2020

Statistical Inference For High Dimensional Models In Genomics And Microbiome

Jiarui Lu
University of Pennsylvania

Statistical Inference For High Dimensional Models In Genomics And Microbiome

Abstract

Human microbiome consists of all living microorganisms that are in and on human body. Largescale microbiome studies such as the NIH Human Microbiome Project (HMP), have shown that this complex ecosystem has large impact on human health through multiple ways. The analysis of these datasets leads to new statistical challenges that require the development of novel methodologies. Motivated by several microbiome studies, we develop several methods of statistical inference for high dimensional models to address the association between microbiome compositions and certain outcomes. The high-dimensionality and compositional nature of the microbiome data make the naive application of the classical regression models invalid. To study the association between microbiome

compositions with a disease's risk, we develop a generalized linear model with linear constraints on regression coefficients and a related debiased procedure to obtain asymptotically unbiased and normally distributed estimates. Application of this method to an inflammatory bowel disease (IBD) study identifies several gut bacterial species that are associated with the risk of IBD. We also consider the post-selection inference for models with linear equality constraints, where we develop methods for constructing the confidence intervals for the selected non-zero coefficients chosen by a Lasso-type estimator with linear constraints. These confidence intervals are shown to have desired coverage probabilities when conditioned on the selected model. Finally, the last chapter of this dissertation presents a method for inference of high dimensional instrumental variable regression. Gene expression and phenotype association can be affected by potential unmeasured confounders, leading to biased estimates of the associations. Using genetic variants as instruments, we consider the problem of hypothesis testing for sparse IV regression models and present methods for testing both single and multiple regression coefficients. A multiple testing procedure is developed for selecting variables and is shown to control the false discovery rate. These methods are illustrated by an analysis of a yeast dataset in order to identify genes that are associated with growth in the presence of hydrogen peroxide.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Epidemiology & Biostatistics

First Advisor

Hongzhe Li

Keywords

Causal inference, Genomcis, High dimensional model, Microbiome, Statistical inference

Subject Categories

Biostatistics

STATISTICAL INFERENCE FOR HIGH DIMENSIONAL MODELS IN GENOMICS AND MICROBIOME

Jiarui Lu

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

Hongzhe Li

Hongzhe Li, Professor of Biostatistics

Graduate Group Chairperson

Nandita Mitra

Nandita Mitra, Professor of Biostatistics

Dissertation Committee

Ian J. Barnett, Assistant Professor of Biostatistics

T.Tony Cai, Professor of Statistics

Weijie Su, Assistant Professor of Statistics

Kai Tan, Associate Professor of Pediatrics

STATISTICAL INFERENCE FOR HIGH DIMENSIONAL MODELS IN GENOMICS AND MICROBIOME

ACKNOWLEDGEMENT

I would like to thank my advisor Dr. Hongzhe Li, for his guidance and support on my research. As a scientist, he is professional and has deep insight in research topics. His enthusiasm and creativity also inspired me when I was working with him. Being a mentor, he is patient and encouraging. During my study at the University of Pennsylvania, he helped me so much in both research and life. His values towards statistical methods and scientific fields have great impact on me. It is my great honor to start my PhD study under the supervision of Dr. Li.

I am grateful to Dr. Ian Barnett for his kind support serving as my committee chair. His offered helpful suggestions on my research projects and future career choices. I would also like to thank my committee members Dr. Tony Cai, Dr. Weijie Su and Dr. Kai Tan for providing valuable suggestions and comments. Being a member in the Department of Biostatistics, Epidemiology and Informatics, I would like to thank all the faculty members and staffs whose kindness and considerateness have helped so much during my five-year PhD life. I also need to thank students in my cohort Ali, Jordan, Eric, Lingjiao and Jiaxin. We spent time together getting through courses, lab rotations and preparing for qualification exams. Many thanks also given to my friends Xiaohan Li, Ruixiang Zhang, Haochang Shou, Xingyan Wang and Robert Li. I could hardly finish my research without the help and support from them.

Finally, I would like to thank my parents and family being supportive to my life and research in the US. I was surrounded by their warm love and encouragement, which supported me to overcome all the difficulties.

ABSTRACT

STATISTICAL INFERENCE FOR HIGH DIMENSIONAL MODELS IN GENOMICS AND MICROBIOME

Jiarui Lu

Hongzhe Li

Human microbiome consists of all living microorganisms that are in and on human body. Large-scale microbiome studies such as the NIH Human Microbiome Project (HMP), have shown that this complex ecosystem has large impact on human health through multiple ways. The analysis of these datasets leads to new statistical challenges that require the development of novel methodologies. Motivated by several microbiome studies, we develop several methods of statistical inference for high dimensional models to address the association between microbiome compositions and certain outcomes.

The high-dimensionality and compositional nature of the microbiome data make the naïve application of the classical regression models invalid. To study the association between microbiome compositions with a disease's risk, we develop a generalized linear model with linear constraints on regression coefficients and a related debiased procedure to obtain asymptotically unbiased and normally distributed estimates. Application of this method to an inflammatory bowel disease (IBD) study identifies several gut bacterial species that are associated with the risk of IBD. We also consider the post-selection inference for models with linear equality constraints, where we develop methods for constructing the confidence intervals for the selected non-zero coefficients chosen by a Lasso-type estimator with linear constraints. These confidence intervals are shown to have desired coverage probabilities when conditioned on the selected model.

