\|_1,
\]

(2.5.1)

for any matrix \(\Theta \in \mathbb{R}^{p \times p}, L \in \mathbb{R}^{n \times p}\) and \(\beta \in \mathbb{R}^p\). Then on events \(\{\|G(\beta)\|_\infty \leq \omega_0\},\ \{\|G(\Theta)\|_\infty \leq \omega_1\}\) and \(\{\|G(L)\|_2 \leq \omega_2\}\), with \(\lambda_1 = \omega_1 + 1\) and \(\lambda_2 = \omega_2 + 1\), then the solution \((\hat{\beta}, \hat{\Theta}, \hat{L})\) in (2.2.7) satisfies

\[
D_f \leq K^2(\omega_0^2 + 4\omega_1^2 + 4\omega_2^2)(s + r),
\]

(2.5.2)
where $D_f$ is the abbreviation of the Bregman divergence $D_f(\hat{\beta}, \hat{\Theta}, \hat{L}; \beta^*, \Theta^*, L^*)$.

Proof. The proof details of Theorem 3 are provided in Appendix A.7. Here we only outline the proof strategy.

(1) Denote $(\beta^*, \Theta^*, L^*)$ as the statistical truth, and $(\hat{\beta}, \hat{\Theta}, \hat{L})$ as the minimizer of (2.2.7), we have the following first-order Taylor expansion of function $f(\beta, \Theta, L)$:

\[
\begin{align*}
f(\hat{\beta}, \hat{\Theta}, \hat{L}) &= f(\beta^*, \Theta^*, L^*) + \langle \nabla_\beta f(\beta^*, \Theta^*, L^*), \hat{\beta} - \beta^* \rangle \\
&\quad + \langle \nabla_\Theta f(\beta^*, \Theta^*, L^*), \hat{\Theta} - \Theta^* \rangle + \langle \nabla_L f(\beta^*, \Theta^*, L^*), \hat{L} - L^* \rangle + D_f,
\end{align*}
\]

where $D_f$ is the abbreviation of Bregman divergence $D_f(\hat{\beta}, \hat{\Theta}, \hat{L}; \beta^*, \Theta^*, L^*)$. Since $(\hat{\beta}, \hat{\Theta}, \hat{L})$ is the minimizer of (2.2.7), we have

\[
f(\hat{\beta}, \hat{\Theta}, \hat{L}) + \lambda_1\|\hat{\Theta}\|_1 + \lambda_2\|\hat{L}\|_* \leq f(\beta^*, \Theta^*, L^*) + \lambda_1\|\Theta^*\|_1 + \lambda_2\|L^*\|_*,
\]

which means

\[
f(\hat{\beta}, \hat{\Theta}, \hat{L}) - f(\beta^*, \Theta^*, L^*) \leq \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \lambda_2(\|L^*\|_* - \|\hat{L}\|_*).
\]

Plugging the equation (2.5.3) into (2.5.4), we have

\[
\begin{align*}
D_f + \langle \nabla_\beta f(\beta^*, \Theta^*, L^*), \hat{\beta} - \beta^* \rangle + \langle \nabla_\Theta f(\beta^*, \Theta^*, L^*), \hat{\Theta} - \Theta^* \rangle + \langle \nabla_L f(\beta^*, \Theta^*, L^*), \hat{L} - L^* \rangle \\
\leq \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \lambda_2(\|L^*\|_* - \|\hat{L}\|_*).
\end{align*}
\]

By the definition of effective noises $G(\beta)$, $G(\Theta)$ and $G(L)$, and on the events \{$\|G(\beta)\|_\infty \leq \omega_0\}$, \{$\|G(\Theta)\|_\infty \leq \omega_1\}$ and \{$\|G(L)\|_2 \leq \omega_2\}$, after some simple mathematical operations, we have

\[
D_f \leq \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \omega_1\|\Theta^* - \hat{\Theta}\|_1 \\
+ \lambda_2(\|L^*\|_* - \|\hat{L}\|_*) + \omega_2\|L^* - \hat{L}\|_* + \omega_0\|\beta^* - \hat{\beta}\|_1.
\]

(2) For the $\ell_1$ norm of $\Theta^*$ and $\hat{\Theta}$ in (2.5.6), by the decomposable property of the $\ell_1$ norm, we have

\[
\begin{align*}
\lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \omega_1\|\Theta^* - \hat{\Theta}\|_1 \leq (\omega_1 + \lambda_1)\|\Theta_S^* - \hat{\Theta}_S\|_1 \\
+ (\omega_1 - \lambda_1)\|\hat{\Theta}_{SC}\|_1,
\end{align*}
\]

where $S$ is the support set of $\Theta^*$, $S^c$ is the complement set, $\Theta_S^*$ is the sub-matrix of $\Theta^*$ with indices in $S$, and the other terms are defined similarly.
(3) For the the nuclear norm of $L^*$ and $\hat{L}$ in (2.5.6), first we define two subspaces of matrices

$$M = \{ A \in \mathbb{R}^{n \times p} \mid \text{rowspan}(A) \subseteq V, \text{colspan}(A) \subseteq U \},$$

and

$$\bar{M}^\perp = \{ B \in \mathbb{R}^{n \times p} \mid \text{rowspan}(B) \subseteq V^\perp, \text{colspan}(B) \subseteq U^\perp \}.$$ 

Here $V$ and $U$ are spaces spanned by the top $r$ left and right singular vectors of matrix $L^*$ respectively. $V^\perp$ and $U^\perp$ denote the subspaces orthogonal to $V$ and $U$ respectively. Then by the decomposable property of the nuclear norm, we have

$$\lambda_2(\|L^*\|_* - \|\hat{L}\|_*) + \omega_2\|L^* - \hat{L}\|_* \leq (\omega_2 + \lambda_2)\|(L^* - \hat{L})_{\bar{M}}\|_*$$

$$+ (\omega_2 - \lambda_2)\|(L^* - \hat{L})_{\bar{M}^\perp}\|_* , \quad (2.5.8)$$

where $L_{\bar{M}}$ denotes the projection of matrix $L$ onto the subspace $\bar{M}$, and the other term is defined similarly.

(4) Combine results (2.5.7) and (2.5.8) and set $\lambda_1 = \omega_1 + 1$, $\lambda_2 = \omega_2 + 1$, then by the regularity condition (2.5.1), we have

$$D_f \leq (\omega_1 + \lambda_1)\|\Theta^*_S - \hat{\Theta}_S\|_1 + (\omega_1 - \lambda_1)\|\hat{\Theta}_{SC}\|_1 + (\omega_2 + \lambda_2)\|(L^* - \hat{L})_{\bar{M}}\|_*$$

$$+ (\omega_2 - \lambda_2)\|(L^* - \hat{L})_{\bar{M}^\perp}\|_* + \omega_0\|\beta^* - \hat{\beta}\|_1,$$

$$= (1 + 2\omega_1)\|\Theta^*_S - \hat{\Theta}_S\|_1 - \|\hat{\Theta}_{SC}\|_1 + (1 + 2\omega_2)\|(L^* - \hat{L})_{\bar{M}}\|_*$$

$$- \|(L^* - \hat{L})_{\bar{M}^\perp}\|_* + \omega_0\|\beta^* - \hat{\beta}\|_1,$$

$$\leq K\{ (\omega_0^2 + 4\omega_1^2 + 4\omega_2^2)(s + r)D_f \}^{1/2},$$

which means

$$D_f \leq K^2(\omega_0^2 + 4\omega_1^2 + 4\omega_2^2)(s + r).$$

The Bregman divergence, which is a measure of distance between two points, plays an important role in the analysis of convex functions. To the best of our knowledge, Theorem 3 is the first theoretical result that employs the Bregman divergence as the error measure of multiple parameters in the graph learning problem with the existence of latent variables. Theorem 3 states that the Bregman divergence between the truth $(\beta^*, \Theta^*, L^*)$ and our estimator $(\hat{\beta}, \hat{\Theta}, \hat{L})$ is bounded under the condition (2.5.1). Here the condition (2.5.1) extends the comparison regularity condition in She et al. (2016, 2019) which considers the $\ell_1$ norm only to our case that
includes multiple penalties. The comparison regularity condition can be viewed as an extension of compatibility and restricted-eigenvalue assumptions (Van De Geer et al., 2009) in the graph setting. However, it is less restrictive, since the restricted-eigenvalue assumption is a sufficient condition for the comparison regularity condition to hold (She et al., 2016). Moreover, the Restricted Isometry Property (RIP), which was first shown in Compressed Sensing (CS) by Candès et al. (2006), is a sufficient condition for both the restricted-eigenvalue condition and the comparison regularity condition to hold (Van De Geer et al., 2009).

From the proof above, we can know that the result in Theorem 3 is deterministic, and it is valid for any function \( f \) which satisfies the condition (2.5.1). The following Corollary 2 is based on our model where the random variable only takes the binary value 0 or 1.

Corollary 2. In our model, where the random variable \( X_{ij} \) only takes value 0 or 1, the global error inequality (2.5.2) holds with probability \( 1 - cp^{-\alpha} \), and the error bound on the right-hand side becomes \( K^2[c_1(1 + \alpha)n \log p + 2c_2(2 + \alpha)\sqrt{n} \log p + 4c_3p + 4c_4 \alpha \log p](s + r) \), where \( c = \max(2, C) \) and \( \alpha, c_1, c_2, c_3, c_4, C \) are positive constants.

Proof. The proof details of Corollary 2 are provided in Appendix A.8. Here we outline the proof strategy. Since the random variable \( X_{ij} \) in the binary Markov random field only takes value 0 or 1, each element in \( G(\beta), G(\Theta), \) and \( G(L) \) is a sub-gaussian random variable. Then the probabilities of events \( \{\|G(\beta)\|_{\infty} \leq \omega_0\} \), \( \{\|G(\Theta)\|_{\infty} \leq \omega_1\} \) and \( \{\|G(L)\|_2 \leq \omega_2\} \) in Theorem 3 can be computed as follows.

1. For the vector \( G(\beta) \), we have

\[
G(\beta) = -X^T 1_n + \left[ \frac{\exp(1_{n} \beta^T + X\Theta^* + L^*)}{1 + \exp(1_{n} \beta^T + X\Theta^* + L^*)} \right] 1_n,
\]

and its \( j \)th component

\[
G(\beta)_j = \sum_{i=1}^{n} \left[ -X + \frac{\exp(1_N \beta^T + X S^* + L^*)}{1 + \exp(1_N \beta^T + X S^* + L^*)} \right]_{ij},
\]

where matrix \( X \) has row independence, each entry in matrix \( X \) takes value 0 or 1, and the sigmoid function \( f(\cdot) = \frac{\exp(\cdot)}{1 + \exp(\cdot)} \in (0, 1) \). Hence \( G(\beta)_j \) is bounded in \( [-n, n] \), and it is a sub-gaussian random variable. By the property of sub-gaussian, we have

\[
P\left[\|G(\beta)\|_{\infty} \geq \omega_0\right] \leq 2 \exp\left(\log p - \frac{\omega_0^2}{c_1 n}\right).
\]
Let $\omega_0^2 = c_1(1 + \alpha)n \log p$, we have the following result

$$P\left[ \|G(\beta)\|_\infty \geq \omega_0 \right] \leq 2p^{-\alpha}, \quad (2.5.9)$$

where $c_1, \alpha$ are positive constants.

(2) For the matrix $G(\Theta)$, we can re-write it as

$$G(\Theta) = -X^T \left( X + \frac{\exp(1_n\beta^T \star X + X\Theta^* + L^*)}{1 + \exp(1_n\beta^T \star X + X\Theta^* + L^*)} \right) = -X^T Z,$$

and the $(i,j)$th component of $G(\Theta)$ is

$$G(\Theta)_{ij} = \sum_{k=1}^{n} -X_{ki}Z_{kj},$$

where $-X_{ki}$ is bounded in $[-1, 0]$ and $Z_{kj}$ is bounded in $[0, 2]$. Hence $-X_{ki}Z_{kj}$ is bounded and is a sub-gaussian random variable. By the row independence of $X$ and the property of sub-gaussian, we have

$$P\left[ \|G(\Theta)\|_\infty \geq \omega_1 \right] \leq 2 \exp(2 \log p - \frac{\omega_1^2}{c_2 n}),$$

where $c_2$ is a positive constant. Let $\omega_1^2 = c_2(2 + \alpha)n \log p$, we have

$$P\left[ \|G(\Theta)\|_\infty \geq \omega_1 \right] \leq 2p^{-\alpha}. \quad (2.5.10)$$

(3) For the matrix $G(L)$, we denote it as $M$, that is

$$G(L) = -X + \left[ \frac{\exp(1_n\beta^T \star X + X\Theta^* + L^*)}{1 + \exp(1_n\beta^T \star X + X\Theta^* + L^*)} \right] = M.$$

Similar to the analysis above, we know that each entry in $G(L)$ is bounded, thus is a sub-gaussian random variable. Let $M_1, \ldots, M_n$ be the $n$ rows of $M$, then the column vector $Mu$ has coefficients $M_iu = \sum_{j=1}^{p} M_{ij}u_j$ for $i = 1, \ldots, n$. Since $M_{ij}$ is a sub-gaussian random variable and $\|u\|_2 = 1$, then $M_iu$ is also a sub-gaussian. By the row independence of $X$ and the property of sub-gaussian, we have

$$P(\|Mu\| \geq \omega_2) \leq B \exp(-b\omega_2^2),$$

where $B$ and $b$ are positive constants. Since the spectral norm of matrix $M$ is defined
as \( \|M\| = \sup_{\|u\|_2 = 1} \|Mu\| \), using an \( \varepsilon \)-net argument, we have the following result

\[
P(\|M\|_2 \geq \omega_2) \leq C \exp(p \log 2 - \frac{b}{4} \omega_2^2).
\]

Let \( \omega_2^2 = \frac{4(p \log 2 + \alpha \log p)}{b} = c_3p + c_4 \alpha \log p \), we have

\[
P(\|G(L)\|_2 \geq \omega_2) \leq Cp^{-\alpha}, \quad (2.5.11)
\]

where \( c_3, c_4 \) and \( C \) are positive constants.

(4) Combining results (2.5.9), (2.5.10) and (2.5.11) with \( \omega_2^0 = c_1(1 + \alpha)n \log p \), \( \omega_1^2 = c_2(2 + \alpha)n \log p \) and \( \omega_2^2 = c_3p + c_4 \alpha \log p \), then we have

\[
D_f \leq K^2(c_1(1 + \alpha)n \log p + 4c_2(2 + \alpha)n \log p + 4c_3p + 4c_4 \alpha \log p)(s + r)
\]

with probability \( 1 - cp^{-\alpha} \), where \( c = \max(2, C) \).

Corollary 2 states that in our case where the random variable takes value 0 or 1, the Bregman divergence between the truth \((\beta^*, \Theta^*, L^*)\) and our estimator \((\hat{\beta}, \hat{\Theta}, \hat{L})\) achieves satisfactory accuracy with high probability. Moreover, the explicit bound of Corollary 2 is non-asymptotic, which means the result holds for all choices of the pair \( (n, p) \). It is worth mentioning that in Corollary 2, the only used fact is that the random variable is bounded. Nothing else is employed with regard to the distribution of the model or other sophisticated assumption. Thus the result in Corollary 2 is nonparametric and can be generalized to models with more complex data structures.

The above Theorem 3 and Corollary 2 focus on the joint error of our parameters. In the following section, we show the result of consistent graph recovery which is important in graph learning problems.

### 2.5.2 Consistent Estimation of Graph Structure

Before stating the theorem, we introduce related notations first. Here \( \text{vec}(\cdot) \) denotes the vectorization of matrix by stacking the columns of the matrix. To characterize the selection outcome of parameter \( \Theta \), we define \( nz(\text{vec}(\Theta)) = \{ j : \text{vec}(\Theta)_j \neq 0 \} \) and \( z(\text{vec}(\Theta)) = \{ j : \text{vec}(\Theta)_j = 0 \} \). The number of nonzero parameters in \( \text{vec}(\Theta^*) \) is denoted as \( d_{nz} = |nz(\text{vec}(\Theta^*))| \), and the number of zeros is \( d_z = |z(\text{vec}(\Theta^*))| \). Denote \( \Sigma = \frac{\partial^2}{\partial \Theta^2} f(\beta^*, \Theta^*, L^*) \). Denote \( \text{vec}(\Theta)_{nz} \) as the sub-vector of \( \text{vec}(\hat{\Theta}) \) with indices in \( nz(\text{vec}(\Theta^*)) \), \( \Sigma_{z,nz} \) as the sub-matrix with the row indices in \( z(\text{vec}(\Theta^*)) \), ...
and column indices in $nz(vec(\Theta^*))$, and $\Sigma_{nz}$ as the sub-matrix with the same column and row indices $nz(vec(\Theta^*))$.

We introduce two useful quantities

$$
\mu := \mu_{\min}(\Sigma_{nz}), \quad \kappa := \frac{\max_{i \in z} \|\Sigma_{i,nz}\|_2}{\sqrt{d_{nz}}}
$$

where $\mu_{\min}$ denotes the smallest eigenvalue. Then we have the following result.

**Theorem 4.** Assume that $\frac{\mu_{\min}d_{nz}}{\mu + \kappa d_{nz}} \lambda_1 \geq t_1 K_1 + t_2 \min |\beta^*| + t_3 + K_2$ and $\mu_{\min} \min_{nz}(vec(\Theta^*)) - \lambda_1 d_{nz} \geq t_1 K_1 + t_2 \min |\beta^*| + t_3 + K_2$, here $t_1 = \sqrt{c_1 n (2 + \alpha) \log p}$, $t_2 = \sqrt{c_2 n (2 + \alpha) \log p}$, $t_3 = \sqrt{c_3 n \log 2 + (2 + \alpha) \log p}$, where $\alpha$, $c_1$, $c_2$, $c_3$, $K_1$, $K_2$ are positive constants. Then in our problem (2.2.7), the probability of successful graph recovery, denoted by $P_s$, i.e., there exists one estimator $(\hat{\beta}, \hat{\Theta}, \hat{L})$ such that $nz(vec(\hat{\Theta})) = nz(vec(\Theta^*))$, is at least $1 - 2p^{-\alpha}$.