Finally, the last chapter of this dissertation presents a method for inference of high dimensional instrumental variable regression. Gene expression and phenotype association can be affected by potential unmeasured confounders, leading to biased estimates of the associations. Using genetic variants as instruments, we consider the problem of hypothesis testing for sparse IV regression models and present methods for testing both single and multiple regression coefficients. A multiple testing procedure is developed for selecting variables and is shown to control the false discovery

rate. These methods are illustrated by an analysis of a yeast dataset in order to identify genes that are associated with growth in the presence of hydrogen peroxide.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iii

ABSTRACT iv

LIST OF TABLES ix

LIST OF ILLUSTRATIONS xi

CHAPTER 1 : INTRODUCTION 1

1.1 Human Microbiome and its Relation to Human Health 1

1.2 Analysis of Microbiome Compositional Data 2

1.3 Integrative Analysis of Multi-omics Data 4

1.4 Organization of the Thesis 4

CHAPTER 2 : GENERALIZED LINEAR MODELS WITH LINEAR CONSTRAINTS FOR MICROBIOME COMPOSITIONAL DATA 6

2.1 Introduction 6

2.2 GLMs with Linear Constraints for Microbiome Compositional Data 7

2.3 De-biased Estimator and its Asymptotic Distribution 11

2.4 Applications to Gut Microbiome Studies 16

2.5 Simulation Studies 20

2.6 Discussion 23

CHAPTER 3 : POST-SELECTION INFERENCE FOR REGRESSION MODELS WITH LINEAR CONSTRAINTS, WITH AN APPLICATION TO MICROBIOME DATA 25

3.1 Introduction 25

3.2 Post-selection inference for high dimensional linear models with linear equality constraints 26

3.3 Optimization algorithm and computational details 34

3.4 Applications: UK twins data 35

3.5 Simulation studies 38

3.6 Discussion 42

CHAPTER 4 : HYPOTHESIS TESTING IN HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES . 44

4.1 Introduction 44

4.2 IV Models and Proposed Methodology 45

4.3 Theoretical Results 50

4.4 Simulations 54

4.5 Application to a Yeast Data Set 58

4.6 Discussion 61

CHAPTER 5 : DISCUSSION 63

CHAPTER A : PROOFS AND ADDITIONAL SIMULATIONS 64

A.1 Proofs for Chapter 2 64

A.2 Additional simulation results for Chapter 2 70

A.3 Proofs for Chapter 3 71

A.4 Proofs for Chapter 4 75

A.5 Additional Simulation Studies Chapter 4 86

APPENDIX 64

BIBLIOGRAPHY 90

LIST OF TABLES

TABLE 2.1 : Selected bacterial species and their corresponding phylum, estimated coefficients (standard errors in the parenthesis) and 95% confidence intervals. Model 1: regression analysis with the compositions of 77 bacterial species as covariates. Model 2: regression analysis with the subcompositions of bacterial species that belong to different genera as covariates. 18

TABLE 2.2 : True /False positive rates of the significant variables selected by the 95% confidence interval using multiple, one, no and misspecified constraints. $p = 50, 100$ and $n = 50, 100, 200, 500$ are considered. 23

TABLE 3.1 : Estimates and confidence intervals of the regression coefficients using different methods applying to the UK twins dataset. The computation for the post-selection confidence interval for *Blautia* fails to converge and hence is not shown. 38

TABLE 3.2 : Average coverage probabilities of the post-selection confidence intervals obtained by conditioned on selected model and signs or conditioned only on the selected model. Three different contraints on coefficients and two different tuning parameters are considered. For each setting, the first row represents the confidence intervals calculated assuming the variance parameter σ^2 is known and the second row represents the confidence intervals calculated when the variance parameter σ^2 is estimated. For $p = 500$, see text for selection of the tuning parameters λ_1 and λ_2. 41

TABLE 3.3 : Average length of the post-selection confidence intervals obtained by conditioned on the selected model and signs or conditioned only on the selected model. Three different contraints on coefficients and two different tuning parameters are considered. For each setting, the first row represents the confidence intervals calculated assuming the variance parameter σ^2 is known and the second row represents the confidence intervals calculated when the variance parameter σ^2 is estimated. For $p = 500$, see text for selection of the tuning parameters λ_1 and λ_2. 42

TABLE 4.1 : Simulation results based on 500 replications. The eFDR and power for multiple testing procedure based on IV regression and naive high dimensional linear regression for different combinations of (n, p, q) and different α levels. 57

TABLE 4.2 : Simulation results based on 500 replications. The eFDV and power for multiple testing procedures based on IV regression and naive high dimensional linear regression for different combinations of (n, p, q) and and different k levels. 57

TABLE 4.3 : Results from analysis of yeast growth yield data. Table shows the selected genes using single test statistics ($p < 0.05$) and multiple testing procedure with FDR< 0.10 and FDV< 2(marked by $*$). The gene names and estimated regression coefficients and refitted values are listed. 60

TABLE A.1 : Comparisons of parameter estimates and CIs using two different methods of replacing zeros. Selected bacteria and their estimated coefficients (standard errors in the parenthesis) and 95% confidence intervals. 70

TABLE A.2 : Sensitivity analysis results based on 500 replications. The eFDR and eFDV for multiple testing procedures based on IV regression for different combinations of (n, p, q) and different α, k levels and weak and strong direct effects. 89

LIST OF ILLUSTRATIONS

FIGURE 1.1 : Scatter plots of two bacterial taxa for control and case group. The first panel (A) is the true abundances, the second panel (B) is the observed compositions and the third panel (C) is the relative abundance comparing to a reference taxon. 2

FIGURE 2.1 : Analysis of the IBD microbiome data using a single constraint on regression coefficients. (a) Lasso estimates, de-biased estimates and 95% confidence intervals of the regression coefficients. Species selected based on the CIs are annotated. (b) Boxplots of log-relative abundances of the five identified species. The red and blue boxplots correspond to controls and case samples, respectively. (c) Fitted probability plot. (d) Selection stability plot. 17

FIGURE 2.2 : Analysis of the IBD microbiome data using multiple constraints. (a) Lasso estimates, de-biased estimates and 95% confidence intervals of the regression coefficients. Species selected based on the CIs are annotated. (b) Boxplots of log-relative abundances of the six identified species. The red and blue boxplots correspond to controls and cases samples respectively. (c) Fitted probability plot. (d) Selection stability plot. 19

FIGURE 2.3 : Coverage probabilities and length of confidence intervals based on 100 simulations for $p = 50$ ((a) and (b)) and $p = 100$ ((c) and (d)) and $n = 50, 100, 200, 500$ (separated by vertical dashed lines). 22

FIGURE 3.1 : Stability selection plot for UK twin data based on Lasso with a zero sum constraint of the regression coefficients. 36

FIGURE 3.2 : Estimates and confidence intervals of the regression coefficients for UK twins dataset. (a): Model with linear constraints; (b): Model without linear constraints 37

FIGURE 4.1 : Box plots of the empirical type I errors for single hypothesis testing based on IV regression and naive Lasso regression under different settings for α-level of 0.05 (left) and 0.01 (right). 56

FIGURE 4.2 : Analysis of yeast eQTL data sets, showing the histogram of the number of genotypes associated with each gene expression (left plot) and the histogram of the estimated regression coefficients in the first stage ($\widehat{\mathbf{\Gamma}}$) based on Lasso regressions (right plot). 58

FIGURE 4.3 : Scatter-plots of the fitted versus the observed yeast growth yield. (a): refitted model using the estimated expression levels of the 15 genes selected by our proposed method; (b): refitted model using expression levels of 34 genes selected using naive test; (c): the refitted model using expression levels of genes selected based on Lasso. 60

FIGURE A.1 : QQ-plots of the test statistic $\widehat{T}_i$ based on the two-stage IV model for several randomly selected variables to demonstrate the validity of its asymptotic distribution. The panels in the first and second row correspond to selected variables whose true value are zero and the third row are variables that are not zero. For different columns, (a)(d)(g), (b)(e)(h) and (c)(f)(i) correspond to different (n, p, q) values as $(200, 100, 100)$, $(400, 200, 200)$ and $(200, 500, 500)$. 87

FIGURE A.2 : Selected QQ-plots of the test statistics $\widehat{T}_i$ developed for fitting naive high dimensional regression models. The panels in the first and second row corresponds to selected variables whose true value are zero and the third row are variables that are not zero. For different columns, (a)(d)(g), (b)(e)(h) and (c)(f)(i) correspond to different (n, p, q) values as $(200, 100, 100)$, $(400, 200, 200)$ and $200, 500, 500)$. 88

CHAPTER 1

INTRODUCTION

1.1. Human Microbiome and its Relation to Human Health

Human microbiome consists of all living microorganisms that are in and on human body. They could be found on the skin, in the gut, oral cavity, lung etc. These microorganisms that colonize human body form a complex ecosystem, which is rich in both the amount of cells and species-diversity (Gilbert et al., 2018). The number of genes they carry as a result is far more than the number of human genes. There are also within and interpersonal heterogeneity of the distributions of the bacteria. For example, the dominate phylum on the skin is *Actinobacteria* while in the esophagus *Firmicutes* has a relative abundance of more than 50%. The compositions of microbiome on the same anatomical site may also be different (Cho and Blaser, 2012). The study of human microbiome dates back to nineteenth century but not until the recent development of DNA-based analysis, could people gain more insights about the compositions and functions of the bacteria and how they interact with host in the content of diets and environmental factors.

Large-scale microbiome studies such as the NIH Human Microbiome Project (HMP), have shown that this complex ecosystem has huge impact on human health through multiple ways, including exchanging molecules with human cells, interacting with human genetics and interacting with immune systems etc (Research Network Consortium, 2019). For example, studies reveals the associations between gut microbiome compositions and inflammatory bowel diseases including Crohn's disease and ulcerative colitis (Lloyd-Price et al., 2019). Researches also demonstrate the contribution of microbiome to cancer (Schwabe and Jobin, 2013), cardiovascular disease (Jie et al., 2017), cystic fibrosis (Surette, 2014) and many other microbiome-linked health states. This ecosystem also has impact on brain through exchanging chemicals among gut microbiota, immune cells and Vagus nerve (Cryan and Dinan, 2012). The interaction between gut microbiota and innate immune system also contributes to obesity, Type I diabetes, non-alcoholic fatty liver diseases etc (Thaiss et al., 2016). Literatures also demonstrate how microbiota interact with the brain through the gut-brain axis and relate to anxiety and depression (Carabotti et al., 2015; Foster and Neufeld, 2013). With the evidences that human microbiome is closely related to human health, it is important to further

investigate the specific roles of microbiome in initiation and progression of diseases.