**Proof.** The proof details of Theorem 4 are provided in Appendix A.9. Here we outline the proof strategy. If $(\hat{\beta}, \hat{\Theta}, \hat{L})$ is the minimizer of our target function, the KKT condition for the $\ell_1$ norm reads

$$
\frac{\partial}{\partial \Theta} f(\hat{\beta}, \hat{\Theta}, \hat{L}) + \lambda_1 \hat{Z} = 0, \quad (2.5.12)
$$

where $\hat{Z}$ is a subgradient of the function $\|\hat{\Theta}\|_1$. Then we have the Taylor expansion of the gradient function

$$
\begin{align*}
vec\left( \frac{\partial}{\partial \Theta} f(\hat{\beta}, \hat{\Theta}, \hat{L}) \right) &= vec\left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) + \frac{\partial^2}{\partial \Theta^2} f(\beta^*, \Theta^*, L^*)(vec(\hat{\Theta}) - vec(\Theta^*)) \\
&+ \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*)(vec(\hat{L}) - vec(L^*)) \\
&+ \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*)(\hat{\beta} - \beta^*) + R_{\Theta}, \quad (2.5.13)
\end{align*}
$$

where $R_{\Theta}$ is the higher order term and $vec(\cdot)$ is the vectorization of matrix. Plugging above function (2.5.13) into the KKT function (2.5.12), we have

$$
0 = \frac{\partial^2}{\partial \Theta^2} f(\beta^*, \Theta^*, L^*)(vec(\hat{\Theta}) - vec(\Theta^*)) + vec\left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) \\
+ \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*)(vec(\hat{L}) - vec(L^*)) + \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*)(\hat{\beta} - \beta^*) \\
+ R_{\Theta} + \lambda_1 vec(\hat{Z}). \quad (2.5.14)
$$
Denote  
\[ \Sigma = \frac{\partial^2}{\partial \Theta^2} f(\beta^*, \Theta^*, L^*), \]
and
\[ \varepsilon = \text{vec} \left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) + \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) (\text{vec}(\hat{L}) - \text{vec}(L^*)) \]
\[ + \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) (\hat{\beta} - \beta^*) + R, \]  \tag{2.5.15}
the function (2.5.14) can be written as the following simple form
\[ \Sigma \text{vec}(\hat{\Theta}) = \Sigma \text{vec}(\Theta^*) - \varepsilon - \lambda_1 \text{vec}(\hat{Z}). \]  \tag{2.5.16}
Denote \( nz = \{ j : \text{vec}(\Theta^*)_j \neq 0 \} \) and \( z = \{ j : \text{vec}(\Theta^*)_j = 0 \} \), where \( \text{vec}(\hat{\Theta})_{nz} \) is the sub-vector with index \( nz \). Then (2.5.16) is equivalent to
\[ S_z \text{vec}(\hat{\Theta})_z = \Sigma_{z, nz}^{-1} \varepsilon_{nz} - \lambda_1 \Sigma_{z, nz}^{-1} \text{vec}(\hat{Z})_{nz} \]
\[ - \lambda_1 \text{vec}(\hat{Z})_z, \]  \tag{2.5.17}
and
\[ \text{vec}(\hat{\Theta})_{nz} = \text{vec}(\Theta^*)_{nz} - \Sigma_{nz}^{-1} \varepsilon_{nz} - \lambda_1 \Sigma_{z, nz}^{-1} \text{vec}(\hat{Z})_{nz} - \Sigma_{nz}^{-1} \Sigma_{nz, z} \text{vec}(\hat{\Theta})_z, \]  \tag{2.5.18}
where \( S_z = \Sigma_z - \Sigma_{z, nz} \Sigma_{nz}^{-1} \Sigma_{nz, z} \).

Here we focus on \( P_s \), the probability that there exists one estimator \( \hat{\Theta} \) such that \( nz(\text{vec}(\hat{\Theta})) = nz(\text{vec}(\Theta^*)) \). Define
\[ A \triangleq \left\{ \left| \Sigma_{z, nz}^{-1} \varepsilon_{nz} - \varepsilon_z + \lambda_1 \Sigma_{z, nz}^{-1} \text{vec}(\hat{Z})_{nz} \right| \leq \lambda_1 \right\}, \]
\[ V \triangleq \left\{ \left| \Sigma_{nz}^{-1} \varepsilon_{nz} \right| + \lambda_1 \left| \Sigma_{nz}^{-1} s \right| < \left| \text{vec}(\Theta^*)_{nz} \right| \text{ for any } s \text{ satisfying } |s| \leq 1 \right\}. \]
Then we have
\[ 1 - P_s \leq P(A^c \cup V^c) \leq P(A^c) + P(V^c). \]
For \( P(A^c) \) and \( P(V^c) \), we have
\[ P(A^c) \leq P\left( \| \varepsilon \|_\infty \geq \frac{\mu - \kappa d_{nz}}{\mu + \kappa d_{nz} \lambda_1} \right), \]
and
\[ P(V^c) \leq P\left( \|\varepsilon\|_\infty \geq \mu \cdot \min \left| \text{vec}(\Theta^*)_{nz} \right| - \lambda_1 d_{nz} \right). \]

By the definition (2.5.15) of \( \varepsilon \), we have the following result
\[
\|\varepsilon\|_\infty \leq \left\| \text{vec} \left( \frac{\partial^2}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) \right\|_\infty + \left\| \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty \left\| \text{vec}(\hat{L}) - \text{vec}(L^*) \right\|_\infty
+ \left\| \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty \|\hat{\beta} - \beta^*\|_\infty + \|R_\Theta\|_\infty.
\]

(1) For the term \( \left\| \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty \), we have

\[
\frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) = (I_{pxp} \otimes X^T) \text{Diag}\{ \text{vec}(b''(1_n\beta^T + XS + L)) \} \}
\]

where \( \otimes \) is the Kronecker product between matrices, and \( \text{Diag} \) converts a vector into a diagonal matrix. Similar to the analysis in the proof of Corollary 2, by the property of sub-gaussian, we can obtain that let \( t_1^2 = c_1 n (2 + \alpha) \log p \), where \( t_1, c_1, \alpha > 0 \), then we have

\[
P\left( \left\| \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty > t_1 \right) \leq p^{-\alpha}.
\]

(2) For the term \( \left\| \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty \), we have

\[
\frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) = (I_{pxp} \otimes 1_n) \{ (I_{pxp} \otimes b''(1_n\beta^T + XS + L)) \} \}
\]

Similarly, by the property of sub-gaussian, we can obtain that let \( t_2^2 = c_2 n (2 + \alpha) \log p \), where \( t_2, c_2 > 0 \), we have

\[
P\left( \left\| \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty > t_2 \right) \leq p^{-\alpha}.
\]

(3) For the term \( \left\| \text{vec} \left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) \right\|_\infty \), by the result (2.5.10), we can know that let \( t_3^2 = c_3 n [\log 2 + (2 + \alpha) \log p] \), where \( t_3, c_3 > 0 \), we have

\[
P\left( \left\| \text{vec} \left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) \right\|_\infty > t_3 \right) \leq p^{-\alpha}.
\]

47
(4) For the rest terms, since $\hat{\beta}$ is assumed sign consistent and $\hat{L}$ is assumed rank consistent, then we have $\|\hat{\beta} - \beta^*\|_\infty < \min |\beta^*|$, $\|\text{vec}(\hat{L}) - \text{vec}(L^*)\|_\infty \leq K_1$, and the higher order term $\|R_\Theta\|_\infty \leq K_2$, where $K_1, K_2$ are some positive constants. Combining the above results (1)-(4), we have the following result

$$P(\|\varepsilon\|_\infty \geq t_1 K_1 + t_2 \min |\beta^*| + t_3 + K_2) \leq p^{-\alpha}.$$ 

Hence if

$$\frac{\mu - \kappa d_{nz}}{\mu + \kappa d_{nz}} \lambda_1 \geq t_1 K_1 + t_2 \min |\beta^*| + t_3 + K_2,$$

then we have

$$P(A^c) \leq P\left(\|\varepsilon\|_\infty \geq \frac{\mu - \kappa d_{nz}}{\mu + \kappa d_{nz}} \lambda_1\right) \leq p^{-\alpha}.$$ 

Similarly, if

$$\mu \cdot \min |\text{vec}(\Theta^*)_{nz}| - \mu_1 d_{nz} \geq t_1 K_1 + t_2 \min |\beta^*| + t_3 + K_2,$$

then we have

$$P(V^c) \leq P\left(\|\varepsilon\|_\infty \geq \mu \cdot \min |\text{vec}(\Theta^*)_{nz}| - \mu_1 d_{nz}\right) \leq p^{-\alpha}.$$ 

Since $1 - P_s \leq P(A^c \cup V^c) \leq P(A^c) + P(V^c)$, we can know that

$$P_s \geq 1 - 2p^{-\alpha}.$$ 

Theorem 4 states that the probability of consistent graph recovery is high under two assumptions. The first assumption about $\kappa$ can be viewed as an extension of the irrepresentable condition (Zhao and Yu, 2006) in our problem. Intuitively, the size of $\kappa$ measures the correlations between the relevant predictors and the irrelevant predictors. The second assumption about $\min |\text{vec}(\Theta^*)_{nz}|$ describes how small the minimum signal strength can be. These two assumptions are quite standard in the analysis of models with an $\ell_1$ penalty (She et al., 2009, 2013; Ravikumar et al., 2010, 2011). Moreover, compared with the conditions in the previous studies of the graph estimation problems with latent variables (Chandrasekaran et al., 2012; Frot et al., 2017; Nussbaum and Giesen, 2019), our assumptions are less restrictive and more straightforward. In addition, it is worth mentioning that if there is no latent variable, and we multiply a term $1/n$ in front of $f(\beta, \Theta, L)$ in (2.2.7), which is the same as the setting in Ravikumar et al. (2010), then the minimum magnitude
nonzero entry in $\Theta^*$ is at the order of $\sqrt{(\log p)/n}$, which has been shown to be the optimal rate in the existing studies of the graphical learning (Ravikumar et al., 2011; Cai et al., 2011).

2.6 Simulation Study

In this section, we describe the simulation studies conducted and present the results under three different settings. Commonly used evaluation criteria are employed to compare the performance of our new method and its competitors.

2.6.1 Simulation Setting

To investigate the performance of our new method, simulation experiments were performed under the following settings: the sample size of $n = 500$, the number of nodes set as $p = 10, 30, 50, 70$ and 100, and the number of latent variables set as $k = 1, 3$ and 10 respectively. We also conducted simulations with a larger sample size $n = 1000$, but since the results were similar, these results are omitted here. This method also worked when the value of $p$ became larger, but a bit more time was needed. Given each setting, 100 independent datasets were generated via the `BMN.samples` function of the `TestBMN` package in R which samples from a binary Markov random field using Gibbs sampling. In each data set, the edges in the graph were uniformly selected from all possible combinations of nodes. In different settings, the number of selected edges in a graph differ. Details are given in the following case I, II and III (in section 2.6.2). For a given graph, the nonzero parameter $\theta_{ij}$ was drawn uniformly from $[-0.5, -0.1] \cup [0.1, 0.5]$. For the `BMN.samples` function, the parameter of the burn-in tolerance was set as $n_0 = 1000$, and the skip parameter kept the default setting skip = 1. All simulation studies were conducted on a machine with 3.0 GHz CPU and 16 GB RAM.

We then compared the performance of our SIMPLE method with the competitors introduced in Section 2.1.5: the node-wise logistic regression with “BOTH” edge criteria (denoted as “Node-wise-BOTH”), the node-wise logistic regression with “OR” edge criteria (denoted as “Node-wise-OR”), and the pseudo-likelihood approach (denoted as “Pseudo”). The SIMPLE method was performed via Algorithm 3 whose details are given in Section 2.3, and the initial values of parameters $\beta^{[0]}$, $\Theta^{[0]}$ and $L^{[0]}$ were set as the description in Section 2.3.6. The “Node-wise-BOTH” and the “Node-wise-OR” approaches were conducted via the `glmnet` package in R. The “Pseudo” method, of which the original `BMN` package is no longer available, was performed via an algorithm written by ourselves.
To reveal the potential of each method and eliminate the interference of different tuning parameter selection schemes, we varied the regularization parameter along some pre-specified ranges. There is one penalty term in the “Node-wise-BOTH”, the “Node-wise-OR” and the “Pseudo” method, and the ranges of the regularization parameters in these three methods were based on the default setting of the glmnet package. There are two penalty terms in the SIMPLE method, and we assigned 10 values for each parameter, thus there were 100 grid values in total for the tuning parameters. It is worth mentioning that under different settings, the grid of the tuning parameters varied. In general, the range of the ℓ₁ norm in the SIMPLE method was (0, 5) and the range of the nuclear norm was (0, 3).

The following criteria were used to compare the graph estimation performances of the aforementioned methods:

- **True Positive Rate (TPR)**, also called recall or sensitivity, defined as

  \[
  TPR = \frac{TP}{TP + FN},
  \]

  where the True Positive (TP) is the outcome where the model correctly identifies the positive class in a binary classification problem, and the False Negative (FN) is the outcome where the model wrongly identifies the positive class as negative. Hence the TPR is the ratio of correctly estimated edges to all true edges. A higher TPR is preferred since a large number of true edges are correctly estimated.

- **False Positive Rate (FPR)**, defined as

  \[
  FPR = \frac{FP}{FP + TN},
  \]

  where the False Positive (FP) is the outcome where the model incorrectly identifies the negative class as positive, and the True Negative (TN) is the outcome where the model correctly identifies the negative class. Hence the FPR is the ratio of wrongly estimated edges to the number of actual non-edges. A lower FPR is preferred since only a small number of non-edges are wrongly marked as edges.

To evaluate the performance of each method, the results were shown by the ROC curve which is a graphical plot that summarizes the relationship of the true positive rate against the false positive rate. It is widely used to compare the performance of different approaches in graphical estimation. In our case, the x-axis
in the ROC curve represents the false positive rate, and the y-axis shows the average of the highest true positive rate at a certain false positive rate level. Thus, the higher the curve, the more accurate the method is. For this study, we plotted the ROC curve with a small range for the false positive rate, since on the one hand, a low FPR is preferred as it indicates that only a small number of non-edges are wrongly estimated, on the other hand, all methods express similar performances when a high FPR is tolerated.

2.6.2 Simulation Results

Case I: one latent variable with strong connections

In the first setting, the latent variable was assumed to be highly connected with the observed variables, and the observed variables themselves had rare connections. Concretely, we assumed that there was only one latent node which had edges with 70% of observed nodes, and there was a total of $30\% \times p$ random edges between the observed nodes themselves. To evaluate the performance of each method, 100 independent datasets were generated and the average results were reported. In each data set, the observed nodes which had edges were uniformly selected.

Figure 2.2 shows the estimation results under different settings ($p = 10, 30, 50, 100$) when one latent variable has a high impact on the observed variables. The red line represents our SIMPLE method, which has the highest performance across all the settings. The yellow line represents the Pseudo method, which employs a similar pseudo-likelihood approach as the SIMPLE method, but ignores the latent variables. Compared with our method, the Pseudo method yields less accurate results, which indicates that it is worthwhile taking the latent variable into account.

The blue line represents the Node-wise-OR approach, in which the edge is counted if it is indicated by any one of the nodes that it links with. This approach produced the lowest performance among all the methods, especially when we focus on the low FPR. This is due to the “OR” constraint, which tends to overestimate the edges. The green line represents the Node-wise-BOTH method, in which the edge is counted if it is indicated by both nodes that it connects with. This strict “BOTH” constraint provided better results than the “OR” constraint when the graph is sparse. However, it remains less accurate than the SIMPLE method.

Case II: three latent variables with moderate connections

In the second setting, we assumed that three latent variables existed. Moreover, the strong connections between these latent and observed variables in the first setting
Figure 2.2: Performance of different estimation methods when there is one latent variable. The red line represents the SIMPLE method, the green line represents the Node-wise-BOTH method, the blue line represents the Node-wise-OR method, and the yellow line represents the Pseudo approach. 

slightly decreased in this case, and the connections within the observed variables increased. In other words, we reduced the impact of the latent variables. Concretely, we assumed the presence of three latent variables with edges to 50% of the observed nodes, and the number of edges within these observed variables was set to 100% × p.

Figure 2.3 shows the performance of these methods under different settings (p = 10, 30, 50, 70) when there are three latent nodes. The SIMPLE method (the red line) has the highest performance in all cases. The Pseudo method (yellow line) does not perform well, because it ignores the latent variables. The Node-wise-OR approach is not accurate at a low FPR as it tends to overestimate the edges. The Node-wise-BOTH method outperforms the Node-wise-OR approach due to its strict edges constraint, but it does not perform as well as our new approach.

Case III: ten latent variables with weak connections

In the third setting, we compared all the methods when there were ten latent variables. Moreover, we further decreased the impact of the latent variables and increased the connections within the observed variables. In this case, we assumed that the ten latent nodes connected with 30% of the observed nodes, and the number of edges within these observed variables was set to 300% × p.
Figure 2.3: Performance of different estimation methods when there are three moderately connected latent variables. The red line represents the SIMPLE method, the green line represents the Node-wise-BOTH method, the blue line represents the Node-wise-OR method, and the yellow line represents the Pseudo approach.

Figure 2.4 shows the performance of these methods under different settings \((p = 30, 50, 70)\) when there are ten latent nodes. Similar to previous settings, the SIMPLE method (red line) has the best performance in all cases. The Pseudo method (yellow line) has the lowest performance when the latent variables are ignored. The Node-wise-OR approach does not perform well at the low FPR. This coincides with its performance under previous settings. Compared with the Node-wise-OR approach, the Node-wise-BOTH method yields better results, but it does not perform as well as the SIMPLE method.

The above simulation results show that the SIMPLE method provides uniformly better results than its competitors when latent variables exist. It is worth mentioning that we also compared the performance of these methods when there was no latent variable, and the SIMPLE method still had comparable results with the other methods. This means the SIMPLE method is the most reliable when taking the latent variables into account, regardless of whether the latent variables exist or not, which makes it more adaptive to complex situations in real-world problems.
Figure 2.4: Performance of different estimation methods when there are ten weakly connected latent variables. The red line represents the SIMPLE method, the green line represents the Node-wise-BOTH method, the blue line represents the Node-wise-OR method, and the yellow line represents the Pseudo approach.

2.7 Real-world Data Applications

To illustrate the value of the SIMPLE method, in this section we apply it to two different datasets which contain the stock prices of oil companies in different countries and the exchange rates of several currencies to the US dollar.

2.7.1 Application 1: the Stock Prices of Oil Companies

This case considers the daily stock prices of oil companies in China (CN), France (FR), Germany (DE), Italy (IT), Japan (JP), the United States (US), and the United Kingdom (UK) from 2016 to 2020. The number of oil companies in each market (CN, FR, DE, IT, JP, US, and the UK) is 5, 4, 3, 2, 6, 25, 14, respectively. The data was first processed by taking logarithm and differencing transformations, and second discretized by converting all positive results to 1 and non-positive results to 0. To incorporate the latent factor in the SIMPLE method, we discarded 14 companies in the UK market, and then modeled the graph structure of the remaining data. The number of samples $n$ is equal to 1000 and the number of variables $p$ is 45.

Figure 2.5 compares graph structures using the SIMPLE method and its com-
The Node-wise-BOTH method, which demonstrated a good performance in the simulation studies in Section 2.6. In Figure 2.5, graph (a) on the top shows the graph structure of the stock prices using the SIMPLE method, which considered companies in the UK market as latent factors, and selected suitable values of regularization parameters $\lambda_1$, $\lambda_2$ with the same number of latent variables. Graph (b) on the bottom shows the graph structure using a standard sparse graphical model selection method: namely, the Node-wise approach with the “BOTH” edge criteria. It does not take latent variables into account, and the penalty parameter was selected via cross-validation. In Figure 2.5, each node in the graph denotes an oil company, and a gray edge between two nodes indicates there are interactions between the corresponded two companies. The nodes with the same color in graphs (a) and (b) belong to the same market. The red, yellow, green, light blue, dark blue and the purple nodes represent companies in China, France, Germany, Italy, Japan and the US, respectively.

Comparing these two graphs in Figure 2.5, it is apparent that graph (a) is less dense than graph (b), which means the SIMPLE method generates a more sparse graph than the Node-wise-BOTH approach. To further clarify the relationships within each graph, we have Figure 2.6 where each market is plotted like a circle. The edges within each circle represent the relationships between the oil companies in the same market. The edges among different circles show the relationships between different markets. Similar to Figure 2.5, graph (a) in Figure 2.6 shows the graph structure in a circular layout via the SIMPLE method, and graph (b) in Figure 2.6 shows the graph structure using the Node-wise-BOTH method. To enable a clear comparison, the nodes with the same color in the four graphs, i.e., graph (a) and (b) in Figure 2.5 and graph (a) and (b) in Figure 2.6, represent the oil companies belong to the same market, and the nodes with the same number in these four graphs represent the same oil company. For example, the purple node with label 31 is the thirty-first company in our list and it is in the US market.

To make a comprehensive comparison between the results of the SIMPLE method and the Node-wise-BOTH approach, we compared them from the following three aspects:

1. The “in-market” level, i.e., the relationships within the same market. In Figure 2.6, these are the edges within a circle with the same color.