1.2. Analysis of Microbiome Compositional Data

Advanced sequencing technologies such as 16S sequencing and shotgun metagenomic sequencing, provide powerful methods to quantify the relative abundance of bacterial taxa in or on human body of a large set of individuals (Xia et al., 2011). Since only the relative abundances are available, the resulting data are compositional with a unit sum constraint. The compositional nature of the data requires additional care in statistical analysis, including linear regression analysis (Shi, Zhang, and Li, 2016) and two-sample tests (Cao, Lin, and Li, 2018).

We conduct a simple simulation to illustrate the impact of observing the compositional data only. We simulate the true abundances of bacterial taxa for control and case groups. The difference in these two groups is the abundance of the second taxon (labeled as "bac2" in figure 1.1). The rest of the taxa share the same distribution between the two groups. As shown in figure 1.1 (A), the

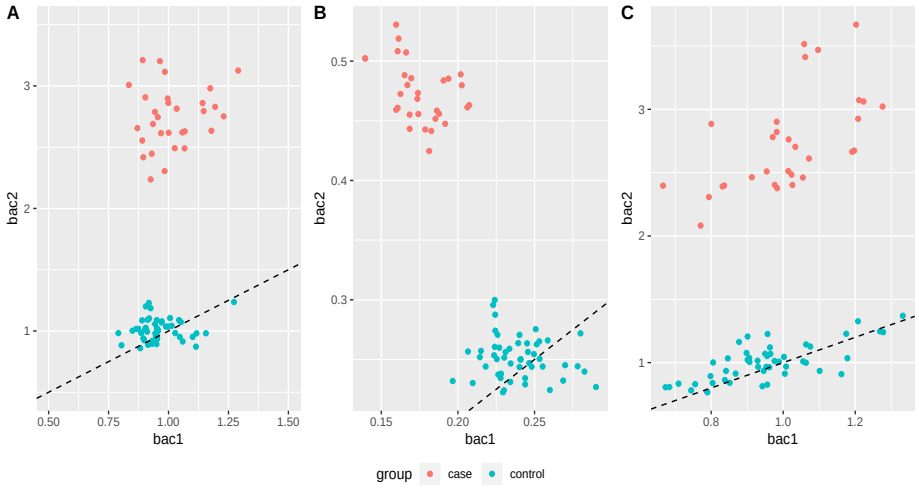


Figure 1.1: Scatter plots of two bacterial taxa for control and case group. The first panel (A) is the true abundances, the second panel (B) is the observed compositions and the third panel (C) is the relative abundance comparing to a reference taxon.

true abundance of the "bac1" is the same between the two groups. However, as we mentioned previously, the true abundances are never observed and the observed compositions are shown in figure 1.1 (B). A problem naturally rises when we conduct simple hypothesis tests comparing the mean compositions of taxa between the two groups. From the observation data only we would detect differential abundance in both "bac1" and "bac2" while we know from the simulation setup

that only “bac2” is different. This indicates that statistical analysis may lead to false information without taking care of the compositional nature of the data. As suggested by Aitchison (1982), a way to account for this issue is to use the ratio of bacterial taxa comparing to a reference bacteria. As demonstrated in figure 1.1 (C), the pattern remains normal again once using the reference taxon. In this case one will not detect a difference in “bac1” and would only find difference in “bac2”. A reference bacteria is usually defined as a taxon that has the same distribution in the population of interests. In practice, the selection of the reference group is a problem that needs extra care.

In the context of regression framework, Aitchison and Bacon-shone (1984) proposed the log-contrast linear model:

$$y = \sum_{i=1}^{p-1} \beta_i \log \frac{X_i}{X_p} + \varepsilon,$$

where the p-th bacteria is assumed to be the reference group. Instead of using the abundances of the bacteria as covariates, Aitchison used the log-ratio. Statistical methods related to linear regression could naturally applied to this model.

As an extension, Lin et al. (2014) proposed the log-contrast linear regression model with constraints.

$$y = \sum_{i=1}^{p} \beta_i \log X_i + \varepsilon, \text{subject to } \sum_{i=1}^{p} \beta_i = 0.$$

From a mathematical point of view, these two models are equivalent. But applying the constraints could bring extra benefits when imposing regularization in high-dimensional settings. Removing the reference group brings symmetrical structure to the model, which leads to the property that the model is scale invariant, permutation invariant and selection invariant (see Lin et al. (2014) for more details). The necessity and benefits of applying sum-zero constraints have been widely discussed in past literatures. Imposing such constraints, however, brings extra challenges in statistical analysis. This motivates us to develop novel methods for analyzing models with linear constraints.

1.3. Integrative Analysis of Multi-omics Data

Along with the metagenomics data, omics data from other sources are also closely related to human health. The joint analysis of gene expression and genetic variants data is one of the most important methods to reveal the link between human genes and phenotypes of interests. Among various methods, association analysis between gene expression and phenotype such as differential gene expression analysis has been widely reported. Such studies have shown that gene expressions are associated with many common human diseases, such as liver disease (Romeo et al., 2008; Speliotes et al., 2011) and heart failure (Liu et al., 2015). However, there are possibly many unmeasured factors that affect both gene expressions and phenotypes of interest (Hoggart et al., 2003; Leek and Storey, 2007). The existence of such unmeasured confounding variables can cause correlation between the error term and one or some of the independent variables and lead to identifying false associations. Particularly, the independence assumption between gene expressions and errors are required in linear regression in order to obtain valid statistical inference of the effects of gene expressions on phenotype. If this assumption is violated, standard methods can lead to biased estimates (Fan and Liao, 2014; Lin, Feng, and Li, 2015). To account for the existence of such unmeasured confounding variables, certain novel statistical methods are needed.