2. The “cross-market” level, i.e., the relationships between different markets. In Figure 2.6, these are the edges between different circles with different colors.

3. The “US-others” level, i.e., the relationships between the US market and the
non-US markets. In Figure 2.6, these are the edges between the purple circle and the circles with other colors. 

In Figure 2.6, we can see that our SIMPLE method produces a sparser graph than the Node-wise-BOTH approach in all the above three aspects. To further show their differences, Figure 2.7 is provided which shows the sparsity patterns of these two methods. Similar to Figures 2.5 and 2.6, graph (a), on the top of Figure 2.7, shows the sparsity pattern using the SIMPLE method, and graph (b), on the bottom of Figure 2.7, shows the sparsity pattern obtained from the Node-wise-BOTH method. The markets along the x-axis (from top to bottom) and y-axis (from left to right) are China, France, Germany, Italy, Japan, and the US. Their colors, which coincide with Figures 2.5 and 2.6, are red, yellow, green, light blue, dark blue, and purple, respectively.

A dark square in the sparsity pattern denotes an edge, while a blank means no edge. Therefore in Figure 2.7, “in-market” interactions can be observed via dark squares close to the diagonal with the same colored labels in the x-axis and y-axis. For instance, “in-market” interactions in China can be obtained from the upper left corner in Figure 2.7 (a) and (b) whose x-axis and y-axis labels are red “CN1” to “CN5”. “Cross-market” interactions can be obtained similarly via dark squares off the diagonal with different colored labels in the x-axis and y-axis. For instance, interactions between markets in France and Japan can be observed from dark squares in Figure 2.7 (a) and (b) whose labels in the x-axis are “FR6” to “FR9” in orange, and in the y-axis are “JP15” to “JP20” in dark blue, or vice versa due to the symmetry of the sparsity pattern. The interactions between the US and non-US markets can be obtained from the upper right or the lower left parts in Figure 2.7 (a) and (b) whose label colors in the x-axis are purple, and in the y-axis are other colors, or vice versa. For example, we can observe the interactions between the markets in the US and Germany from the dark squares in Figure 2.7 (a) and (b) whose labels in the x-axis are “US21” to “US45” in purple, and in the y-axis are “DE10” to “DE12” in green, or vice versa.

Extensive studies show that the changes in the stock price of a company are mainly determined by internal determinants of a firm. Buying and selling stocks by investors causes stock price movements, and investors actions are decided on available information they have about a company, like its book value of equity, dividend yield, the growth rate of the company, operating cash flows, total assets, earnings per share, capital structure and book value per share (Collins, 1957; Gordon, 1959; Allen and Rachim, 1996; Malhotra and Prakash, 2001; Irfan et al., 2002; Al-Tamimi et al., 2011; Sanju et al., 2011; Tandon and Malhotra, 2013; Singh, 2018).
Apart from internal determinants, macroeconomic factors in the same market have also been found to have an impact on stock price fluctuations. Although macroeconomic determinants may vary in different markets, existing studies demonstrate that common macroeconomic factors affecting stock prices include the growth rate of the GDP, the inflation rate, the unemployment rate, government policies, and interest rates (Campbell, 1987; Campbell and Ammer, 1993; Tsoukalas, 2003; Joseph and Vezos, 2006; Kurihara, 2006b; Ratanapakorn and Sharma, 2007; Baharumshah et al., 2009; Mondal and Imran, 2010).

In addition to domestic determinants, external factors, such as oil prices and exchange rate fluctuations, have also been noticed to influence stock prices (Sadorsky, 1999; Tsoukalas, 2003; Kim, 2003; Joseph and Vezos, 2006; Kurihara, 2006a; Ratanapakorn and Sharma, 2007; Park and Ratti, 2008; Kilian and Park, 2009; Zhao, 2010; Taiwo et al., 2012; Tsai, 2012). Besides, the US stock market has been observed to have spillover effects on other markets (Kim and Rogers, 1995; Liu and Pan, 1997; Xiao and Dhesi, 2010; Tsai, 2014; Sui and Sun, 2016; Liu et al., 2017).

From Figures 2.5, 2.6 and 2.7, it can be seen that for the Node-wise-BOTH method which ignores the latent variables, the resulting dense graphs cannot deliver the above information since nearly all nodes are connected. However, the graphs corresponding to the SIMPLE method coincide with the above results. There are some interactions between the stocks in the same market due to macroeconomic factors, which can be easily observed from the dark squares close to the diagonal of the sparsity pattern in Figure 2.7 (a). Aside from these, there are also some interactions between different markets due to global factors and spillover effects from the US stock market. However, cross-market connections are sparser than in-market connections, since, as mentioned before, stock price fluctuations are mainly based on domestic information. This can be observed from the off-diagonal dark squares in Figure 2.7 (a). Based on these results, we can determine that when latent variables are present, our SIMPLE method is more accurate in capturing the structural information.
Figure 2.5: Graph (a) and (b) show graph structures of stock prices using the SIMPLE method and the Node-wise-BOTH method.
Figure 2.6: Graph (a) and (b) show the circular layouts of graph structures using the SIMPLE method and the Node-wise-BOTH method.
Figure 2.7: Figure (a) and (b) show the sparsity patterns of stock prices using the SIMPLE method and the Node-wise-BOTH method.
2.7.2 Application 2: the Exchange Rates to the US Dollar

The second application considers the daily exchange rates of several currencies to the US dollar in six continents from 2015 to 2020. Similar to Application 1, the data was first processed by taking logarithm and differencing transformations, and second discretized by converting all positive results to 1 and non-positive results to 0. The total number of samples \( n \) is equal to 1304, and the number of currencies \( p \) is 47. As follows the first three capital letters of a country are currency abbreviations and the names of the countries or regions that these currencies belong to are shown in brackets.

- Asia: KWD (Kuwait), THB (Thailand), ILS (Israel), LKR (Sri Lanka), LBP (Lebanon), IDR (Indonesia), INR (India), JOD (Jordan), JPY (Japan), PKR (Pakistan), RUB (Russia), PHP (Philippines), KRW (South Korea), AED (United Arab Emirates), HKD (Hong Kong), CNY (China), MYR (Malaysia), SGD (Singapore)

- South America: COP (Colombia), PEN (Peru), CLP (Chile), BRL (Brazil), TTD (Trinidad and Tobago), ARS (Argentina)

- North America: HNL (Honduras), CAD (Canada), JMD (Jamaica), MXN (Mexico)

- Africa: ZAR (South Africa), GHS (Ghana), DZD (Algeria), EGP (Egypt)

- Oceania: FJD (Fiji), NZD (New Zealand), AUD (Australia)

- Europe: ISK (Iceland), BGN (Bulgaria), RON (Romania), TRY (Turkey), SEK (Sweden), NOK (Norway), CZK (Czech Republic), HUF (Hungary), PLN (Poland), GBP (United Kingdom), EUR (European Union), DKK (Denmark)

Unlike in Application 1 where we manually removed the UK market and treated them as latent variables, this case considers a variety of exchange rates to the US dollar. Thus the US dollar itself and the information related to the US economy are naturally the latent factors; for instance, the GDP growth rate, the unemployment rate, the financial index of the US and so on.

Figure 2.8 shows the resulting graph structures of the exchange rates to the US dollar using the SIMPLE and the Node-wise-BOTH methods. Figure 2.8 (a), on the top, was obtained by the SIMPLE method which considers the US dollar and the US economy as latent factors, and suitable regularization parameters were selected with \( k = 6 \) latent variables. Figure 2.8 (b), on the bottom, was obtained
via the Node-wise-BOTH approach without introducing any latent variables, and its penalty parameters were selected by cross-validation. In Figure 2.8, a single node in the graph represents one currency, and an edge between two nodes indicates interaction between the exchange rates of these two currencies to the US dollar. Moreover, the nodes with the same color in Figure 2.8 (a) and (b) belong to the same continent. The red, yellow, purple, green, pink and blue nodes represent the currencies in Asia, South America, North America, Africa, Oceania and Europe, respectively.

Comparing graphs (a) and (b) in Figure 2.8, we can see that graph (a) displays less edges than graph (b). This means, similar to the first application in the last section, our SIMPLE method which considers latent factors produces a sparser graph than the Node-wise-BOTH method. To be close to real-world systems, in Figure 2.9 each node is located in its geographic position. The currencies that originate from the same continent are marked with the same color, and these colors are consistent with Figure 2.8.

Similar to Application 1, to further reveal the relationships in each graph, Figure 2.10 shows the sparsity patterns obtained from these two methods. Figure 2.10 (a) on the top and (b) on the bottom display the sparsity patterns using the SIMPLE and the Node-wise-BOTH methods, respectively. The continents along the $x$-axis (from top to bottom) and the $y$-axis (from left to right) are Asia, South America, North America, Africa, Oceania, and Europe. Their colors, which are the same as in Figures 2.8 and 2.9, are red, yellow, purple, green, pink, and blue, respectively.

By the properties of the sparsity pattern, (i.e., a dark square indicates an edge and a blank means no edge), interactions between currencies can be easily obtained by checking their names in the $x$-axis and $y$-axis. For example, the interaction between DKK and EUR can be observed by finding DKK in the $x$-axis and EUR in the $y$-axis, or vice versa due to the symmetry of sparsity pattern. The interactions between currencies in the same continent can also be observed via the dark squares close to the diagonal with the same colored labels in the $x$-axis and $y$-axis. For instance, the interactions of currencies in Asia can be obtained from the upper left parts in Figure 2.10 (a) and (b), whose labels in the $x$-axis and $y$-axis are red.

The currency exchange rate to the US dollar, one of the most important indexes of economic health at a country’s level, is among the most watched, analyzed and governmentally manipulated economic measures. Due to the uniqueness of the US dollar, which is the de facto global currency dominating foreign exchange reserves in most central banks in the world, the currency exchange rate to the US dollar plays
a vital role in country-to-country levels of trade.

The currency exchange rate to the US dollar is, of course, influenced by the value of the US dollar. As prices of any other products, the value of a currency is determined by demand and supply. If the demand of the US dollar increases, then its value would appreciate, and thus the value of another currency against it would depreciate. Previous studies have showed that the performance of the US economy is at the heart of investors’ decisions to buy or sell dollars. If investors worry about the future of the US economy, they tend to sell dollars which leads its value to depreciate. In contrast, on the one hand, a strong economy attracts investors from all over the world who seek high and safe yields. Such increasing investments, especially from abroad, create a strong demand for dollars which results in a higher value. On the other hand, US business becomes relatively more competitive which increases the demand for US goods and would cause the value of the dollar appreciates (PCE, 2012; Labonte, 2015; Jackson, 2019; Lioudis, 2019). In addition to the confidence of the US economy, there are other factors that affect the value of the US dollar, such as monetary policy, the sentiment which includes foreign policies and relationships, and the market psychology (Alton, 2015; Gardner, 2017; Mcguinness, 2019; Evrensel, 2020).

Another determinant of the exchange rate to the US dollar is the behavior of a local currency. If this currency appreciates, then the value of it against the US dollar increases. Existing literature has showed that there are a variety of reasons which affect currency value. Common determinants are inflation rates, interest rates, current account deficits, public debts, and economic growth/recession (Edwards, 1988; Faruqee, 1995; Maeso-Fernandez et al., 2002; Pettinger, 2008; Twin, 2020). In addition to economic factors, government policies have been found to play a more important role. Countries may adopt different exchange rate regimes which depend on how much currency flexibility they want. Sharp exchange rate fluctuations often cause financial crises, such as the financial crisis in Japan during the 1990s and the Asian financial crisis in 1997. To avoid such disastrous consequences, many countries have managed to implement floating exchange rate regimes, i.e., they allow their exchange rates to respond to markets conditionally and intervene in them occasionally to limit the changes (Dornbusch, 1976; Krugman, 1979; Obstfeld, 1984; Chang and Velasco, 1998; Radelet et al., 1998; Calvo and Reinhart, 2002; Hayashi and Prescott, 2002).

Aside from the domestic determinants of the local currency and the US dollar, some external factors, such as spillover effects from other countries, may also affect the exchange rate. This has been noticed in places with deep and com-
plex economic relationships, such as Europe and Asia (Wyplosz, 2004; Égert and Morales-Zumaquero, 2008; Schnabl, 2009).

Figures 2.8, 2.9, and 2.10 show the results obtained from the Node-wise-BOTH method which does not take the latent variables into account are contradictory to the findings in literature discussed above. When countries collaborate in the international trade, their governments ensure their currencies remain relatively stable to safeguard the strong competitiveness of their products and services; especially, to the US dollar since most financial transactions and international trade agreements are made with it (Ito et al., 1998; Zucchi, 2019; Amadeo, 2020). Therefore the dense graphs produced by the Node-wise-BOTH method which imply that currencies are highly connected with each other, do not match current global markets. However, the graphs corresponding to the SIMPLE method represent the aforementioned information. These sparse graphs show the stabilities and independences of these currencies to the US dollar. Besides, graphs produced by the SIMPLE method also capture the relatively close relationships within Europe and Asia, as in the upper left and lower right corners of Figure 2.10 (a). Moreover, among all 47 currencies, four of them are pegged to the US dollar, which means their exchange rates to the US dollar are fixed. They are LBP (Lebanon), JOD (Jordan), HKD (Hong Kong), and AED (United Arab Emirates). The results obtained from the SIMPLE method deliver this important information. In Figure 2.10 (a), there is no connection between these four currencies with the rest, which demonstrates the independence of these pegged currencies. However, the results obtained from the Node-wise-BOTH method contradict this information. Overall, these results indicate that the SIMPLE method performs better when latent variables exist.
Figure 2.8: Graph (a) and (b) show the graph structures of exchange rates using the SIMPLE method and the Node-wise-BOTH method.
Figure 2.9: Graph (a) and (b) show the maps of graph structures of exchange rates using the SIMPLE method and the Node-wise-BOTH method.
Figure 2.10: Figure (a) and (b) show the sparsity patterns of exchange rates using the SIMPLE method and the Node-wise-BOTH method.
Chapter 3

Decorrelated Lasso and Interaction Screening

3.1 Introduction

We propose a new method, called DECO-IS, to solve the variable selection problem of the linear model with two-way interactions in the high-dimensional setting. We design a two-stage approach that can identify important variables and maintain a strong hierarchical structure simultaneously. Theoretical results about the consistent estimation of parameters in the model are provided. Numerical results show the good performance of this new method.

3.1.1 Our Setup

Consider the usual regression situation where we have a data set that contains \( n \) observations, \( p \) predictors, and the response to those predictors. To model the response in terms of predictors, one may consider a widely used linear model which assumes that the predictors and the response are linearly related, that is

\[
Y = X\beta + \varepsilon, \tag{3.1.1}
\]

where \( Y \in \mathbb{R}^n \) is the response vector, \( X \in \mathbb{R}^{n\times p} \) is the design matrix, \( \beta \in \mathbb{R}^p \) is the parameter vector, and \( \varepsilon \) is the error term.

However in practice, predictors often work together, and thus the linear model is not enough to capture the relationship between the response and the predictors. In such a situation, two or higher-order interaction terms are often of interest to researchers because of their crucial roles in model prediction and inter-
interpretation. In general, a linear model with two-way interaction terms is expressed as

\[ Y = \beta_1 X_1 + \ldots + \beta_p X_p + \beta_{1,2} X_1 X_2 + \beta_{1,3} X_1 X_3 + \ldots + \beta_{p-1,p} X_{p-1} X_p + \varepsilon, \]

(3.1.2)

where \( X_1, \ldots, X_p \) are main effects, \( X_j X_k \) \((j \neq k)\) is the two-way interaction term and it is the entry-wise product of \( X_j \) and \( X_k \), \( \beta_i \) is the parameter of main effect \( X_i \), \( \beta_{j,k} \) is the parameter of interaction term \( X_j X_k \), \( Y \) is the response, and \( \varepsilon \) is the error term.

The model (3.1.2) extends the linear model by considering interaction effects between predictors can improve prediction ability when predictors work together. The aim of this study is to discover which terms, especially which interaction terms, have important effects on the response in the high-dimensional setting where the number of parameters is larger than the sample size. In other words, the goal is to select a subset of the main effects and interaction terms that is predictive to the response and estimate the associated parameters.

However, this is a nontrivial task. There are two fundamental challenges here. First, for a large \( p \) which is comparable to or much larger than \( n \), there are a total number of \((p^2 + p)/2\) parameters in the model with interactions, which is enormous and computationally challenging. For example, if \( n = 100 \) and \( p = 500 \), then the total number of interaction terms is 124750, which cannot be handled easily. Second, certain hierarchy principles are required to be maintained to ensure model interpretability (Nelder, 1977; Peixoto, 1987; Hamada and Wu, 1992; Chipman, 1996), such as the “strong” (“weak”) hierarchy restrictions; that is, an interaction is included in the model if both (at least one) of the associated main effects are included. This means extra effort is needed during the selection process.

The first challenge can be solved by variable selection methods. To deal with large and complex datasets, variables selections become a central problem in the past several years. In the last two decades, changes in modern science and technology have generated complex datasets featuring a large number of variables compared to sample size, which is the so-called “high-dimensional setting”. Since classical statistical inference methods are often infeasible in such a case, statisticians propose applying a “sparsity” assumption on the underlying coefficients. That is, only a selection of the parameters in the model are nonzero, which means only a small number of the variables affect the response. The aim of sparse high-dimensional regression is to select some of the most relevant variables for the response and to discard the rest. For this high-dimensional linear regression, many approaches have been proposed.
based on a sparsity assumption via the principle of regularization, with the aim of estimating the support of the regression coefficient and its value simultaneously, including Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), group Lasso (Yuan and Lin, 2006), adaptive Lasso (Zou, 2006), and Dantzig selector (Candes et al., 2007).

The second problem can be addressed via a two-stage method: at Stage 1, only selecting the main effects; at Stage 2, selecting the interactions of the main effects identified during Stage 1. Such a two-stage procedure ensures hierarchy conditions to be satisfied. It is worth mentioning that the two-stage method is able to maintain hierarchy structure, and is, possibly, the only feasible approach for interaction selection when the data dimension is large (Wu et al., 2009, 2010; Hao and Zhang, 2014). Interaction selection has attracted much attention in recent years. Related studies include Yuan et al. (2007, 2009), Choi et al. (2010), Radchenko and James (2010), and Bien et al. (2013). These approaches, described as joint analysis where main effects and interactions are amalgamated and a global search is performed, become infeasible when the number of parameters is large due to high computational cost.

Based on the information above, we propose the following new method, called DECO-IS, for interaction selection:

- Stage 1: intentionally leave the interactions out, and only estimate the main effects via the Decorrelated Lasso (DECO-Lasso) method.

- Stage 2: estimates the associated interaction terms of the main effects selected at Stage 1 by the DECO-Lasso method.

The DECO-Lasso method first appeared in Wang and Leng (2016). As the name suggests, its main idea is to first decorrelate the design matrix of the linear model (3.1.1), and then apply the Lasso (Tibshirani, 1996) to the transformed data. The Lasso, which is able to estimate parameters and select variables in the linear model simultaneously via introducing an \( \ell_1 \) penalty, is one of the most important and popular techniques to solve high-dimensional estimation problems. However, it requires a stringent “Irrepresentable Condition” (Zhao and Yu, 2006) that needs weak variable dependence to achieve the “sign consistency”, i.e., the Lasso selects the correct set of true parameters asymptotically. This means highly correlated columns in the design matrix can cause severe detrimental effects to the Lasso estimators. What is more, if the columns in the design matrix are correlated in a way that does not satisfy the Irrepresentable Condition, then the Lasso estimator will not be sign consistent. However, such a “Irrepresentable Condition” is not easy to be verified.
and satisfied in reality. The DECO-Lasso was proposed to avoid this condition. It achieves this by first decorrelating the design matrix of the linear model by column. After that, the transformed columns in the new design matrix are orthogonal, where the “Irrepresentable Condition” holds naturally.

3.1.2 Related Work

1. Jia and Rohe (2012)

In a linear model, to conform to the irrepresentable condition when columns in the design matrix are correlated in a way that not satisfy the irrepresentable condition, Jia and Rohe (2012) proposed a preconditioning method. The main idea is (1) pre-multiplying the term $UD^{-1}U^T$ on both sides of the linear model to obtain the transformed data $(UD^{-1}UY, UD^{-1}UX)$, where matrices $U$, $D$ come from the singular value decomposition $X = UDVT$, and (2) applying the Lasso to the transformed data. Details are given as follows.

Consider a linear model
\[ Y = X\beta + \varepsilon, \quad (3.1.3) \]
where $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix, $\beta \in \mathbb{R}^p$ is the parameter vector, and $\varepsilon$ is the error term.

Pre-multiplying the term $UD^{-1}U^T$ on both sides of (3.1.3), we have
\[
UD^{-1}UY = UD^{-1}UX\beta + UD^{-1}U^T\varepsilon, \\
= UV^T\beta + UD^{-1}U^T\varepsilon.
\]

The Lasso estimator based on the transformed data $(UD^{-1}UY, UV^T)$ is
\[
\hat{\beta}_1 = \arg \min_{\beta} \left\{ \|UD^{-1}UY - UV^T\beta\|_2^2 + \lambda_1\|\beta\|_1 \right\}.
\]

Denoting the transformed response vector as $\tilde{Y} = UD^{-1}UY$, the transformed design matrix as $\tilde{X} = UV^T$, Jia and Rohe (2012) demonstrated that the columns of $\tilde{X} = UV^T$ were exactly orthogonal when $n > p$, then they satisfied the “Irrepresentable Condition” and $\hat{\beta}_1$ had sign consistency. When $n < p$, they showed the columns in the transformed design matrix were also more amenable than the original design matrix to this condition.

2. Wang and Leng (2016)
A similar idea was proposed by Wang and Leng (2016) to solve the problem of feature space partition. In the high-dimensional setting, fitting statistical models is computationally challenging. An attractive approach is to first divide the dataset into subsets and then fit statistical algorithms to these subsets. The dataset can be partitioned either horizontally (in the sample space) or vertically (in the feature space). In the high-dimensional setting, the feature space partition is more effective. However, the feature space partitioning is challenging because of the correlations between features. To solve this problem, Wang and Leng (2016) proposed a new method called DECO. The main idea is to first decorrelate the variables such that the transformed variables are roughly orthogonal to each other, then partition the transformed feature space into \( q \) subsets, and finally apply estimation methods to each subset.

In detail, the decorrelation step of DECO in the linear model is to pre-multiply the term \( \sqrt{pUD^{-1}U^T} \) on both sides of model (3.1.3) which is similar to Jia and Rohe (2012). The transformed model is

\[
\sqrt{pUD^{-1}U^T}Y = \sqrt{pUD^{-1}U^T}X\beta + \sqrt{pUD^{-1}U^T}\epsilon, \\
= \sqrt{pUV}T\beta + \sqrt{pUD^{-1}U^T}\epsilon.
\]

Applying the Lasso to the transformed model, the DECO-Lasso estimator based on the transformed data \( (\sqrt{pUD^{-1}U^T}Y, \sqrt{pUV}T) \) is

\[
\hat{\beta}_{DECO-Lasso} = \arg\min_{\beta} \left\{ \|\sqrt{pUD^{-1}U^T}Y - \sqrt{pUV}T\beta\|_2^2 + \lambda_2\|\beta\|_1 \right\}.
\]

Wang and Leng (2016) demonstrated that under some mild conditions, the estimator \( \hat{\beta}_{DECO-Lasso} \) is sign consistent in both low-dimensional and high-dimensional settings.

### 3.1.3 Contributions

Interaction selection in the high-dimensional setting has drawn much attention in recent years. Many studies have been done to obtain reliable estimation results. Among them, most methods can be described as joint analysis, i.e., they consider the main and interaction effects altogether, and then conduct a global search. These methods are effective in identifying important main and interaction effects when the number of parameters \( p \) is moderate. However, they become infeasible if \( p \) is large due to an enormous number of interaction terms in this situation.
Our contributions are threefold. First, different from previous approaches, we propose a new two-stage method that not only selects main and interaction effects efficiently, even when the number of parameters of main effects $p$ is considerably large, but also can maintain the hierarchy structure of interaction terms. Second, we provide theoretical results of the consistent estimation of main effects. Third, simulation studies show the value of our new method.

3.2 The DECO-IS Method

The main purpose of this study is to propose a new method called DECO-IS which estimates the following linear model with two-way interactions model in the high-dimensional setting

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_{1,2} X_1 X_2 + \ldots + \beta_{p-1,p} X_{p-1} X_p + \epsilon \quad (3.2.1)$$

where $X_1, \ldots, X_p$ are main effects, $X_j X_k$ $(j \neq k)$ is the two-way interaction term and it is the entry-wise product of $X_j$ and $X_k$, $\beta_i$ is the parameter of main effect $X_i$, $\beta_{j,k}$ is the parameter of interaction term $X_j X_k$, $Y$ is the response, and $\epsilon$ is the error term.

An important consideration when it comes to estimating interaction effects is to apply certain hierarchical principles. In this study, we focus on the “strong” hierarchy principle, i.e., the coefficient of $X_j X_k$ is nonzero unless both coefficients of $X_j$ and $X_k$ are nonzeros.

In general, it is not an easy task to identify important main and interaction effects when the number of parameters is relatively large compared with the sample size due to high computational requirements. Moreover, maintaining the hierarchy principle introduces extra difficulties to the estimation process. In order to solve this problem, we propose a two-stage method that only selects the main effects at Stage 1, and then selects associated interactions at Stage 2.

The following result also indicates the utilization of a two-stage method. Model (3.2.1) can be written as the following simplified way

$$Y = X \beta^{(M)} + Z \beta^{(I)} + \epsilon, \quad (3.2.2)$$

where $Y \in R^n$ is the response vector, $X \in R^{n \times p}$ is a design matrix for the main effects, $\beta^{(M)}$ is the parameter of main effects, $Z \in R^{n \times q}$, where $q = \frac{p(p-1)}{2}$, is the design matrix for the two-way interaction terms, $\beta^{(I)}$ is the parameter of interactions, and $\epsilon \in R^n$ is the error term.
Assume that $E(X_j) = 0$, and $X$ follows a multivariate normal distribution $N(0, \Sigma)$, Hao et al. (2018) observed that the joint covariance matrix for the main effects and all the interactions admits a block diagonal structure as

$$
\begin{pmatrix}
\Sigma^{(M)} & 0 \\
0 & \Sigma^{(I)}
\end{pmatrix},
$$

where $M$ and $I$ indicate main effects and interactions respectively.

The discussion above motivates the following new method, called DECO-IS, for interaction selection:

- **Stage 1:** treating the interaction term and the original error term as a new noise, then model (3.2.2) can be written as

  $$
  Y = X\beta^{(M)} + Z\beta^{(I)} + \varepsilon \\
  = X\beta^{(M)} + W
  $$

  (3.2.3)

  where $Y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is a design matrix for the main effects, and $W$ is the new noise.

  Pre-multiplying the term $UD^{-1}U^T$ on both sides of the model (3.2.3) where matrices $U$ and $D$ come from the singular value decomposition $X = UDV^T$, that is

  $$
  UD^{-1}U^TY = UD^{-1}U^TX\beta^{(M)} + UD^{-1}U^TW \\
  = UV^T\beta^{(M)} + UD^{-1}U^TW.
  $$

  (3.2.4)

  Denoting the transformed response vector as $\tilde{Y} = UD^{-1}U^TY$, the transformed design matrix as $\tilde{X} = UV^T$, and the transformed error term as $\tilde{W} = UD^{-1}U^TW$, then (3.2.4) can be written as

  $$
  \tilde{Y} = UV^T\beta^{(M)} + \tilde{W} \\
  = \tilde{X}\beta^{(M)} + \tilde{W}. 
  $$

  (3.2.5)

  Applying the Lasso to the transformed data $(\tilde{Y}, \tilde{X})$ to select the main effect, the DECO-IS estimator at Stage 1 is

  $$
  \hat{\beta}^{(M)} = \arg \min_{\beta^{(M)}} \left\{ (\tilde{Y} - \tilde{X}\beta^{(M)})^T(\tilde{Y} - \tilde{X}\beta^{(M)}) + \lambda||\beta^{(M)}||_1 \right\},
  $$

  (3.2.6)

  where $||\beta^{(M)}||_1 = \sum_{j=1}^{p} |\beta_j^{(M)}|$.
Stage 2: denoting the estimated parameters of main effects at Stage 1 as \( \hat{\beta}^{(M)} \), we have the following model for two-way interactions only

\[
Y - X \hat{\beta}^{(M)} = Z \beta^{(I)} + \varepsilon.
\]  
(3.2.7)

Pre-multiplying the term \( U_Z D_Z^{-1} U_Z^T \) on both sides of (3.2.7) where matrices \( U_Z \) and \( D_Z \) come from the singular value decomposition \( Z = U_Z D_Z V_Z^T \), that is

\[
U_Z D_Z^{-1} U_Z^T (Y - X \hat{\beta}^{(M)}) = U_Z D_Z^{-1} U_Z^T Z \beta^{(I)} + U_Z D_Z^{-1} U_Z^T \varepsilon
= U_Z V_Z \beta^{(I)} + U_Z D_Z^{-1} U_Z^T \varepsilon
= \tilde{Z} \beta^{(I)} + U_Z D_Z^{-1} U_Z^T \varepsilon.
\]  
(3.2.8)

Denoting \( \tilde{Y}_I = U_Z D_Z^{-1} U_Z^T (Y - X \hat{\beta}^{(M)}) \) and \( \tilde{Z} = U_Z D_Z^{-1} U_Z^T Z \), similar to Stage 1, the DECO-IS estimator at Stage 2 is

\[
\hat{\beta}^{(I)} = \arg \min_{\beta^{(I)}} \left\{ (\tilde{Y}_I - \tilde{Z} \beta^{(I)})^T (\tilde{Y}_I - \tilde{Z} \beta^{(I)}) + \lambda \| \beta^{(I)} \|_1 \right\},
\]  
(3.2.9)

where \( \| \beta^{(I)} \|_1 = \sum_{j=1}^{p} |\beta_j^{(I)}| \).

### 3.3 Theoretical Analysis

In this section, we provide the consistency result of the estimator (3.2.6) of the DECO-IS method at Stage 1. Define \( Q = \{1, 2, ..., p\} \) and let \( A^C \) be \( Q \setminus A \) for any set \( A \subseteq Q \). The following Theorem 5 shows the sign consistency result for the estimator (3.2.6) of DECO-IS at Stage 1.

**Theorem 5.** Consider the interaction model

\[
Y = X \beta^{(M)} + Z \beta^{(I)} + \varepsilon
= X \beta^{(M)} + Y_I + \varepsilon \text{ (denote } Y_I = Z \beta^{(I)} \text{)}
\]  
(3.3.1)

where \( Y \in \mathbb{R}^n \) is the response vector, \( X \in \mathbb{R}^{n \times p} \) is a design matrix for the main effects, \( \beta^{(M)} \) is the parameter of main effects, \( Z \in \mathbb{R}^{n \times q} \) is the design matrix for the two-way interaction terms, \( \beta^{(I)} \) is the parameter of interactions, and \( \varepsilon \in \mathbb{R}^n \) is the error term.

Assume that \( X_i \) (the \( i \)-th row of the design matrix \( X \)), \( 1 \leq i \leq n \), is independent and identically distributed from \( N(0, \Sigma) \), \( \varepsilon \sim N(0, \sigma^2 I) \) is independent of
\{X_i\}_{i=1}^n. When \( n > p \), if \( \lambda > \frac{9\|\beta(M)^*\|_2\sqrt{\log p}}{C_{\text{min}}n^{\frac{3}{4}}} + 20\sqrt{\frac{\sigma^2}{n \min_{i} n_i}} \) and \( M(\beta(M)^*) > 2\lambda \), then the probability that there exists a DECO-IS estimator at Stage 1, i.e.,

\[
\beta^{(M)} = \arg\min_{\beta(M)} \left\{ (\hat{Y} - \hat{X}\beta^{(M)})^T(\hat{Y} - \hat{X}\beta^{(M)}) + \lambda\|\beta^{(M)}\|_1 \right\},
\]

which satisfies \( \text{sign}(\hat{\beta}^{(M)}) = \text{sign}(\beta(M)^*) \), is at least \( 1 - cp^{-\alpha} \). Here \( c \) and \( \alpha \) are positive constants, \( \beta(M)^* \) is the true parameter of the interaction term, \( \beta(I)^* \) is the true parameter of the main effects term, \( M(\beta(M)^*) = \min_{j \in S}\|\beta_j^{(M)^*}\|_1 \), and \( C_{\text{min}} \) denotes the minimal eigenvalue of \( X^TX \).

**Proof.** The full proof of Theorem 5 is given in Appendix B.1. Here we outline the proof strategy. The proof is based on the following Lemma 6 which is first given in Wainwright (2006). It provides necessary and sufficient conditions for the sign consistency of the \( \ell_1\)-regularized estimator in the linear model. Hence for the DECO-IS estimator (3.3.2) which is based on the transformed data \( \hat{Y} \) and \( \hat{X} \), we only need to verify these conditions stated in Lemma 6 to obtain the consistency result.

**Lemma 6.** (Lemma 1 in Wainwright (2006)) For linear model \( Y = X\beta^* + \epsilon \), assume that the matrix \( X(S)^TX(S) \) is invertible, where \( S \) is the support set of \( \beta^* \) and and \( X(S) \) is the sub-matrix of \( X \) with index \( S \). Then for any given \( \lambda > 0 \) and any noise term \( \epsilon \in \mathbb{R}^n \), there exists a Lasso estimate \( \hat{\beta}(\lambda) \)

\[
\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\},
\]

which satisfies \( \text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*) \), if and only if the following two conditions hold

(a) \(|X(S^C)^TX(S)(X(S)^TX(S))^{-1}[X(S)^T\epsilon - \lambda\text{sign}(\beta^*(S))] - X(S^C)^T\epsilon| \leq \lambda \).

(b) \( \text{sign}(\beta^*(S) + (X(S)^TX(S))^{-1}[X(S)^T\epsilon - \lambda\text{sign}(\beta^*(S))]) = \text{sign}(\beta^*(S)) \).

where the vector inequality and equality are taken elementwise.

Wainwright (2006) provided the following sufficient conditions for (a) and (b). Define

\[
V_j = X_j^T \left\{ X(S)(X(S)^TX(S))^{-1}\lambda\text{sign}(\beta^*(S)) - [X(S)(X(S)^TX(S))^{-1}X(S)^T - I] \epsilon \right\},
\]

and

\[
U_i = e_i^T (X(S)^TX(S))^{-1}[X(S)^T\epsilon - \lambda\text{sign}(\beta^*(S))],
\]

76
where \( e_i \) is the vector with 1 in the \( i \)th position and zeros elsewhere. By simple calculations, the condition (a) holds strictly if and only if the following strict dual feasibility condition holds

\[
\mathcal{M}(V) = \left\{ \max_{j \in SC} |V_j| < \lambda \right\}, \tag{3.3.3}
\]

and the following sign consistency condition is sufficient to guarantee that condition (b) holds

\[
\mathcal{M}(U) = \left\{ \max_{i \in S} |U_i| < M(\beta^*) \right\}, \tag{3.3.4}
\]

where \( M(\beta^*) = \min_{j \in S} |\beta_j^*| \), and \( S = \left\{ j : \beta_j^* \neq 0 \right\} \). Thus our goal is to check conditions (3.3.3) and (3.3.4) on the transformed data \((\tilde{Y}, \tilde{X})\) in (3.2.5) when \( n > p \).

(Part I) Verifying the strict dual feasibility condition. The interaction model (3.2.5) has the transformed response vector \( \tilde{Y} \) and the transformed design matrix \( \tilde{X} \), therefore the strict dual feasibility condition in (3.3.3) takes the following form

\[
\mathcal{M}(\tilde{V}) = \left\{ \max_{j \in SC} |\tilde{V}_{\tilde{j}}| < \lambda \right\} \tag{3.3.5}
\]

where

\[
\tilde{V}_{\tilde{j}} = \tilde{X}_j^T \left\{ \tilde{X}(S)(\tilde{X}(S)^T\tilde{X}(S))^{-1}\lambda \text{sign}(\beta(M)^*(S)) - \left[ \tilde{X}(S)(\tilde{X}(S)^T\tilde{X}(S))^{-1}\tilde{X}(S)^T - I \right] \tilde{W} \right\}
\]

Above condition (3.3.5) can be verified via the following Lemma 7. The proof details are in Appendix B.1.

Lemma 7. Under the interaction model (3.3.1), assume that \( X_i \) (the \( i \)th row of design matrix \( X \)), \( 1 \leq i \leq n \) are independent and identically distributed from \( N(0, \Sigma) \), \( \varepsilon \sim N(0, \sigma^2 I) \) is independent of \( \{X_i\}_{i=1}^n \), if \( \lambda > \frac{9 \|\beta(I)^*\|_2 \sqrt{\log p}}{C_{\min} n^{\frac{1}{2}}} + 20 \sqrt{\frac{\sigma^2 \log p}{C_{\min} n}} \), we have

\[
P\left( \max_{j \in SC} |\tilde{V}_j| > \lambda \right) \leq c p^{-\alpha}, \tag{3.3.6}
\]

where \( \beta(M)^* \) is the true parameter of the main effects term, \( \beta(I)^* \) is the true parameter of the interaction term, \( C_{\min} \) denotes the minimal eigenvalue of \( X^T X \), \( c \) and \( \alpha \) are positive constants.

(Part II) Verify the sign consistency condition. Based on the transformed data
(\hat{Y}, \hat{X})$, the sign consistency condition in (3.3.4) has the following form:

$$\mathcal{M}(\hat{U}) = \left\{ \max_{i \in S} |\hat{U}_i| < M(\beta^{(M)*}) \right\}, \tag{3.3.7}$$

where

$$\hat{U}_i = e_i^T \left( \tilde{X}(S)^T \tilde{X}(S) \right)^{-1} \left[ \tilde{X}(S)^T \tilde{W} - \lambda \text{sign}(\beta^{(M)*}(S)) \right].$$

Above condition (3.3.7) can be verified via the following Lemma 8 and the proof details are in Appendix B.1.

**Lemma 8.** Under the same setting of Lemma 7, if

$$M(\beta^{(M)*}) > \frac{9\|\beta^{(I)*}\|_2 \sqrt{\log P}}{C_{\text{min}} n^{\frac{3}{2}}} + 20\sqrt{\frac{\sigma^2 \log P}{C_{\text{min}} n}} + \lambda,$$

which means

$$M(\beta^{(M)*}) > 2\lambda,$$

then we have

$$P \left( \max_{i \in S} |\hat{U}_i| < M(\beta^{(M)*}) \right) \leq cp^{-\alpha}, \tag{3.3.8}$$

Where $\beta^{(M)*}$, $\beta^{(I)*}$, $C_{\text{min}}$, $c$ and $\alpha$ are the same as defined in Lemma 7.

Theorem 7 states that the DECO-IS method selects the main effects consistently with high probability under mild conditions. It is important to notice that in Theorem 7, there is no the irrepresentable condition (Zhao and Yu, 2006) which is hard to be satisfied in reality. It is worth mentioning that in Theorem 7, the interaction term $Y_I$ and the original error term $\varepsilon$ are treated together as a new model noise $W$ as shown in (3.2.3). However, since the interaction term $Y_I$ is correlated with $X$, the new noise term $W$ is correlated with $X$ as well. This may cause the issue of biasness of the estimator $\hat{\beta}^{(M)}$ and the failure of the DECO-IS method.