1.4. Organization of the Thesis

My thesis mainly focused on the analysis of metagenomics data and joint analysis of genetic variates, gene expression and phenotypical data. In Chapter 2[1], we developed a generalized linear model with linear constraints to study the association between microbiome compositions and a disease's risk. A group of linear constraints on the regression coefficients are imposed to account for the compositional nature of the data and to achieve subcompositional coherence. The regression coefficients were estimated by a constrained L1-penalized likelihood method computed via a generalized accelerated proximal gradient algorithm. A de-biased procedure was developed to obtain asymptotically unbiased and normally distributed estimates, which leads to valid confidence intervals of the regression coefficients. Simulation results showed the correctness of the coverage probability of the confidence intervals and smaller variances of the estimates when the appropriate linear constraints are imposed. Application of this method on the PLEASE study identified several

[1] This part of the thesis is based on paper Lu, Shi, and Li (2019)

gut bacterial species that are associated with the risk of IBD.

In Chapter 3[2], we considered the post-selection inference method for models with linear equality constraints. We developed methods for constructing the confidence intervals for the selected non-zero coefficients chosen by a Lasso-type estimator with linear constraints. These confidence intervals were proofed to have desired coverage probabilities when conditioned on the selected model. Simulations were conducted to demonstrate the validity of our method in providing valid confidence intervals after variable selection step. We applied this procedure to a UK Twins microbiome dataset identifying several key bacterial genera whose compositions are associated with chronological age.

Finally, the last chapter this dissertation presents a method for inference of high dimensional instrumental variable regression[3]. Gene expression and phenotype association can be affected by potential unmeasured confounders from multiple sources, leading to biased estimates of the associations. Since genetic variants largely explain gene expression variations, they can be used as instruments in studying the association between gene expressions and phenotype in the framework of high dimensional instrumental variable (IV) regression. However, because the dimensions of both genetic variants and gene expressions are often larger than the sample size, statistical inferences such as hypothesis testing for such high dimensional IV models are not trivial and have not been investigated in literature. The problem is more challenging since the instrumental variables (e.g., genetic variants) have to be selected among a large set of genetic variants. We consider the problem of hypothesis testing for sparse IV regression models and present methods for testing single regression coefficient and multiple testing of multiple coefficients, where the test statistic for each single coefficient is constructed based on an inverse regression. A multiple testing procedure is developed for selecting variables and is shown to control the false discovery rate. Simulations are conducted to evaluate the performance of our proposed methods. These methods are illustrated by an analysis of a yeast dataset in order to identify genes that are associated with growth in the presence of hydrogen peroxide.

[2]This part of the thesis is based on the submitted paper Lu and Li (2020b).
[3]This part of the thesis is based on the submitted paper Lu and Li (2020a).

CHAPTER 2

GENERALIZED LINEAR MODELS WITH LINEAR CONSTRAINTS FOR MICROBIOME COMPOSITIONAL DATA

2.1. Introduction

In this chapter, we considered the general regression problems where the covariates include composition of a set of bacterial taxa. The goal of such regression analysis is to identify a subset of the bacteria whose relative abundances are associated with a response variable. The main challenges of analyzing compositional data are to account for the unit sum structure and to achieve subcompositional coherence (Aitchison, 1982), which requires that the same results are obtained regardless of the way the data is normalized into proportions based on the whole compositions or only a subcomposition. To explore the association between a response and the compositional data, Aitchison and Bacon-shone (1984) proposed a linear log-contrast model to link the response and the log of the compositional data for continuous and normally distributed response variable. This model was further extended by Lin et al. (2014) and considered variable selection problem by a ℓ_1-penalized estimation procedure. To achieve subcompositional coherence, Shi, Zhang, and Li (2016) extended the linear regression model by imposing a set of linear constraints. Lin et al. (2014) and Shi, Zhang, and Li (2016) showed the connection between these models and the regression models with centered log-ratio transformed proportions (Aitchison and Bacon-shone, 1984) as covariates and showed that the logarithmic transformation of the proportions is necessary for subcompositional coherence.

In this chapter, the generalized linear regression models (GLMs) with linear constraints in the regression coefficients were proposed for microbiome compositional data, where a group of linear constraints were imposed to achieve subcompositional coherence. In order to identify the bacterial taxa that are associated with the response, a penalized estimation procedure for the regression coefficients via a ℓ_1 penalty was introduced. To solve the computational problem, a generalized accelerated proximal gradient method was developed, which extended the standard accelerated proximal gradient method (Nesterov, 2013) to account for linear constraints. The proposed method could efficiently solve the optimization problem of minimizing the penalized negative log-likelihood

subjects to a group of linear constraints.

Previous works on the inference of Lasso for the generalized linear models include Bühlmann and Van De Geer (2011), which provided properties of the penalized estimates such as bound for ℓ_1 loss and oracle inequality. However, the methods cannot be applied directly to the setting with linear constraints. Furthermore, it is known that the ℓ_1 penalized estimates are biased and do not have a tractable asymptotic distribution. In order to correct such biases, works have been done for the Lasso estimate, including Zhang and Zhang (2014), who proposed a low-dimensional projection estimator to correct the bias and Javanmard and Montanari (2014), who used a quadratic programming method to carry out the task. Geer et al. (2014) considered an extension to generalized linear models. However, these methods still cannot be directly applied to our problem due to the linear constraints.