### 3.4 Simulation Study

In this section, we report the conducted simulation studies which show the performance of the DECO-IS method and its competitor with respect to the variable selection of the linear model with two-way interactions.

We used the GLMNET package in R to carry out the simulations for the interaction model and compute the Lasso and DECO-IS estimator separately. The tuning parameter was chosen by the BIC. In each dataset, the following two-stage method was used to calculate the DECO-IS estimator and its competitor: the Lasso estimator.

Stage 1:
1.1) generate one training dataset and one test dataset with equal size.

1.2) in the training dataset, compute the Lasso estimator

\[ \hat{\beta}^{(M)} (\text{lasso}) = \arg \min_{\beta^{(M)}} \left\{ \| Y - X \beta^{(M)} \|_2^2 + \lambda \| \beta^{(M)} \|_1 \right\}, \]

and the DECO-IS estimator

\[ \hat{\beta}^{(M)} (\text{DECO - IS}) = \arg \min_{\beta^{(M)}} \left\{ \| \tilde{Y} - UV^T \beta^{(M)} \|_2^2 + \lambda \| \beta^{(M)} \|_1 \right\}. \]

separately.

1.3) in the test set, calculate the BIC values for the models using the Lasso and the DECO-IS respectively.

1.4) choose the tuning parameter corresponding to the smallest BIC value for each method.

1.5) obtain the estimators \( \hat{\beta}^{(M)}_{\text{Lasso}} \) and \( \hat{\beta}^{(M)}_{\text{DECO - IS}} \) corresponding to the selected tuning parameters, and obtain the estimated main effects sets \( \hat{S}^{(M)}_{\text{Lasso}} \) and \( \hat{S}^{(M)}_{\text{DECO - IS}} \) for both methods, where \( \hat{S}^{(M)} = \{ j : \hat{\beta}^{(M)}_j \neq 0 \} \).

Stage 2:

2.1) generate one training dataset and one test dataset with equal size,

2.2) in the training dataset, compute the Lasso estimator

\[ \hat{\beta}^{(I)} (\text{Lasso}) = \arg \min_{\beta^{(I)}} \left\{ \| Y_I - X \beta^{(I)} \|_2^2 + \lambda \| \beta^{(I)} \|_1 \right\}, \]

and the DECO-IS estimator

\[ \hat{\beta}^{(I)} (\text{DECO - IS}) = \arg \min_{\beta^{(I)}} \left\{ \| \tilde{Y}_I - U_Z V_Z^T \beta^{(I)} \|_2^2 + \lambda \| \beta^{(I)} \|_1 \right\}. \]

Here \( Y_I = Y - X \hat{\beta}^{(M)}_{\text{Lasso}}, \quad \tilde{Y}_I = U_Z D_Z^{-1} U_Z^T (Y - X \hat{\beta}^{(M)}_{\text{DECO - IS}}). \) The matrices \( U_Z, D_Z \) come from \( Z \hat{S}^{(M)}_{\text{DECO - IS}} = U_Z D_Z V_Z^T, \) where \( Z \hat{S}^{(M)}_{\text{DECO - IS}} \) is the sub-matrix of \( Z \) with index \( \hat{S}^{(M)}_{\text{DECO - IS}}. \)

2.3) in the test set, calculate the BIC values for the models using the Lasso and the DECO-IS respectively.
2.4) choose the tuning parameter corresponding to the smallest BIC value for each method.

2.5) obtain the interaction estimators $\hat{\beta}^{(I)}_{\text{Lasso}}$ and $\hat{\beta}^{(I)}_{\text{DECO-IS}}$ corresponding to the selected tuning parameters, and obtain the estimated interaction effects sets $\hat{S}^{(I)}_{\text{Lasso}}$ and $\hat{S}^{(I)}_{\text{DECO-IS}}$ for both methods, where $\hat{S}^{(I)} = \{j : \hat{\beta}^{(I)}_j \neq 0\}$.

In each setting, 100 i.i.d datasets were generated, and we recorded the frequency of choosing the true model, i.e., the frequency of the estimated parameter set is the same as the true support set. The results in the following Example 1 and 2 show the good performance of the DECO-IS method.

**Example 1. (Independent Predictors; Example 1 in Hao and Zhang (2014))** Let $p=500$, $X_i$ ($i = 1,...,n$) are iid from $\text{MVN}(0, I_p)$, the noise $\varepsilon_i$ is iid generated from $N(0, \sigma^2)$. The true underlying regression coefficients $\beta^{(M)} = (3, 0, 3, 0, 0, 3, 0, 0, 0, 3, 0)$, so $S^{(M)} = \{1, 3, 6, 10\}$. The interaction set $S^{(I)} = \{(1, 3), (1, 6), (3, 10), (6, 10)\}$ with coefficient 2. The following frequency table 3.1 shows the results of stage 1 in example 1 with different $\sigma = 1, 2, 3$.

<table>
<thead>
<tr>
<th>method</th>
<th>n=75</th>
<th>n=100</th>
<th>n=125</th>
<th>n=150</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma=1$</td>
<td>Lasso</td>
<td>61</td>
<td>64</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>DECO-Lasso</td>
<td>91</td>
<td>97</td>
<td>99</td>
</tr>
<tr>
<td>$\sigma=2$</td>
<td>Lasso</td>
<td>66</td>
<td>62</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>DECO-Lasso</td>
<td>79</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>$\sigma=3$</td>
<td>Lasso</td>
<td>71</td>
<td>68</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>DECO-Lasso</td>
<td>75</td>
<td>74</td>
<td>75</td>
</tr>
</tbody>
</table>

**Example 2. (Autoregressive Correlation; Example 2 in Hao and Zhang (2014))** Consider the same setup as Example 1, except that $X_i$ ($i = 1,...,n$) follows Multivariate Normal Distribution with zero mean and $\text{Cov}(x_j, x_k) = 0.5^{|j-k|}$ for $1 \leq j, k \leq p$. The following frequency table 3.2 shows results of stage 1 in example 2 with different $\sigma = 1, 2, 3$. 

Table 3.1: The frequency (%) of choosing the true model in Example 1

Table 3.2: The frequency (%) of choosing the true model in Example 2
Table 3.2: The frequency (%) of choosing the true model in Example 2

<table>
<thead>
<tr>
<th>method</th>
<th>n=75</th>
<th>n=100</th>
<th>n=125</th>
<th>n=150</th>
</tr>
</thead>
<tbody>
<tr>
<td>σ=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>48</td>
<td>29</td>
<td>39</td>
<td>37</td>
</tr>
<tr>
<td>DECO-Lasso</td>
<td>84</td>
<td>92</td>
<td>89</td>
<td>92</td>
</tr>
<tr>
<td>σ=2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>49</td>
<td>32</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>DECO-Lasso</td>
<td>70</td>
<td>73</td>
<td>71</td>
<td>74</td>
</tr>
<tr>
<td>σ=3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>60</td>
<td>35</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>DECO-Lasso</td>
<td>60</td>
<td>59</td>
<td>57</td>
<td>63</td>
</tr>
</tbody>
</table>
Chapter 4

Conclusion

In the first part of this thesis, we studied the problem of the graphical modeling of binary random variables with a few latent components. As we know, compared with the continuous case, the discrete Markov random field is more difficult to handle due to the complex form of its normalization constant. Moreover, the existence of latent variables introduces extra difficulties. However, to solve this problem we proposed a relatively simple method which employs a pseudo-likelihood function with $\ell_1$ and nuclear norms. Such a convex “sparse plus low rank” optimization problem can be efficiently solved by the proposed randomized block coordinate descent algorithm. Acceleration techniques have been provided, as well as a convergence analysis. We also showed theoretical guarantees of the global error and described conditions under which the binary graphical model with latent variables is identifiable given samples of observed variables only. A set of experiments on synthetic data showed superior performances of our model regardless of the importance of latent variables varies. Two real-world data applications, including the stock prices of oil companies across several markets and the exchange rates of different currencies to the US dollar, demonstrated the value and accuracy of our method.

In the second part of this thesis, we focused on the estimation problem of a two-way interaction model in the high-dimensional setting. The task of interaction selection is both computationally and theoretically challenging. To ensure the reasonability and interpretability of the model structure, its interaction terms need to satisfy certain logical relationships, such as a “strong” hierarchy, i.e., the interaction term exists only if both of its associated main effects presented, or the “weak” hierarchy, i.e., the interaction term exists if at least one of its associated main effects presented. On the other hand, potential correlations between main effects and interaction terms pose difficulties in the procedure of variable selection, especially
in the high-dimensional setting. In this study, we proposed a new method, called DECO-IS, to solve this problem. It is a two-stage procedure which selects main effects at Stage 1 and then select associated interactions at Stage 2. It is computationally cheaper and easier to implement than wrapping all terms together. At each stage, we decorrelated the variables to satisfy the stringent “Irrepresentable Condition” when adopting an \( \ell_1 \) penalty. Experiments on synthetic data showed the advantages of our new method. Theoretical results were also provided which guarantee the sign consistency about parameters of main effects.

Several questions arose that are worthy of further investigation. First, the proposed convex function of the \( \ell_1 \) and nuclear norms regularized pseudo-likelihood can be extended to different graphical models with latent variables, such as graphical models with mixed data types and quantile graphical models. Second, it is of interest to develop an efficient one stage method to further estimate the interaction model.
Appendix A

Proofs in Chapter 2

A.1 Proof of Lemma 1

Proof. The Cauchy–Schwarz inequality
\[ u^T v \leq \|u\|_2 \|v\|_2 \]
implies that
\[
(\nabla f(x) - \nabla f(y))^T (x - y) \leq \|\nabla f(x) - \nabla f(y)\|_2 \|x - y\|_2 \\
\leq M \|x - y\|_2^2
\]

Consider the function \( g(t) = f(x + t(y - x)) \), we have
\[
g'(t) - g'(0) = (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) \\
\leq tM \|y - x\|_2^2
\]

Since \( f(y) = g(1) \) and \( f(x) = g(0) \), it implies
\[
f(y) - f(x) = g(1) - g(0) \\
= \int_0^1 g'(t) dt \\
= \int_0^1 (\nabla f(x + t(y - x))^T (y - x)) dt \\
\leq \int_0^1 (tM \|y - x\|_2^2 + \nabla f(x)^T (y - x)) dt \\
= \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2
\]
A.2 Proof of Lemma 2

Proof. The optimality condition for the minimization problem (2.3.11) is

\[ 0 \in \frac{1}{t} (z - x) + \lambda_1 \partial \|z\|_1 \quad (A.2.1) \]

where \( \partial \|z\|_1 \) is the set of subgradients of the \( \ell_1 \) norm. The \( \ell_1 \) norm is separable, thus we can consider its components separately. First, if \( z_i = 0 \), then \( \partial \|z_i\|_1 = \text{sign}(z_i) \), and the optimum \( z_i^* \) is obtained as

\[ 0 = \frac{1}{t} (z_i - x_i) + \lambda_1 \text{sign}(z_i) \Leftrightarrow z_i^* = x_i - \lambda_1 t \text{sign}(z_i^*) \quad (A.2.2) \]

Note that if \( z_i^* < 0 \), then \( x_i < -\lambda_1 t \) and if \( z_i^* > 0 \), then \( x_i > \lambda_1 t \).

Second, if \( z_i = 0 \), the subdifferential of the \( \ell_1 \) norm is the interval \([-1, 1]\).

Then the optimality condition is

\[ 0 \in -\frac{1}{t} x_i + \lambda_1 [-1, 1], \quad (A.2.3) \]

which is equivalent to

\[ x_i \in [-\lambda_1 t, \lambda_1 t] \quad (A.2.4) \]

■

A.3 Proof of Lemma 3

Proof. Since the target function is strictly convex, there exists a unique solution. We thus need to prove it is equal to \( D_{\lambda_2}(X) \). The optimality condition for the minimizer \( Z^* \) of the problem (2.3.12) is

\[ 0 \in \frac{1}{t} (Z^* - X) + \lambda_2 \partial \|Z^*\|_* \quad (A.3.1) \]

where \( \partial \|Z\|_* \) is the set of subgradients of the nuclear norm. Let \( Z \in \mathbb{R}^{n_1 \times n_2} \), its subgradient has the following form (Watson, 1992)

\[ \partial \|Z\|_* = \{ U_Z V_Z^T + W_Z : W_Z \in \mathbb{R}^{n_1 \times n_2}, U_Z^T W_Z = 0, W_Z V_Z = 0, \|W_Z\|_2 \leq 1 \} \]
Now set $Z^* = \mathcal{D}_{\lambda_2 t}(X)$, we want to show that $Z^*$ satisfies the condition (A.3.1). The SVD of $X$ can be written as

$$X = U_0 D_0 V_0^T + U_1 D_1 V_1^T$$

where $U_0$, $V_0$ (resp. $U_1$, $V_1$) are the singular vectors associated with singular values greater than $\lambda_2 t$ (resp. smaller or equal to $\lambda_2 t$). With this notation, we have

$$Z^* = U_0 (D_0 - \lambda_2 t I) V_0^T$$

and therefore,

$$X - Z^* = \lambda_2 t (U_0 V_0^T + W), \quad W = \frac{1}{\lambda_2 t} U_1 D_1 V_1^T$$

By definition, we have $U_0^T W = 0$, $W V_0 = 0$. Since the diagonal elements of $D_1$ are bounded by $\lambda_2 t$, we also have $\|W\|_2 \leq 1$. Hence $X - Z^* \in \lambda_2 t \partial \|X^*\|_*$, which completes the proof.

A.4 Proof of Lemma 4

Proof. From (2.3.14), we have

$$F(x) - F(x^*) \geq F(x) - Q(x^*, y). \quad (A.4.1)$$

Now, since $f$, $g$ are convex, we have

$$f(x) \geq f(y) + \langle x - y, \nabla f(y) \rangle, \quad (A.4.2)$$

and

$$g(x) \geq g(x^*) + \langle x - x^*, \partial g(x^*) \rangle, \quad (A.4.3)$$

where $\partial g(\cdot)$ denotes the subgradient of $g(\cdot)$. Summing the above inequalities (A.4.2) and (A.4.3) yields

$$F(x) = f(x) + g(x) \geq f(y) + \langle x - y, \nabla f(y) \rangle + g(x^*) + \langle x - x^*, \partial g(x^*) \rangle \quad (A.4.4)$$

On the other hand, by the definition of $x^*$ and (2.3.14), we have

$$Q(x^*, y) = f(y) + \langle x^* - y, \nabla f(y) \rangle + \frac{\mu}{2} \|x^* - y\|^2 + g(x^*). \quad (A.4.5)$$

86
Therefore, plugging (A.4.4) and (A.4.5) into (A.4.1), it follows that

\[
F(x) - F(x^*) \geq -\frac{\mu}{2} \|x^* - y\|^2 + \langle x - x^*, \nabla f(y) + \partial g(x^*) \rangle \\
= -\frac{\mu}{2} \|x^* - y\|^2 + \mu(x - x^*, y - x^*) \\
= -\frac{\mu}{2} \|x^* - y\|^2 + \mu(y - x, x^* - y),
\]

where the first equality above is based on the fact \(\nabla f(y)\). 

\[ \blacksquare \]

### A.5 Proof of Corollary 1

**Proof.** If \(F(x^{[k_0]}) = F(\bar{x})\) at some \(k_0\), then \(x^{[k]} = x^{[k_0]}\) for all \(k \geq k_0\), since \(F(x^{[k]})\) is strictly less than \(F(x^{[k-1]})\) as long as \(x^{[k]} \neq x^{[k-1]}\). If \(F(x^{[k_0]}) > F(\bar{x})\) for all \(k \geq 0\), thus Lemma 5 indicates that \(x^{[k]}\) converges to some critical point \(x^*\) if \(x^{[0]}\) is sufficiently close to \(\bar{x}\), where \(x^{[0]}, \bar{x}, x^* \in \mathcal{B}\). If \(F(x^*) > F(\bar{x})\), then the KL inequality implies \(\phi'(F(x^*) - F(\bar{x})) \text{dist}(0, \partial F(x^*)) \geq 1\), which is impossible since \(0 \in \partial F(x^*)\).

\[ \blacksquare \]

### A.6 Proof of Theorem 1

**Proof.** Since \(F(x^{[k]})\) is monotonically nonincreasing, if \(F(x^{[k_0]}) = F(\bar{x})\) at some \(k_0\), then \(x^{[k]} = x^{[k_0]} = \bar{x}\) for all \(k \geq k_0\). Now consider \(F(x^{[k]}) > F(\bar{x})\) for all \(k \geq 0\). Since \(\bar{x}\) is a limit point and \(F(x^k) \rightarrow F(\bar{x})\), there must exist an integer \(k_0\) such that \(x^{[k_0]}\) is sufficiently close to \(\bar{x}\) as required in Lemma 5. Hence the entire sequence \(x^{[k]}\) converges. Since \(\bar{x}\) is a limit point of \(\{x^{[k]}\}\), we have \(x^{[k]}\) converges to \(\bar{x}\).

\[ \blacksquare \]

### A.7 Proof of Theorem 3

**Proof.** Denote \((\beta^*, \Theta^*, L^*)\) as the statistical truths, and \((\hat{\beta}, \hat{\Theta}, \hat{L})\) as the minimizer of (2.2.7), we have the following first-order Taylor expansion of \(f(\beta, \Theta, L)\):

\[
D_f(\hat{\beta}, \hat{\Theta}, \hat{L}; \beta^*, \Theta^*, L^*) = f(\hat{\beta}, \hat{\Theta}, \hat{L}) - f(\beta^*, \Theta^*, L^*) - \langle \nabla_{\beta} f(\beta^*, \Theta^*, L^*), \hat{\beta} - \beta^* \rangle \\
- \langle \nabla_{\Theta} f(\beta^*, \Theta^*, L^*), \hat{\Theta} - \Theta^* \rangle - \langle \nabla_{L} f(\beta^*, \Theta^*, L^*), \hat{L} - L^* \rangle,
\]

(A.7.1)

where \(D_f\) is the abbreviation of Bregman divergence \(D_f(\hat{\beta}, \hat{\Theta}, \hat{L}; \beta^*, \Theta^*, L^*)\).
Since \((\hat{\beta}, \hat{\Theta}, \hat{L})\) is the minimizer of (2.2.7), we have
\[
f(\hat{\beta}, \hat{\Theta}, \hat{L}) + \lambda_1\|\hat{\Theta}\|_1 + \lambda_2\|\hat{L}\|_* \leq f(\beta^*, \Theta^*, L^*) + \lambda_1\|\Theta^*\|_1 + \lambda_2\|L^*\|_*,
\]
which means
\[
f(\hat{\beta}, \hat{\Theta}, \hat{L}) - f(\beta^*, \Theta^*, L^*) \leq \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \lambda_2(\|L^*\|_* - \|\hat{L}\|_*). \tag{A.7.2}
\]
Plugging the equation (A.7.1) into (A.7.2), we have
\[
D_f + \langle \nabla_\beta f(\beta^*, \Theta^*, L^*), \hat{\beta} - \beta^* \rangle + \langle \nabla_\Theta f(\beta^*, \Theta^*, L^*), \hat{\Theta} - \Theta^* \rangle + \langle \nabla_L f(\beta^*, \Theta^*, L^*), \hat{L} - L^* \rangle \\
\leq \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \lambda_2(\|L^*\|_* - \|\hat{L}\|_*). \tag{A.7.3}
\]
Define the effective noises \(E_\beta\), \(E_\Theta\) and \(E_L\) as
\[
G_\beta = \nabla_\beta f(\beta^*, \Theta^*, L^*), \\
G_\Theta = \nabla_\Theta f(\beta^*, \Theta^*, L^*), \\
G_L = \nabla_L f(\beta^*, \Theta^*, L^*),
\]
then on the events \(\{\|G_\beta\|_* \leq \omega_0\}\), \(\{\|G_\Theta\|_* \leq \omega_1\}\) and \(\{\|G_L\|_2 \leq \omega_2\}\), using the cauchy-schwarz inequality for the inner product, (A.7.3) can be written as
\[
D_f \leq \|\nabla_\beta f(\beta^*, \Theta^*, L^*)\|_* \|\beta^* - \hat{\beta}\|_1 + \|\nabla_\Theta f(\beta^*, \Theta^*, L^*)\|_* \|\Theta^* - \hat{\Theta}\|_1 \\
+ \|\nabla_L f(\beta^*, \Theta^*, L^*)\|_2 \|L^* - \hat{L}\|_* + \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \lambda_2(\|L^*\|_* - \|\hat{L}\|_*), \\
= \|G_\beta\|_* \|\beta^* - \hat{\beta}\|_1 + \|G_\Theta\|_* \|\Theta^* - \hat{\Theta}\|_1 + \|G_L\|_2 \|L^* - \hat{L}\|_* \\
+ \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \lambda_2(\|L^*\|_* - \|\hat{L}\|_*), \\
\leq \lambda_1(\|\Theta^*\|_1 - \|\hat{\Theta}\|_1) + \omega_1\|\Theta^* - \hat{\Theta}\|_1 + \lambda_2(\|L^*\|_* - \|\hat{L}\|_* + \omega_2\|L^* - \hat{L}\|_* + \omega_0\|\beta^* - \hat{\beta}\|_1. \tag{A.7.4}
\]
(1) First we focus on the \(\ell_1\) norm of \(\Theta^*\) and \(\hat{\Theta}\) in (A.7.4). Define the support of \(\Theta^*\) by \(S = \{i, j : \Theta^*_{i,j} \neq 0\}\) and the cardinality \(s = |S|\), then we have
\[
\|\Theta^*_S - \hat{\Theta}_S\|_1 \geq \|\Theta^*_S\|_1 - \|\hat{\Theta}_S\|_1,
\]
and
\[
\|\hat{\Theta}\|_1 = \|\hat{\Theta}_S\|_1 + \|\hat{\Theta}_{S^c}\|_1 \geq \|\Theta^*_S\|_1 - \|\Theta^*_S - \hat{\Theta}_S\|_1 + \|\hat{\Theta}_{S^c}\|_1,
\]
88
which means
\[ \| \Theta^* \|_1 - \| \hat{\Theta} \|_1 \leq \| \Theta^*_S - \hat{\Theta}_S \|_1 - \| \hat{\Theta}_{SC} \|_1. \]

Then the $\ell_1$ norm of $\Theta^*$ and $\hat{\Theta}$ in (A.7.4) has the following property:
\[
\lambda_1(\| \Theta^* \|_1 - \| \hat{\Theta} \|_1) + \omega_1(\| \Theta^* - \hat{\Theta} \|_1) \leq \lambda_1(\| \Theta^*_S - \hat{\Theta}_S \|_1 - \| \hat{\Theta}_{SC} \|_1) + \omega_1(\| \Theta^*_S - \hat{\Theta}_S \|_1)
\leq (\omega_1 + \lambda_1)(\| \Theta^*_S - \hat{\Theta}_S \|_1 + (\omega_1 - \lambda_1)(\| \hat{\Theta}_{SC} \|_1)
\]
(A.7.5)

(2) Second we consider the nuclear norm of $L^*$ and $\hat{L}$. Assume that the target matrix $L^*$ has rank $r$, we define two subspaces of matrix
\[
M = \{ A \in \mathbb{R}^{n \times p} \mid \text{rowspan}(A) \subseteq V, \text{colspan}(A) \subseteq U \},
\]
and
\[
\bar{M}^\perp = \{ B \in \mathbb{R}^{n \times p} \mid \text{rowspan}(B) \subseteq V^\perp, \text{colspan}(B) \subseteq U^\perp \}.
\]
Here $V$ and $U$ are spaces spanned by the top $r$ left and right singular vectors of $L^*$ respectively. $V^\perp$ and $U^\perp$ denote the subspaces orthogonal to $V$ and $U$ respectively.

For an arbitrary pair of matrices $A \in M$ and $B \in \bar{M}^\perp$, the nuclear norm is decomposable, i.e.,
\[
\|A + B\|_* = \|A\|_* + \|B\|_*.
\]

By this important property, we have
\[
\| \hat{L} \|_* = \| L^* + \hat{L} - L^* \|_* = \| L^*_M + (\hat{L} - L^*)_{\bar{M}^\perp} + (\hat{L} - L^*)_{\bar{M}^\perp} \|_*
\geq \| L^*_M + (\hat{L} - L^*)_{\bar{M}^\perp} \|_* - \| (\hat{L} - L^*)_{\bar{M}^\perp} \|_*
\leq \| L^*_M \|_* + \| (\hat{L} - L^*)_{\bar{M}^\perp} \|_* - \| (\hat{L} - L^*)_{\bar{M}^\perp} \|_*
\]
which means
\[
\| L^* \|_* - \| \hat{L} \|_* \leq \| (L^* - \hat{L})_{\bar{M}^\perp} \|_* - \| (L^* - \hat{L})_{\bar{M}^\perp} \|_*.
\]

Then the nuclear norm of $L^*$ and $\hat{L}$ in (A.7.4) has the following property:
\[
\lambda_2(\| L^* \|_* - \| \hat{L} \|_*) + \omega_2(\| L^* - \hat{L} \|_*) \leq \lambda_2(\| (L^* - \hat{L})_{\bar{M}^\perp} \|_* - \| (L^* - \hat{L})_{\bar{M}^\perp} \|_*
\leq (\omega_2 + \lambda_2)(\| (L^* - \hat{L})_{\bar{M}^\perp} \|_* + (\omega_2 - \lambda_2)(\| (L^* - \hat{L})_{\bar{M}^\perp} \|_*
\]
(A.7.6)
Combing results (A.7.5) and (A.7.6), and using the regularity condition

\[ K \left\{ (\vartheta_1^2 + \vartheta_2^2 + \vartheta_3^2)(s + r)D_f \right\}^{1/2} \geq (1 + \vartheta_1)(\|\Theta^* - \Theta\|_1 - \|\Theta_{SC}\|_1
+ (1 + \vartheta_2)(\|L^* - L\|_{\overline{M}} \|_\ast - \|L^* - L\|_{\overline{M}} \|_\ast
+ \vartheta_3\|\beta^* - \beta\|_1
\]

then the function \( D_f \) in (A.7.4) has the following property

\[ D_f \leq (\omega_1 + \lambda_1)\|\Theta^*_S - \hat{\Theta}_S\|_1 + (\omega_1 - \lambda_1)\|\hat{\Theta}_{SC}\|_1 + (\omega_2 + \lambda_2)(\|L^* - \hat{L}\|_{\overline{M}} \|_\ast
+ (\omega_2 - \lambda_2)(\|L^* - \hat{L}\|_{\overline{M}} \|_\ast + \omega_0\|\beta^* - \hat{\beta}\|_1
\]

\[ = (1 + 2\omega_1)\|\Theta^*_S - \hat{\Theta}_S\|_1 - \|\Theta_{SC}\|_1 + (1 + 2\omega_2)(\|L^* - \hat{L}\|_{\overline{M}} \|_\ast
- \|L^* - \hat{L}\|_{\overline{M}} \|_\ast + \omega_0\|\beta^* - \hat{\beta}\|_1
\]

\[ \leq K \{ (\omega_0^2 + 4\omega_1^2 + 4\omega_2^2)(s + r)D_f \}^{1/2}
\]

where we set \( \lambda_1 = \omega_1 + 1 \) and \( \lambda_2 = \omega_2 + 1 \). It means

\[ D_f \leq K^2(\omega_0^2 + 4\omega_1^2 + 4\omega_2^2)(s + r).
\]

\[ \square \]

### A.8 Proof of Corollary 2

**Proof.** To prove Corollary 2, first we define the sub-gaussian random variable, the sub-exponential random variable and the orlicz norm.

**Definition 4.** A centered random variable \( X \in \mathbb{R} \) is said to be a sub-gaussian random variable \( X \sim \text{subG}(\sigma^2) \) with parameter \( \sigma > 0 \), if its moment generating function satisfies

\[ \mathbb{E}[\exp(tX)] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right). \]

A basic property of \( X \sim \text{subG}(\sigma^2) \) is that \( P(|X| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) \) for any \( t > 0 \).

**Definition 5.** A centered random variable \( X \in \mathbb{R} \) is said to be a sub-exponential random variable \( X \sim \text{subE}(\nu^2, b) \) with parameter \( \nu, b > 0 \), if its moment generating function satisfies

\[ \mathbb{E}[\exp(tX)] \leq \exp\left(\frac{\nu^2 t^2}{2}\right), \forall t : |t| < \frac{1}{b} \]

A basic property of \( X \sim \text{subE}(\nu^2, b) \) is that \( P(|X| > t) \leq 2\exp\left(-\frac{t^2}{2\nu^2}\right) \) for any...
0 \leq t \leq \frac{\nu^2}{b}.

**Definition 6.** Let \( \psi \) be a nondecreasing, convex function with \( \psi(0) = 0 \), and \( X \) is a random variable, then the orlicz norm \( \|X\|_\psi \) is defined as

\[
\|X\|_\psi = \inf \left\{ C > 0 : \mathbb{E}\psi\left(\frac{|X|}{C}\right) \leq 1 \right\}.
\]

The orlicz norm has the following properties:

1. Let \( \psi_p(x) = e^{xp} - 1 \), \( p \geq 1 \),

   (a) \( p = 2 \): \( X \) is a sub-gaussian random variable is equivalent to \( \|X\|_{\psi_2} < \infty \), and

   \[
   X \sim \text{subG}(\sigma^2) \Rightarrow \|X\|_{\psi_2} \lesssim \sigma
   \]

   where \( \lesssim \) means the left side is bounded by a positive constant times the right side.

   (b) \( p = 1 \): \( X \) is a sub-exponential random variable is equivalent to \( \|X\|_{\psi_1} < \infty \), and

   \[
   X \sim \text{subE}(\nu^2, b) \Rightarrow \|X\|_{\psi_1} \lesssim \nu \lor b
   \]

   where \( \lor \) means the maximum.

2. Any orlicz norm can be used to obtain an estimate of the tail of a distribution.

By Markov’s inequality, we have

\[
P(|X| > t) \leq P\left(\psi\left(\frac{|X|}{\|X\|_\psi}\right) \geq \psi\left(\frac{t}{\|X\|_\psi}\right)\right) \leq \frac{1 + c}{\psi\left(\frac{t}{\|X\|_\psi}\right) + c} \tag{A.8.1}
\]

Recall the partial derivatives in our problem are

\[
G(\beta) = \nabla_\beta f(\beta^*, \Theta^*, L^*) = -X^T 1_n + \left[ \frac{\exp(1_n \beta^* T + X \Theta^* + L^*)}{1 + \exp(1_n \beta^* T + X \Theta^* + L^*)} \right]^T 1_n,
\]

\[
G(\Theta) = \nabla_\Theta f(\beta^*, \Theta^*, L^*) = -X^T X + X^T \frac{\exp(1_n \beta^* T + X \Theta^* + L^*)}{1 + \exp(1_n \beta^* T + X \Theta^* + L^*)},
\]

\[
G(L) = \nabla_L f(\beta^*, \Theta^*, L^*) = -X + \left[ \frac{\exp(1_n \beta^* T + X \Theta^* + L^*)}{1 + \exp(1_n \beta^* T + X \Theta^* + L^*)} \right].
\]
Here the matrix $X$ has row independence, each entry in matrix $X$ takes value 0 or 1, and the sigmoid function $f(x) = \frac{e^x}{1+e^x} \in (0, 1)$.

(1) For the vector $G(\beta)$, we have its $j$th component

$$G(\beta)_j = \left[ \nabla_{\beta} f(\beta^*, \Theta^*, L^*) \right]_j = \sum_{i=1}^n \left[ -X + \frac{\exp(1_n\beta^*T + XS^* + L^*)}{1 + \exp(1_n\beta^*T + XS^* + L^*)} \right]_{ij}.$$ 

Since $\left[ -X + \frac{\exp(1_n\beta^*T + XS^* + L^*)}{1 + \exp(1_n\beta^*T + XS^* + L^*)} \right]_{ij}$ is bounded in $[-1, 1]$, its centered form is a subG(4). By the property of row independence, we have $G(\beta)_j \sim \text{subG}(4n)$, which means $\|G(\beta)_j\|_{\psi_2} \lesssim \sqrt{n}$. Then we have

$$P\left[\|G(\beta)\|_\infty \geq \omega_0\right] = P\left[\max_{1 \leq j \leq p} |G(\beta)_j| \geq \omega_0\right] \leq \sum_{1 \leq j \leq p} P[|G(\beta)_j| \geq \omega_0], \text{ (by the union bound)}$$

$$\leq \sum_{1 \leq j \leq p} \frac{2}{\psi_2\left(\frac{\omega_0}{\|G(\beta)_j\|_{\psi_2}}\right) + 1}, \text{ (by inequality (A.8.1) and take } c = 1$$

$$\leq 2p \exp\left(-\left(\frac{\omega_0}{\|G(\beta)_j\|_{\psi_2}}\right)^2\right), \text{ (by } \psi_2(x) = \exp(x^2) - 1)$$

$$\leq 2 \exp\left(\log p - \frac{\omega_0^2}{c_1n}\right). \text{ (since } \|G(\beta)_j\|_{\psi_2} \lesssim \sqrt{n})$$

Let $\omega_0^2 = c_1(1 + \alpha)n \log p$, we have

$$P\left[\|G(\beta)\|_\infty \geq \omega_0\right] \leq 2p^{-\alpha}, \quad (A.8.2)$$

where $c_1, \alpha$ are positive constants.

(2) For the matrix $G(\Theta)$, we can re-write it as

$$G(\Theta) = -X^T \left( X + \frac{\exp(1_N\beta^*T + X\Theta^* + L^*)}{1 + \exp(1_N\beta^*T + X\Theta^* + L^*)} \right) = -X^T Z.$$ 

Then the $(i, j)$th component of $G(\Theta)$ can be written as

$$G(\Theta)_{ij} = \sum_{k=1}^n -X_{ki}Z_{kj}.$$ 

Since $-X_{ki}$ is bounded in $[-1, 0]$ and $Z_{kj}$ is bounded in $[0, 2]$, we can know that
\(-X_{ki}Z_{kj}\) is bounded and \(\| -X_{ki}Z_{kj}\|\psi_2 < \infty\). This means \(-X_{ki}Z_{kj}\) is a sub-gaussian random variable, denoted as \(\text{subG}(b)\) where \(b\) is a positive constant. By the row independence of \(X\), we have

\[
G(\Theta)_{ij} = \sum_{k=1}^{n} -X_{ki}Z_{kj} \sim \text{subG}(nb),
\]

which means

\[
\|G(\Theta)_{ij}\|\psi_2 \lesssim \sqrt{n}.
\]

Then we have

\[
P\left[\|G(\Theta)\|_\infty \geq \omega_1\right] = P\left[\max_{1 \leq i,j \leq p} |G(\Theta)_{ij}| \geq \omega_1\right] \\
\leq \sum_{1 \leq i,j \leq p} P\left[|G(\Theta)_{ij}| \geq \omega_1\right] \\
\leq \sum_{1 \leq i,j \leq p} \frac{2}{\psi_2\left(\|G(\Theta)_{ij}\|_\psi_2\right) + 1} \\
\leq 2p^2\exp\left(-\frac{\omega_1^2}{\|G(\Theta)_{ij}\|_\psi_2^2}\right) \\
\leq 2p^2\exp(2\log p - \frac{\omega_1^2}{c_2n}).
\]

Let \(\omega_1^2 = c_2(2 + \alpha)n\log p\), we have

\[
P\left[\|G(\Theta)\|_\infty \geq \omega_1\right] \leq 2p^{-\alpha}. \tag{A.8.3}
\]

(3) For the matrix \(G(L)\), we denote it as \(M\)

\[
G(L) = -X + \left[\frac{\exp(1N\beta^*T + X\Theta^* + L^*)}{1 + \exp(1N\beta^*T + X\Theta^* + L^*)}\right] = M.
\]

Here we focus on the operator norm of matrix \(\|G(L)\|_2 = \|M\|_2\). From the definition of the operator norm, we can view the upper tail \(\|M\|_2 > t\) as a union of many simpler events:

\[
P(\|M\|_2 > t) \leq P(\cup_{u \in S}\|Mu\| > t),
\]

where \(S := \{u \in \mathbb{R}^p : \|u\| = 1\}\) is the unit sphere in \(\mathbb{R}^p\). Hence the first step we need to obtain the tail bound of \(P(\|Mu\| > t)\).

(Step 1) Let \(M_1, ..., M_n\) be the \(n\) rows of \(M\), then the the column vector \(Mu\) has coefficients \(M_iu = \sum_{j=1}^{p} M_{ij}u_j\) for \(i = 1, ..., n\). Since \(M_{ij}\) is a sub-Gaussian...
random variable and \( \|u\|_2 = 1 \), then \( M_i u \) is also a sub-Gaussian variable. Thus we have
\[
E \exp(b|M_i u|^2) \leq 2,
\]
for constant \( b > 0 \). By the row independence, multiplying above inequalities for all \( i \), we obtain
\[
E \exp(b\|Mu\|^2) \leq 2^n.
\]
Then the Markov property gives us
\[
P(\|Mu\| \geq \omega_2) = P(\exp(b\|Mu\|^2) \geq \exp(b\omega_2^2))
\leq \frac{E[\exp(b\|Mu\|)]}{\exp(b\omega_2^2)}
\leq 2^n \exp(-b\omega_2^2)
= B \exp(-b\omega_2^2)
\]

(Step 2) As the discussion before, a natural idea to obtain the tail bound of \( \|M\|_2 \) is to use the simple union bound on all points of the unit sphere for the operator norm:
\[
P(\|M\|_2 \geq \omega_2) \leq P\left( \bigcup_{u \in S} \|Mu\| \geq \omega_2 \right).
\]
However the unit sphere \( S \) is uncountable. To solve this issue, we can use the trick of maximal \( \varepsilon \)-net of the unit sphere, denoted by \( \Sigma(\varepsilon) \), since the maximal \( \varepsilon \)-net of the sphere \( S \) is countable. This leads to the next step.

(Step 3) We have the following Lemma 9 which uses the maximal \( \varepsilon \)-net of the sphere \( S \). It is Lemma 2.3.2 in Tao (2012).

**Lemma 9.** Let \( 0 < \varepsilon < 1 \) and \( \Sigma(\varepsilon) \) be the maximal \( \varepsilon \)-net of the sphere \( S \), that is the set of points in \( S \) separated from each other by a distance of at least \( \varepsilon \), and which is maximal with respect to set inclusion. Then for the matrix \( M \in \mathbb{R}^{n \times p} \) and any \( \omega_2 > 0 \), we have
\[
P(\|M\|_2 \geq \omega_2) \leq P\left( \bigcup_{u \in \Sigma(\varepsilon)} \|Mu\| \geq \omega_2 (1 - \varepsilon) \right)
\]

(Step 4) Since the maximal \( \varepsilon \)-net of the sphere \( S \) is countable, the following Lemma 10 gives its cardinality. It is Lemma 2.3.4 in Tao (2012).