In order to make statistical inference on the regression coefficients, we propose a bias correction procedure for GLMs with linear constraints by extending the method of Javanmard and Montanari (2014). Such a debiased procedure provided asymptotically unbiased and normal distributed estimates of the regression coefficients, which can be used to construct confidence intervals. Our simulations results showed the correctness of the coverage probability of the confidence intervals and smaller variances of the estimates when the appropriate linear constraints are imposed.

2.2. GLMs with Linear Constraints for Microbiome Compositional Data

2.2.1. GLMs with linear constraints

Consider a microbiome study with outcome y_i and a p dimensional compositional covariates $\mathbf{X}_i = (x_{i1}, \cdots, x_{ip})$ with the unit sum constraint $\sum_j x_{ij} = 1$ for $i = 1, \cdots, n$, where x_{ij} represents the relative abundance of the jth taxon of the ith samples. To account for compositional nature of the covariates, Lin et al. (2014) proposed the linear model with constraint:

$$y_i = \mathbf{Z}_i^\top \boldsymbol{\beta} + \epsilon_i, \text{ subject to } \mathbf{1}^\top \boldsymbol{\beta} = 0, \tag{2.1}$$

where $\mathbf{Z}_i = \{\log(x_{ij})\} \in \mathbb{R}^{n \times p}$ and $\mathbf{1} = (1, 1, \ldots, 1)^\top$. Such a zero-sum constraint ensured that the regression coefficients are independent of an arbitrary scaling of the basis count from which a composition is obtained, and remain unaffected by correctly excluding some or all of the zero

components (Lin et al., 2014). This subcompositional coherence property is one of the principals of compositional data analysis (Aitchison, 1982). Because of the linear constraints, the interpretation of a given β_j has to be in the context of other none-zero βs.

Shi, Zhang, and Li (2016) further developed this method to allow r linear constraints by specifying the $p \times r$ constraint matrix C. For example, if we are interested in studying whether the composition of taxa that belong to a given taxon at a higher rank is associated with the response, in which case subcompositions of taxa under a given high rank are calculated. Suppose r taxa at a given rank are considered with m_g taxa at the lower rank that belong to taxon g. We define the subcomposition of these m_g taxa, which is simply a subvector of the p dimensional compositions, rescaled so that its components sum to unity. Specifically, let X_{gs} be the relative abundance of the sth taxon that belong to the gth taxon at a higher rank, for $g = 1, \cdots, r$, $s = 1, \cdots, m_g$ such that

$$\sum_{s=1}^{m_g} X_{gs} = 1, \text{for } g = 1, \cdots, r.$$

Suppose we have n samples and let $n \times m_g$ matrix $\mathbf{X}_g$ represents n samples of the subcomposition of m_g taxa. Shi, Zhang, and Li (2016) proposed to associate the subcompositions to a continuous response Y via the following linear model,

$$\begin{aligned} Y &= \sum_{g=1}^{r} \mathbf{Z}_g \beta_g + \epsilon, \\ \text{such that } \mathbf{1}_{m_g}^\top \beta_g &= \sum_{s=1}^{m_g} \beta_{gs} = 0 \text{ for } g = 1 \cdots, r, \end{aligned} \tag{2.2}$$

where $\mathbf{Z}_g = (Z_{g1}, \ldots, Z_{gm_g}) = (\log X_{g1}, \ldots, \log X_{gm_g}) \in \mathbb{R}^{n \times m_g}$, and $\beta_g = (\beta_{g1}, \cdots, \beta_{gm_g})^\top$. For a given group of species that belong to the g-th genus, the regression coefficient β_{gs} has to be interpreted together with other species that belong to the g-th genus. In other words, the expected response depends on the subcomposition via the parameter vector β_g, not just simply a single component of β_g. The parameter vector β_g determines how the expected response changes as the subcomposition moves away from the center of the $m_g - 1$ dimensional simplex.

For general outcome, we extended the linear model (2.1) to the generalized linear model with its

density function specified as

$$
\begin{aligned}
f(y_i|\boldsymbol{\beta}, \mathbf{Z}_i) &= h(y_i) \exp\left\{\eta_i y_i - A(\eta_i)\right\}, \quad \eta_i = \mathbf{Z}_i^\top \boldsymbol{\beta}, \\
\mathbb{E} y_i = \nabla_{\eta_i} A(\eta_i) &\equiv \mu(\boldsymbol{\beta}, \mathbf{Z}_i), \quad \mathrm{Var} y_i = \nabla^2_{\eta_i} A(\eta_i) \equiv v(\boldsymbol{\beta}, \mathbf{Z}_i),
\end{aligned}
\tag{2.3}
$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots \beta_p)^\top \in \mathbb{R}^p$ and satisfies $C^\top \boldsymbol{\beta} = 0$, and $\mathbf{Z}_i^\top = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$. For simplicity, we assumed the intercept being zero, though our formal justification will allow for an intercept. Although Model (2.3) does not explicitly include other covariates, it can handle covariates by simply including columns of all zeros in the C matrix that correspond to these covariates. All the results in the rest of the Chapter still hold with covariates. For binary outcome and logistic regression, we have

$$
A(\eta) = \log(1 + e^\eta), \mu(\boldsymbol{\beta}, \mathbf{Z}_i) = \frac{e^{\mathbf{Z}_i^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{Z}_i^\top \boldsymbol{\beta}}}, v(\boldsymbol{\beta}, \mathbf{Z}_i) = \frac{e^{\mathbf{Z}_i^\top \boldsymbol{\beta}}}{(1 + e^{\mathbf{Z}_i^\top \boldsymbol{\beta}})^2}.
$$