**Lemma 10.** Let \( 0 < \varepsilon < 1 \), then \( \Sigma(\varepsilon) \) has cardinality at most \( \frac{C_1}{\varepsilon^p} \) for some constant \( C_1 > 0 \).
(Step 5) Based on Lemma 9 and 10, now we have

\[
P(\|M\|_2 \geq \omega_2) \leq P\left( \bigcup_{u \in \Sigma(\varepsilon)} \|Mu\| \geq \omega_2(1 - \varepsilon) \right)
\leq \frac{C_1}{\varepsilon^{p-1}} P(\|Mu\| \geq \omega_2(1 - \varepsilon)), \quad \text{(Since the cardinality of } \Sigma(\varepsilon) \text{ is bounded by } \frac{C_1}{\varepsilon^{p-1}})
\leq C_1 B \cdot 2^{p-1} \exp\left(-\frac{b}{4}\omega_2^2\right), \quad \text{(set } \varepsilon = 1/2)
\leq C \exp(p \log 2 - \frac{b}{4}\omega_2^2).
\]

Let \( \omega_2^2 = \frac{4(\rho \log 2 + \alpha \log p)}{\varepsilon} = c_3 p + c_4 \alpha \log p \), we have

\[
P\left[\|G(L)\|_2 \geq \omega_2\right] \leq C p^{-\alpha}. \tag{A.8.4}
\]

(4) Combine results (A.8.2), (A.8.3) and (A.8.4) with \( \omega_0^2 = c_1 (1 + \alpha) n \log p \), \( \omega_1^2 = c_2 (2 + \alpha) n \log p \) and \( \omega_2^2 = c_3 p + c_4 \alpha \log p \), we have

\[
D_f \leq K^2(c_1 (1 + \alpha) n \log p + 4c_2 (2 + \alpha) n \log p + 4c_3 p + 4c_4 \alpha \log p)(s + r),
\]

with probability \( 1 - cp^{-\alpha} \), where \( c = \max(2, C) \).

\[\blacksquare\]

**A.9 Proof of Theorem 4**

*Proof.* Denote \( (\hat{\beta}, \hat{\Theta}, \hat{L}) \) as the solution of our target function (2.2.7), the KKT condition for the \( \ell_1 \) norm reads

\[
\frac{\partial}{\partial \Theta} f(\hat{\beta}, \hat{\Theta}, \hat{L}) + \lambda_1 \hat{Z} = 0, \tag{A.9.1}
\]

where \( \hat{Z} \) is a subgradient of the function \( \|\hat{\Theta}\|_1 \), i.e.,

\[
\hat{Z}_{ij} = \begin{cases} 
0 & \text{if } i = j \\
n\sign(\hat{\Theta}_{ij}) & \text{if } i \neq j \text{ and } \hat{\Theta}_{ij} \neq 0 \\
\in [-1, +1] & \text{if } i \neq j \text{ and } \hat{\Theta}_{ij} = 0 
\end{cases}
\]
Then we have the Taylor expansion of the gradient function
\[
vec \left( \frac{\partial}{\partial \Theta} f(\hat{\beta}, \hat{\Theta}, \hat{L}) \right) = vec \left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) \\
+ \frac{\partial^2}{\partial \Theta^2} f(\beta^*, \Theta^*, L^*)(vec(\hat{\Theta}) - vec(\Theta^*)) \\
+ \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*)(vec(\hat{L}) - vec(L^*)) \\
+ \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*)(\hat{\beta} - \beta^*) + R_\Theta,
\]
(A.9.2)

where \( R_\Theta \) is the higher order term and \( vec(\cdot) \) is the vectorization of matrix. Plugging above function (A.9.2) to the KKT function (A.9.1), we have
\[
0 = \frac{\partial^2}{\partial \Theta^2} f(\beta^*, \Theta^*, L^*)(vec(\hat{\Theta}) - vec(\Theta^*)) + vec \left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) \\
+ \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*)(vec(\hat{L}) - vec(L^*)) + \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*)(\hat{\beta} - \beta^*) + R_\Theta + \lambda_1 vec(\hat{Z}).
\]
(A.9.3)

Denote
\[
\Sigma = \frac{\partial^2}{\partial \Theta^2} f(\beta^*, \Theta^*, L^*),
\]
and
\[
\varepsilon = vec \left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) \right) + \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*)(vec(\hat{L}) - vec(L^*)) \\
+ \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*)(vec(\hat{L}) - vec(L^*)) + \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*)(\hat{\beta} - \beta^*) + R_\Theta,
\]
the function (A.9.3) can be written as the following simple form
\[
\Sigma vec(\hat{\Theta}) = \Sigma vec(\Theta^*) - \varepsilon - \lambda_1 vec(\hat{Z}).
\]
(A.9.4)

Denote \( nz = \{ j : vec(\Theta^*)_j \neq 0 \} \) and \( z = \{ j : vec(\Theta^*)_j = 0 \} \), here \( vec(\Theta^*)_{nz} \) is the sub-vector with index \( nz \), we have the following result

**Lemma 11.** Assume \( \Sigma_{nz} \) is nonsingular, then (A.9.4) is equivalent to
\[
S_z vec(\hat{\Theta})_z = \Sigma_{z,nz}^{-1} \Sigma_{nz}^{-1} \varepsilon_{nz} - \varepsilon_z + \lambda_1 \Sigma_{z,nz}^{-1} vec(\hat{Z})_{nz} - \lambda_1 vec(\hat{Z})_z \\
vec(\hat{\Theta})_{nz} = vec(\Theta^*)_{nz} - \Sigma_{nz}^{-1} \varepsilon_{nz} - \lambda_1 \Sigma_{nz}^{-1} vec(\hat{Z})_{nz} - \Sigma_{nz}^{-1} \Sigma_{nz,z}^{-1} vec(\hat{\Theta})_z
\]
(A.9.5)
(A.9.6)
where \( S_z = \Sigma_z - \Sigma_{z,nz}\Sigma^{-1}_{nz}\Sigma_{nz,z} \).

**Proof.** Written the matrix \( \Sigma \) as a block form

\[
\Sigma = \begin{bmatrix}
\Sigma_{nz} & \Sigma_{nz,z} \\
\Sigma_{z,nz} & \Sigma_z
\end{bmatrix}
\]

then equations (A.9.5) and (A.9.6) can be obtained directly from the block matrix multiplication.

Here we focus on \( P_s \) the probability that there exists one estimator \( \hat{\Theta} \) such that \( nz(\text{vec}(\hat{\Theta})) = nz(\text{vec}(\Theta^*)) \). Define

\[
A \triangleq \left\{ \left| \Sigma_{z,nz}\Sigma^{-1}_{nz}\varepsilon - \varepsilon_z + \lambda_1\Sigma_{z,nz}\Sigma^{-1}_{nz}\text{vec}(\hat{Z})_{nz} \right| \leq \lambda_1 \right\},
\]

and

\[
V \triangleq \left\{ |\Sigma^{-1}_{nz}\varepsilon| + \lambda_1|\Sigma^{-1}_{nz}s| < |\text{vec}(\Theta^*)_{nz}| \text{ for any } s \text{ satisfying } |s| \leq 1 \right\}.
\]

Then we have

\[
1 - P_s \leq P(A^c \cup V^c) \leq P(A^c) + P(V^c)
\]

(1) To bound \( P(A^c) \), define

\[
\mu := \mu_{\min}(\Sigma_{nz}), \text{ and } \kappa := \frac{\max_{i \in nz} \|\Sigma_{i,nz}\|_2}{\sqrt{d_{nz}}}
\]

where \( \mu_{\min} \) denotes the smallest eigenvalue, and \( d_{nz} = |nz| \). Let \( \Sigma_{z,nz} = [v_1, \ldots, v_{d_z}]^T \), then \( \|v_i\|_2 \leq \kappa \sqrt{d_{nz}} \), hence

\[
|v_i^T\Sigma^{-1}_{nz}\text{vec}(\hat{Z})_{nz}| \leq \|v_i\|_2 \cdot \|\Sigma^{-1}_{nz}\|_2 \cdot \|\text{vec}(\hat{Z})_{nz}\|_2 \leq \frac{\kappa d_{nz}}{\mu},
\]

and

\[
|\Sigma_{z,nz}\Sigma^{-1}_{nz}\text{vec}(\hat{Z})_{nz}| \leq \frac{\kappa d_{nz}}{\mu}.
\]
It follows that

\[ P(A^c) \leq P\left( \max |\Sigma_{nz}^{-1}z_n - \varepsilon| \geq (1 - \frac{\kappa d_{nz}}{\mu}) \lambda_1 \right) \]

\[ \leq P\left( \|\Sigma_{nz}^{-1}z_n\|_{\infty} + \|\varepsilon\|_{\infty} \geq (1 - \frac{\kappa d_{nz}}{\mu}) \lambda_1 \right) \]

\[ \leq P\left( \|\Sigma_{nz}^{-1}z_n\|_{\infty} \cdot \|z_n\|_{\infty} + \|\varepsilon\|_{\infty} \geq (1 - \frac{\kappa d_{nz}}{\mu}) \lambda_1 \right) \]

\[ \leq P\left( \frac{1 + \kappa d_{nz}}{\mu} \cdot \|\varepsilon\|_{\infty} \geq (1 - \frac{\kappa d_{nz}}{\mu}) \lambda_1 \right) \]

\[ = P\left( \|\varepsilon\|_{\infty} \geq \frac{\mu}{\mu + \kappa d_{nz}} \lambda_1 \right) \quad \text{(A.9.7)} \]

(2) To bound \( P(V^c) \), suppose the spectral decomposition of \( \Sigma_{nz} \) is given by \( \Sigma_{nz} = UDU^T \) with \( U = [u_1, ..., u_{d_{nz}}]^T \), then we can represent \( \Sigma_{nz}^{-1} \) as \( [u_1^T D^{-1} u_1]_{d_{nz} \times d_{nz}} \), and \( \Sigma_{nz}^{-1}s \) as \( \sum_{j=1}^{d_{nz}} s_j u_1^T D^{-1} u_j \) \( d_{nz} \times 1 \). It follows that

\[ \text{diag}(\Sigma_{nz}^{-1}) \leq \frac{1}{\mu}, \text{ and } |\Sigma_{nz}^{-1}s| \leq \frac{d_{nz}}{\mu}. \]

Therefore

\[ P(V^c) \leq P\left( \max |\Sigma_{nz}^{-1}z_n| \geq \min |\text{vec}(\Theta^*)_{nz}| - \frac{\lambda_1 d_{nz}}{\mu} \right) \]

\[ \leq P\left( \|\Sigma_{nz}^{-1}\|_2 \cdot \|\varepsilon\|_{\infty} \geq \min |\text{vec}(\Theta^*)_{nz}| - \frac{\lambda_1 d_{nz}}{\mu} \right) \]

\[ = P\left( \|\varepsilon\|_{\infty} \geq \mu \cdot \min |\text{vec}(\Theta^*)_{nz}| - \lambda_1 d_{nz} \right) \quad \text{(A.9.8)} \]

(3) From equations (A.9.7) and (A.9.8), we can know that \( \|\varepsilon\|_{\infty} \) plays an important role in the analysis. Recall that

\[ \varepsilon = \text{vec}\left( \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) + \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) (\text{vec}(\hat{L}) - \text{vec}(L^*)) \right) \]

\[ + \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) (\text{vec}(\hat{L}) - \text{vec}(L^*)) + \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) (\hat{\beta} - \beta^*) + R_{\Theta}, \]

Here we have

\[ \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*) = -X^T X + X^T \exp (1_n \beta^* + XS^* + L^*) \cdot \frac{\exp (1_n \beta^* + XS^* + L^*)}{1 + \exp (1_n \beta^* + XS^* + L^*)}. \]

98
Let 
\[ F(L) = X^T b'(1_n \beta^T + XS + L), \text{ where } b(\cdot) = \log(1 + \exp(\cdot)), \]
then we have
\[
dF(L) = X^T [b''(1_n \beta^T + XS + L) \circ dL]
dvecF(L) = (I_{p \times p} \otimes X^T) vec[b''(1_n \beta^T + XS + L) \circ dvecL]
\]
therefore
\[
\frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) = (I_{p \times p} \otimes X^T) Diag\{vec(b''(1_n \beta^T + XS + L))\}.
\]
Similarly, let
\[ F(\beta) = X^T b'(1_n \beta^T + XS + L), \text{ where } b(\cdot) = \log(1 + \exp(\cdot)), \]
then we have
\[
dF(\beta) = X^T [b''(1_n \beta^T + XS + L) \circ 1_n d(\beta^T)]
dvecF(\beta) = (I_{p \times p} \otimes X^T) vec[b''(1_n \beta^T + XS + L) \circ 1_n dvec(\beta)]
\]
therefore
\[
\frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) = (I_{p \times p} \otimes X^T) Diag\{vec(b''(1_n \beta^T + XS + L))\} (I_{p \times p} \otimes 1_n) dvec(\beta)
\]
Since
\[
\|\varepsilon\|_\infty \leq \left\| vec\left(\frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*)\right)\right\|_\infty + \left\| \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*)\right\|_\infty \left\| vec(\hat{L}) - vec(L^*)\right\|_\infty + \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) \|\hat{\beta} - \beta^*\|_\infty + \|R_\Theta\|_\infty,
\]
we just need to bound the terms at the l.h.s of (A.9.9).
(1) For the terms $\left\| \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty$, we denote the matrix by $U$, i.e.,

$$U = \frac{\partial^2}{\partial L \partial \Theta} f(\beta^*, \Theta^*, L^*) = (I_{p \times p} \otimes X^T) \text{Diag}\{\text{vec}(b''(1_n \beta^T + XS^* + L^*))\}.$$  

Since function $b''(\cdot)$ is the derivative of the sigmoid function $\frac{\exp(\cdot)}{1 + \exp(\cdot)} \in (0, 1)$, we have $b''(\cdot) \in (0, \frac{1}{4})$. Recall that each entry in $X$ takes value 0 or 1, then we have the nonzero element $U_{ij}$ in bounded in $(0, \frac{1}{4})$, and the orlicz norm $\|U_{ij}\|_{\psi_2} = c < \infty$, where $c$ is a small positive constant, $i = 1, \ldots, p^2$ and $j = 1, \ldots, np$. By the structure of matrix $U$, we can know that there are at most $n$ nonzero elements in each row, thus for any $t_1 > 0$, we have

$$P(\|U\|_\infty > t_1) = P(\max_i \sum_{j=1}^{np} |U_{ij}| > t_1) \leq p^2 \cdot P(\sum_{j=1}^{np} |U_{ij}| > t_1) \leq \exp\left(2 \log p - \frac{t_1^2}{c_1 n}\right),$$

(by the row independence and the Hoeffding bound)

where $c_1$ is a positive constant. Let $t_1^2 = c_1 n(2 + \alpha) \log p$, where $\alpha > 0$, we have

$$P(\|U\|_\infty > t_1) \leq p^{-\alpha}.$$

(2) For the terms $\left\| \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) \right\|_\infty$, we denote the matrix by $V$, i.e.,

$$V = \frac{\partial^2}{\partial \beta \partial \Theta} f(\beta^*, \Theta^*, L^*) = (I_{p \times p} \otimes X^T) \text{Diag}\{\text{vec}(b''(1_n \beta^T + XS^* + L^*))\}(I_{p \times p} \otimes 1_n).$$

By the structure of matrix $V$, we can know that there is only one nonzero element $V_{st}$ in each row, which means $\sum_{t=1}^{p} |V_{st}| = |V_{st}|$, where $s = 1, \ldots, p^2$ and $t = 1, \ldots, p$. Similar to the analysis above, since $b''(\cdot) \in (0, \frac{1}{4})$ and each entry in $X$ takes value 0 or 1, we have $V_{st} \in (0, \frac{n}{4})$ and the orlicz norm $\|V_{st}\|_{\psi_2} \lesssim \sqrt{n}$. Then
for any $t_2 > 0$, we have

$$P(\|V\|_\infty > t_2) = P(\max_s \sum_{t=1}^p |V_{st}| > t_2)$$

$$\leq p^2 \cdot P(\sum_{t=1}^p |V_{st}| > t_2)$$

$$= p^2 \cdot P(V_{st} > t_2)$$

$$\leq \exp (2 \log p - \frac{t_2^2}{c_2n})$$

where $c_2$ is a positive constant. Let $t_2^2 = c_2n(2 + \alpha)$ log $p$, we have

$$P(\|V\|_\infty > t_2) \leq p^{-\alpha}.$$  

(3) For the term $\|vec(\frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*))\|_\infty$, let $G(\Theta) = \frac{\partial}{\partial \Theta} f(\beta^*, \Theta^*, L^*)$, as we show in (A.8.3), for any $t_3 > 0$, we have

$$P\left[\|G(\Theta)\|_\infty \geq t_3\right] = P\left[\max_{1 \leq i,j \leq p} |G(\Theta)_{ij}| \geq t_3\right]$$

$$\leq \sum_{1 \leq i,j \leq p} P\left[|G(\Theta)_{ij}| \geq t_3\right]$$

$$\leq 2 \exp(2 \log p - \frac{t_3^2}{c_3n}).$$

Let $t_3^2 = c_3n[\log 2 + (2 + \alpha) \log p]$, we have

$$P\left[\|G(\Theta)\|_\infty \geq t_3\right] \leq p^{-\alpha}. \tag{A.9.10}$$

(4) For the rest terms, since $\hat{\beta}$ is assumed sign consistent and $\hat{L}$ is assumed rank consistent, then we have $\|\hat{\beta} - \beta^*\|_\infty < \min |\beta^*|$, $\|vec(\hat{L}) - vec(L^*)\|_\infty \leq K_1$, and the higher order term $\|R_0\|_\infty \leq K_2$, where $K_1, K_2$ are some positive constants. Combine the above results (1)-(4), we have the following result

$$P(\|\epsilon\|_\infty \geq t_1K_1 + t_2 \min |\beta^*| + t_3 + K_2) \leq p^{-\alpha}.$$  

Hence if

$$\frac{\mu - \kappa d_{nz}}{\mu + \kappa d_{nz}} \lambda_1 \geq t_1K_1 + t_2 \min |\beta^*| + t_3 + K_2,$$

101
then we have
\[ P(A^c) \leq P\left( \|\varepsilon\|_{\infty} \geq \frac{\mu - \kappa d_{nz}}{\mu + \kappa d_{nz}} \lambda_1 \right) \leq p^{-\alpha}. \]

Similarly, if
\[ \mu \cdot \min |\text{vec}(\Theta^*)_{nz}| - \lambda_1 d_{nz} \geq t_1 K_1 + t_2 \min |\beta^*| + t_3 + K_2, \]
then we have
\[ P(V^c) \leq P\left( \|\varepsilon\|_{\infty} \geq \mu \cdot \min |\text{vec}(\Theta^*)_{nz}| - \lambda_1 d_{nz} \right) \leq p^{-\alpha}. \]

Since \( 1 - P_s \leq P(A^c \cup V^c) \leq P(A^c) + P(V^c) \), we can know that
\[ P_s \geq 1 - 2p^{-\alpha}. \]
Appendix B

Proofs in Chapter 3

B.1 Proof of Theorem 5

Proof. In stage 1, we treat the interaction term and the original error term as a new noise $W$, then the transformed model (3.2.5) is

$$UD^{-1}U^TY = UD^{-1}U^T X \beta^{(M)} + UD^{-1}U^T W$$

$$\tilde{Y} = \tilde{X} \beta^{(M)} + \tilde{W}$$  \hspace{1cm} (B.1.1)

where $\tilde{Y} = UD^{-1}U^TY$, $\tilde{X} = UD^{-1}U^TX = UV^T$, $\tilde{W} = UD^{-1}U^TW$ denote the transformed response vector, the transformed design matrix and the transformed error term, respectively.