2.2.2. ℓ_1 penalized estimation with constraints

The log-likelihood function based on model (2.3) is given by

$$
\ell(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{Z}) = \sum_{i=1}^n \log h(y_i) + \mathbf{Y}^\top \mathbf{Z} \boldsymbol{\beta} - \sum_{i=1}^n A(\mathbf{Z}_i^\top \boldsymbol{\beta}), \tag{2.4}
$$

with score function and information matrix:

$$
\nabla_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{Z}) = \{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}, \mathbf{Z})\}^\top \mathbf{Z}, \ \nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{Z}) = -\mathbf{Z}^\top \mathbf{V}(\boldsymbol{\beta}, \mathbf{Z}) \mathbf{Z},
$$

where $\mathbf{V}(\boldsymbol{\beta}, \mathbf{Z}) = \mathrm{diag}\{v(\boldsymbol{\beta}, Z_1), \ldots, v(\boldsymbol{\beta}, Z_n)\}$. The constraints on $\boldsymbol{\beta}$ are given by $C^\top \boldsymbol{\beta} = 0$, where C is a $p \times r$ matrix. Without lose of generality, the columns of C are assumed to be orthonormal. Define $P_C = CC^\top$, $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - P_C)$ and $\widetilde{Z}_i = (\mathbf{I}_p - P_C)\mathbf{Z}_i$, then under the constraints of $C^\top \boldsymbol{\beta} = 0$, all the $\mathbf{Z}$ and Z_i can be replaced by $\widetilde{\mathbf{Z}}$ and $\widetilde{Z}_i$ because $\mathbf{Z}\boldsymbol{\beta} = \widetilde{\mathbf{Z}}\boldsymbol{\beta}$.

In high-dimensional settings, $\boldsymbol{\beta}$ is assumed to be s-sparse, where $s = \#\{i : \boldsymbol{\beta}_i \neq 0\}$ and $s = o(\sqrt{n}/\log p)$. The ℓ_1 penalized estimates of $\boldsymbol{\beta}$ is given as the solution to the following problem:

$$
\hat{\boldsymbol{\beta}}^n = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \left[-\frac{1}{n} \left\{ \mathbf{Y}^\top \widetilde{\mathbf{Z}} \boldsymbol{\beta} - \sum_{i=1}^n A(\widetilde{\mathbf{Z}}_i^\top \boldsymbol{\beta}) \right\} + \lambda ||\boldsymbol{\beta}||_1 \right] \text{ subject to } C^\top \boldsymbol{\beta} = 0, \tag{2.5}
$$

where λ is a tuning parameter.

2.2.3. Generalized accelerated proximal gradient method

Due to the linear constraints in the optimization problem (2.5), the standard coordinate descent algorithm cannot be applied directly. We develop a generalized accelerated proximal gradient algorithm. Specifically, define g, h as following

$$g(\boldsymbol{\beta}) = -\frac{1}{n}\left\{Y^\top \widetilde{\mathbf{Z}}\boldsymbol{\beta} - \sum_{i=1}^{n} A(\widetilde{Z}_i^\top \boldsymbol{\beta})\right\}, \quad h(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1$$

so the optimization problem (2.5) becomes

$$\hat{\boldsymbol{\beta}}^n = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{g(\boldsymbol{\beta}) + h(\boldsymbol{\beta})\} \text{ subject to } C^\top \boldsymbol{\beta} = 0.$$

Since g is convex and differentiable and h is convex, the standard accelerated proximal gradient method (Nesterov, 2013) is given by the following iterations:

$$\begin{aligned}
\boldsymbol{\beta}^{(k)} &= \mathbf{prox}_{t_k h}\left(y^{(k-1)} - t_k \nabla g(y^{(k-1)})\right), \\
y^{(k)} &= \boldsymbol{\beta}^{(k)} + \frac{k-1}{k+r-1}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}),
\end{aligned}$$

where t_k is the step size in the k-th iteration and r is a friction parameter. The proximal mapping of a convex function h, which is the key ingredient of this algorithm, is defined as:

$$\mathbf{prox}_h(x) = \underset{u}{\operatorname{argmin}}\left\{h(u) + \frac{1}{2}||x-u||_2^2\right\}.$$

We generalize this method to handle the linear constraints. Denote $S_C = \{\boldsymbol{\beta} \in \mathbb{R}^p \mid C^\top \boldsymbol{\beta} = 0\}$, a linear subspace of $\mathbb{R}^p$. The generalized accelerated proximal gradient method becomes

$$\boldsymbol{\beta}^{(k)} = \underset{\boldsymbol{\beta} \in S_C}{\operatorname{argmin}}\left\{\lambda t_k \|\boldsymbol{\beta}\|_1 + \frac{1}{2}||y^{(k-1)} - t_k \nabla g(y^{(k-1)}) - \boldsymbol{\beta}||_2^2\right\}, \tag{2.6}$$

$$y^{(k)} = \boldsymbol{\beta}^{(k)} + \frac{k-1}{k+r-1}(\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}). \tag{2.7}$$