The proof is based on the following Lemma 12 which is first shown in Wainwright (2006). It gives necessary and sufficient conditions for the sign consistency of the $\ell_1$-regularized estimator in the linear model. In the low-dimensional setting, for the transformed data $\tilde{Y}$ and $\tilde{X}$ in the DECO-IS estimator, i.e., in model (B.1.1), we only need to verify these conditions stated in Lemma 12 to obtain the consistency result of the DECO-IS estimator.

Lemma 12. For linear model $Y = X\beta^* + \epsilon$, assume that the matrix $X(S)^TX(S)$ is invertible, where $S$ is the support set of $\beta^*$ and $X(S)$ is the sub-matrix of $X$ with the column index $S$. Then for any given $\lambda > 0$ and any noise term $\epsilon \in \mathbb{R}^n$, there exists a Lasso estimate $\hat{\beta}(\lambda)$

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\},$$

which satisfies $\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)$, if and only if the following two conditions
\begin{align*}
&\text{(a)} \ |X(S^C)^TX(S)(X(S)^TX(S))^{-1}[X(S)^T\epsilon - \lambda \text{sign}(\beta^*(S))] - X(S^C)^T\epsilon| \leq \lambda \\
&\text{(b)} \ \text{sign} \left(\beta^*(S) + (X(S)^TX(S))^{-1}[X(S)^T\epsilon - \lambda \text{sign}(\beta^*(S))]\right) = \text{sign}(\beta^*(S))
\end{align*}

where the vector inequality and equality are taken elementwise.

Wainwright (2006) stated sufficient conditions for (a) and (b) in Lemma 12. Define

\begin{align*}
V_j &= X_j^T \left\{ X(S)(X(S)^TX(S))^{-1}\lambda \text{sign}(\beta^*(S)) - [X(S)(X(S)^TX(S))^{-1}X(S)^T - I] \epsilon \right\}, \\
U_i &= e_i^T (X(S)^TX(S))^{-1}[X(S)^T\epsilon - \lambda \text{sign}(\beta^*(S))],
\end{align*}

where \(e_i\) is the vector with 1 in the \(i\)th position and zeros elsewhere. By simple calculations, condition (a) holds strictly if and only if the following \textit{strict dual feasibility} condition holds

\begin{equation}
\mathcal{M}(V) = \left\{ \max_{j \in S^C} |V_j| < \lambda \right\}, \tag{B.1.2}
\end{equation}

and the following \textit{sign consistency} condition is sufficient to guarantee that condition (b) holds

\begin{equation}
\mathcal{M}(U) = \left\{ \max_{i \in S} |U_i| < M(\beta^*) \right\}, \tag{B.1.3}
\end{equation}

where \(M(\beta^*) = \min_{j \in S} |\beta_j^*|\), and \(S = \left\{ j : \beta_j^* \neq 0 \right\}\).

Then our goal is to check conditions (B.1.2) and (B.1.3) on the transformed data \((\tilde{Y}, \tilde{X})\) in (B.1.1) when \(n > p\).

\textbf{(Part I) Verifying the \textit{strict dual feasibility} condition.}

Our interaction model (B.1.1) has the transformed response vector \(\tilde{Y}\) and the transformed design matrix \(\tilde{X}\), therefore the \textit{strict dual feasibility} condition in (B.1.2) takes the following form

\begin{equation}
\mathcal{M}(\tilde{V}) = \left\{ \max_{j \in S^C} |\tilde{V}_j| < \lambda \right\}. \tag{B.1.4}
\end{equation}
Here we have
\[
\tilde{V}_j = \tilde{X}_j^T \left\{ \tilde{X}(S)(\tilde{X}(S)^T \tilde{X}(S))^{-1}\text{sign}(\beta(M)^{(S)}) - \tilde{X}(S)(\tilde{X}(S)^T - I) \tilde{W} \right\} \\
= \tilde{X}_j^T \left\{ \tilde{X}(S)\text{sign}(\beta(M)^{(S)}) - \tilde{X}(S)\tilde{X}(S)^T - I \tilde{W} \right\} \quad (\text{Since } \tilde{X}^T \tilde{X} = I \text{ when } n > p) \\
= e_j^T \left\{ \tilde{X}(S^C)^T \tilde{X}(S)\text{sign}(\beta(M)^{(S)}) - \tilde{X}(S^C)^T \tilde{X}(S)\tilde{X}(S)^T - I \tilde{W} \right\} \\
= 0 - e_j^T \left[ \tilde{X}(S^C)^T \tilde{X}(S)\tilde{X}(S)^T - \tilde{X}(S^C)^T \tilde{X}(S) \right] \tilde{W} \quad (\text{Since } \tilde{X}(S^C)^T \tilde{X}(S) = 0 \text{ when } n > p) \\
= e_j^T \tilde{X}(S^C)^T \tilde{W}.
\]

Then the condition in (B.1.4)
\[
\max_{j \in S^C} |\tilde{V}_j| < \lambda, 
\]
is equivalent to
\[
\|\tilde{X}(S^C)^T \tilde{W}\|_\infty < \lambda. \quad (B.1.5)
\]

To check the above condition (B.1.5), first we state the following definition and properties of Moore-Penrose pseudo-inverse.

**Definition 7.** For every matrix \( X \in \mathbb{R}^{n \times p} \), there always exits a unique pseudo-inverse, denoted by \( X^+ \), and the matrix \( G = X^+ \) if and only if

(P1) \( XGX = X \)

(P2) \( GXG = G \)

(P3) \( (XG)^T = XG \)

(P4) \( (GX)^T = GX \)

**Property 1.** (1) \( (A^T)^+ = (A^+)^T \); (2) \( (AB)^+ = B^+A^+ \).

We have the pseudo-inverse of \( X \) is \( X^+ = VD^{-1}U^T \). Besides, since \( XX^T = UDV^TVDU^T = UD^2U^T \), then \( UD^{-1}U^T = (XX^T)^\frac{1}{2} \) when \( n > p \), where \( UD^{-1}U^T \) is an \( n \times n \) matrix. Now the condition (B.1.5) can be written as

\[
\|\tilde{X}(S^C)^T \tilde{W}\|_\infty = \|X(S^C)^T(XX^T)^+ W\|_\infty \\
\leq \|X(S^C)^T(XX^T)^+ Y_f\|_\infty + \|X(S^C)^T(XX^T)^+ \varepsilon\|_\infty \quad (B.1.6)
\]

Assume that \( X(S) \) consists of the first \( q \) columns of the design matrix \( X \), and \( X(S^C) \)
consists of the rest \( p - q \) columns of \( X \), then we have

\[
X(S^C) = XE_1, \text{ where } E_1 = \begin{pmatrix}
0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots \\
0 & 0 & \cdots & 0 \\
1 & 0 & \cdots & 0 \\
\vdots \\
0 & 0 & \cdots & 1 \\
\end{pmatrix}_{p-q \text{ columns}}
\] (B.1.7)

Hence we can bound the two parts in (B.1.6) separately,

\[
\|X(S^C)^T (XX^T)^+ Y_I\|_\infty = \|E_1^T X^T (XX^T)^+ Y_I\|_\infty \\
\leq \|E_1\|_\infty \|X^T (XX^T)^+ Y_I\|_\infty \\
= \|X^T (XX^T)^+ Y_I\|_\infty
\]

and similarly we have

\[
\|X(S^C)^T (XX^T)^+ \varepsilon\|_\infty \leq \|X^T (XX^T)^+ \varepsilon\|_\infty. \tag{B.1.9}
\]

By the properties of Moore-Penrose pseudo-inverse, we have

\[
(XX^T)^+ = (X^T)^+ X^+ \\
= (X^+)^T X^+ \\
= [(X^T X)^{-1} X^T] (X^T X)^{-1} X^T \quad \text{(Since } X^+ = (X^T X)^{-1} X^T \text{ when } n \geq p) \\
= X (X^T X)^{-1} (X^T X)^{-1} X^T.
\]

Hence we have

\[
X^T (XX^T)^+ = X^T X (X^T X)^{-1} (X^T X)^{-1} X^T = (X^T X)^{-1} X^T,
\]

which means the upper bounds (B.1.8) and (B.1.9) can be written as

\[
\|X(S^C)^T (XX^T)^+ Y_I\|_\infty \leq \|X^T (XX^T)^+ Y_I\|_\infty \\
= \|(X^T X)^{-1} X^T Y_I\|_\infty, \tag{B.1.10}
\]

106
and

\[ \|X(S^C)^T(XX^T)^+\|_\infty \leq \|X^T(XX^T)^+\|_\infty = \|(X^T X)^{-1}X^T \|_\infty. \]  

(B.1.11)

In summary, to bound the target \( \|\tilde{X}(S^C)^T\tilde{W}\|_\infty \) in (B.1.5), we only need to bound the term \( \|\tilde{(X^T X)^{-1}X^T Y^I}\|_\infty \) in (B.1.10) and \( \|\tilde{(X^T X)^{-1}X^T \varepsilon}\|_\infty \) in (B.1.11).

1. For the term \( \|\tilde{(X^T X)^{-1}X^T Y^I}\|_\infty \), we have the following Lemma 13 (Hao et al., 2018).

**Lemma 13.** Under the interaction model (3.3.1), assume that \( X_i \) (the \( i \)th row of design matrix \( X \)), \( 1 \leq i \leq n \) is independent and identically distributed from \( N(0, \Sigma) \), \( \varepsilon \sim N(0, \sigma^2 I) \) is independent of \( \{X_i\}_{i=1}^n \), then we have

\[ P\left( \|\tilde{(X^T X)^{-1}X^T Y^I}\|_\infty \geq \frac{9\|\beta^{(I)*}\|_2 \sqrt{\log p}}{C_{\min} n^{\frac{1}{4}}} \right) \leq c_1 \exp(-c_2 \log p) \]  

(B.1.12)

where \( \beta^{(I)*} \) is the true parameter of the interaction term, \( C_{\min} \) denotes the minimal eigenvalue of \( X^T X \), and \( c_1, c_2 \) are positive constants.

2. For the term \( \|\tilde{(X^T X)^{-1}X^T \varepsilon}\|_\infty \), we know that \( (X^T X)^{-1}X^T \varepsilon \) is zero-mean Gaussian with variance at most

\[ \hat{\sigma}^2(X) := \frac{\sigma^2}{n} \|\left(\frac{X^T X}{n}\right)^{-1}\|_2. \]

Define the event

\[ \mathcal{T}(X) := \left\{ \hat{\sigma}^2(X) \geq \frac{9\sigma^2}{nC_{\min}} \right\}, \]

when \( n \geq p \), let \( X \in \mathbb{R}^{nxp} \) have i.i.d rows \( X_i \sim N(0, \Sigma) \), we have the following result (Wainwright, 2009)

\[ P\left( \frac{\|X^T X \|_2}{n} - 1 \right) \geq \frac{9}{C_{\min}} \leq 2 \exp\left(-\frac{n}{2}\right), \]

which means

\[ P[\mathcal{T}(X)] \leq 2 \exp\left(-\frac{n}{2}\right). \]

By the total probability rule, we have

\[ P\left[ (X^T X)^{-1}X^T \varepsilon > t \right] \leq P\left[ (X^T X)^{-1}X^T \varepsilon > t | \mathcal{T}(X) \right] + P[\mathcal{T}(X)] \]

Conditioned on \( \mathcal{T}(X) \), the random variable \( (X^T X)^{-1}X^T \varepsilon \) is zero-mean Gaussian.
with variance at most \( \frac{9\sigma^2}{nC_{\min}} \), so that by Gaussian tail bounds, we have

\[
P\left( \left\| (X^T X)^{-1}X^T \varepsilon \right\|_\infty \geq t \right) \leq 2p\exp\left(-\frac{C_{\min}nt^2}{162\sigma^2}\right).
\]

Let \( t = 20\sqrt{\frac{\sigma^2 \log p}{C_{\min}n}} \), overall we have the following result

\[
P\left( \left\| (X^T X)^{-1}X^T \varepsilon \right\|_\infty \geq 20\sqrt{\frac{\sigma^2 \log p}{C_{\min}n}} \right) \leq 4\exp(-c_3 \log p), \quad (B.1.13)
\]

where \( c_3 \) is positive constant.

Now we are ready to show the following target inequality that we need

\[
\max_{j \in S^C} |\tilde{V}_j| < \lambda,
\]

that is

\[
\|\tilde{X}(S^C)^T \tilde{W}\|_\infty < \lambda.
\]

Since

\[
\|\tilde{X}(S^C)^T \tilde{W}\|_\infty \leq \|(X^T X)^{-1}X^T \gamma\|_\infty + \|(X^T X)^{-1}X^T \varepsilon\|_\infty,
\]

combining the results (B.1.12) and (B.1.13), i.e.,

\[
P\left( \left\| (X^T X)^{-1}X^T \gamma\right\|_\infty \geq \frac{9\|\beta(I)^{*}\|_2 \sqrt{\log p}}{C_{\min}n^{3/4}} \right) \leq c_1 \exp(-c_2 \log p),
\]

and

\[
P\left( \left\| (X^T X)^{-1}X^T \varepsilon\right\|_\infty \geq 20\sqrt{\frac{\sigma^2 \log p}{C_{\min}n}} \right) \leq 4\exp(-c_3 \log p),
\]

once \( \lambda > \frac{9\|\beta(I)^{*}\|_2 \sqrt{\log p}}{C_{\min}n^{3/4}} + 20\sqrt{\frac{\sigma^2 \log p}{C_{\min}n}} \), we have

\[
P\left( \max_{j \in S^C} |\tilde{V}_j| > \lambda \right) \leq cp^{-\alpha},
\]

where \( \beta(I)^{*} \) is the true parameter of the interaction term, \( C_{\min} \) denotes the minimal eigenvalue of \( X^T X \), \( c \) and \( \alpha \) are positive constants. This is the end to verify the strict dual feasibility condition.

**(Part II) Verify the sign consistency condition.**

Based on the transformed data \((\tilde{Y}, \tilde{X})\), the sign consistency condition in
(B.1.3) has the following form:

\[ \mathcal{M}(\hat{U}) = \left\{ \max_{i \in S} |\hat{U}_i| < M(\beta^{(M)*}) \right\}, \tag{B.1.14} \]

where \( M(\beta^{(M)*}) = \min_{j \in S} |\beta_j^{(M)*}| \), \( S = \left\{ j : \beta_j^{(M)*} \neq 0 \right\} \) and

\[ \hat{U}_i = \epsilon_i^T (\tilde{X}(S)^T \tilde{X}(S))^{-1} \left[ \tilde{X}(S)^T \tilde{W} - \lambda \text{sign}(\beta^{(M)*}(S)) \right]. \tag{B.1.15} \]

Here we have the following result

\[
\max_{i \in S} |\hat{U}_i| \leq \| (\tilde{X}(S)^T \tilde{X}(S))^{-1} \tilde{X}(S)^T \tilde{W} \|_\infty + \lambda \| (\tilde{X}(S)^T \tilde{X}(S))^{-1} \text{sign}(\beta^{(M)*}(S)) \|_\infty
\]

\[
= \| \tilde{X}(S)^T \tilde{W} \|_\infty + \lambda \| \text{sign}(\beta^{(M)*}(S)) \|_\infty
\]

then the condition (B.1.14) is equivalent to

\[
\| \tilde{X}(S)^T \tilde{W} \|_\infty + \lambda < M(\beta^{(M)*}). \tag{B.1.16}
\]

Since \( UD^{-1}U^T = (XX^T)^{1/2} \) when \( n > p \), we have

\[
\| \tilde{X}(S)^T \tilde{W} \|_\infty = \| X(S)^T (XX^T)^{1/2} W \|_\infty
\]

\[
\leq \| X(S)^T (XX^T)^{1/2} Y_I \|_\infty + \| X(S)^T (XX^T)^{1/2} \varepsilon \|_\infty. \tag{B.1.17}
\]

Similar to the analysis in Part I, we assume that \( X(S) \) consists of the first \( q \) columns of design matrix \( X \), then we have

\[
X(S) = XE_2, \quad \text{where } E_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}_{(n-q) \text{ rows}} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{q \text{ columns}}
\]
Hence the two terms in (B.1.17) can be written as
\[ \|X(S)^T (XX^T)^+ Y_I\|_\infty = \|E_2^T X^T (XX^T)^+ Y_I\|_\infty \]
\[ \leq \|E_2^T\|_\infty \|X^T (XX^T)^+ Y_I\|_\infty \]
\[ = \|X^T (XX^T)^+ Y_I\|_\infty, \quad (B.1.18) \]
and similarly,
\[ \|X(S)^T (XX^T)^+ \varepsilon\|_\infty \leq \|X^T (XX^T)^+ \varepsilon\|_\infty. \quad (B.1.19) \]

By the properties of Moore-Penrose pseudo-inverse, we have
\[ (XX^T)^+ = X(X^T X)^{-1}(X^T X)^{-1}X^T, \]
and
\[ X^T (XX^T)^+ = X^T X(X^T X)^{-1}(X^T X)^{-1}X^T = (X^T X)^{-1}X^T. \]
Hence the upper bounds (B.1.18) and (B.1.19) can be written as
\[ \|X(S)^T (XX^T)^+ Y_I\|_\infty \leq \|X^T (XX^T)^+ Y_I\|_\infty \]
\[ = \|(X^T X)^{-1}X^T Y_I\|_\infty \quad (B.1.20) \]
and
\[ \|X(S)^T (XX^T)^+ \varepsilon\|_\infty \leq \|X^T (XX^T)^+ \varepsilon\|_\infty = \|(X^T X)^{-1}X^T \varepsilon\|_\infty \quad (B.1.21) \]

Now we want to show
\[ \max_{i \in S} |\tilde{U}_i| < M(\beta(M)^*), \]
and we have
\[ \max_{i \in S} |\tilde{U}_i| \leq \|\tilde{X} (S)^T \tilde{W}\|_\infty + \lambda \]
\[ \leq \|X(S)^T (XX^T)^+ Y_I\|_\infty + \|X(S)^T (XX^T)^+ \varepsilon\|_\infty + \lambda \]
\[ \leq \|X^T (XX^T)^+ Y_I\|_\infty + \|X^T (XX^T)^+ \varepsilon\|_\infty + \lambda \]
\[ = \|(X^T X)^{-1}X^T Y_I\|_\infty + \|(X^T X)^{-1}X^T \varepsilon\|_\infty + \lambda. \]
Based on the results (B.1.12) and (B.1.13) in Part I, i.e.,

\[ P\left( \| (X^T X)^{-1} X^T Y \|_\infty \geq \frac{9 \| \beta^{(I)}_\ast \|_2 \sqrt{\log p}}{C_{\text{min}} n^{\frac{1}{3}}} \right) \leq c_1 \exp(-c_2 \log p), \]

and

\[ P\left( \| (X^T X)^{-1} X^T \varepsilon \|_\infty \geq 20 \sqrt{\frac{\sigma^2 \log p}{C_{\text{min}} n}} \right) \leq 4 \exp(-c_3 \log p), \]

thus if 

\[ M(\beta^{(M)}_\ast) > \frac{9 \| \beta^{(I)}_\ast \|_2 \sqrt{\log p}}{C_{\text{min}} n^{\frac{1}{3}}} + 20 \sqrt{\frac{\sigma^2 \log p}{C_{\text{min}} n}} + \lambda, \]

which means 

\[ M(\beta^{(M)}_\ast) > 2\lambda, \]

then we have

\[ P\left( \max_{i \in S} |\tilde{U}_i| < M(\beta^{(M)}_\ast) \right) \leq c p^{-\alpha}, \]

where \( \beta^{(I)}_\ast \) is the true parameter of the interaction term, \( C_{\text{min}} \) denotes the minimal eigenvalue of \( X^T X \), \( c \) and \( \alpha \) are positive constants. This is the end to verify the sign consistency condition.

\[ \blacksquare \]
Bibliography


PCE, P. C. E. (2012). Don’t expect consumer spending to be the engine of economic growth it once was. Differences, 56(21.4):2–4.