The minimization of (2.6) can be solved by soft thresholding and projection:

$$\boldsymbol{\beta}^{(k)} = \Pi_{S_C}\left(S_{t_k\lambda}\left(y^{(k-1)} - t_k\nabla g(y^{(k-1)})\right)\right),$$

where linear operator $\Pi_{S_C}(u)$ projects u onto space S_C. Since $C^\top$ is a matrix and can be regarded as a linear mapping from $R^p \mapsto R^r$, we have $S_C = \ker(C^\top)$. Denote $u_p = \Pi_{S_C}(u)$, we have:

$$C^\top(u - u_p) = C^\top u.$$

So $u - u_p$ is given by least square estimates: $u - u_p = (CC^\top)^\dagger CC^\top u$, where $A^\dagger$ is the Moore-Penrose pseudo inverse of a matrix A. Hence,

$$\Pi_{S_C}(u) = u - (CC^\top)^\dagger CC^\top u.$$

The step size t_k can be fixed or chosen by line search. The procedure of line search consists of the following iterations: we start with a initial $t = t_{k-1}$ and repeat $t = 0.5t$ until the following inequality holds:

$$g(y - tG_t(y)) \leq g(y) - t\nabla g(y)^\top G_t(y) + \frac{t}{2}\|G_t(y)\|_2^2,$$

where $y = y^{(k-1)}$. For the friction parameter r, Su, Boyd, and Candes (2014) suggested that $r > 4.5$ will lead to fast convergence rate and is set to 10.

2.3. De-biased Estimator and its Asymptotic Distribution

2.3.1. A de-biased Estimator

Since $\widehat{\boldsymbol{\beta}}^n$ in equation (2.5) is a biased estimator for $\boldsymbol{\beta}$ due to ℓ_1 penalization, we propose the following de-biased procedure, detailed as Algorithm 1, to obtain asymptotically unbiased estimates of $\boldsymbol{\beta}$. This algorithm has the same general steps but differs from that for linear models (Shi, Zhang, and Li, 2016) in two aspects: (1) the $\widehat{\boldsymbol{\Sigma}}$ matrix defined in our algorithm (Step 2) is different from that for linear models, which is simply the sample covariance matrix. The matrix $\widehat{\boldsymbol{\Sigma}}$ is the information matrix that involves the Lasso-estimated regression coefficients, which makes the theoretical analysis harder. (2) The final de-biased estimator (Step 6) is different, where the mean of $\mathbf{Y}$ is a non-linear

function of the Lasso-estimated coefficients.

Algorithm 1 Constructing a de-biased estimator

Input: $\mathbf{Y}$, $\mathbf{Z}$, $\widehat{\boldsymbol{\beta}}^n$, and γ. **Output:** $\widehat{\boldsymbol{\beta}}^u$

1: Let $\widehat{\boldsymbol{\beta}}^n$ be the regularized estimator from optimization problem (2.5).
2: Set $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - P_C)$, $\widehat{\boldsymbol{\Sigma}} = (\widetilde{\mathbf{Z}}^\top \mathbf{V}(\hat{\boldsymbol{\beta}}^n, \widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}})/n$.
3: **for** $i = 1, 2, \ldots, p$ **do**
4: Let m_i be a solution of the convex program:

$$\begin{aligned} &\text{minimize } m^\top \widehat{\boldsymbol{\Sigma}} m \\ &\text{subject to } ||\widehat{\boldsymbol{\Sigma}} m - (\mathbf{I}_p - P_C)e_i||_\infty \leq \gamma. \end{aligned} \tag{2.8}$$

where $e_i \in \mathbb{R}^p$ is the vector with one at the i-th position and zero everywhere else.
5: Set $M = (m_1, \ldots, m_p)^\top$, set

$$\widetilde{M} = (\mathbf{I}_p - P_C)M. \tag{2.9}$$

6: Define the estimator $\widehat{\boldsymbol{\beta}}^u$ as follows:

$$\widehat{\boldsymbol{\beta}}^u = \widehat{\boldsymbol{\beta}}^n + \frac{1}{n}\widetilde{M}\widetilde{\mathbf{Z}}^\top(\mathbf{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}^n, \widetilde{\mathbf{Z}})). \tag{2.10}$$

From the construction of $\widehat{\boldsymbol{\beta}}^u$, it is easy to check that $\widehat{\boldsymbol{\beta}}^u$ still satisfies $C^\top \widehat{\boldsymbol{\beta}}^u = 0$. To provide insights into this algorithm, using the mean value theorem, there exists $\boldsymbol{\beta}_i^0$ such that

$$\mu(\hat{\boldsymbol{\beta}}^n, \mathbf{Z}_i) - \mu(\boldsymbol{\beta}, \mathbf{Z}_i) = v(\boldsymbol{\beta}_i^0, \mathbf{Z}_i)\mathbf{Z}_i^\top(\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}),\ i = 1, 2, \ldots, n.$$

Define $\widehat{\boldsymbol{\Sigma}}^0 = (\widetilde{\mathbf{Z}}^\top \mathbf{V}(\boldsymbol{\beta}^0, \widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}})/n$, where $\mathbf{V}(\boldsymbol{\beta}^0, \widetilde{\mathbf{Z}}) = \mathrm{diag}\{v(\boldsymbol{\beta}_1^0, Z_1), \ldots, v(\boldsymbol{\beta}_n^0, Z_n)\}$, we have

$$\begin{aligned} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^u - \boldsymbol{\beta}\right) &= \sqrt{n}\left\{(\mathbf{I}_p - P_C) - \widetilde{M}\widehat{\boldsymbol{\Sigma}}^0\right\}(\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}) + \frac{1}{\sqrt{n}}\widetilde{M}\widetilde{\mathbf{Z}}^\top(Y - \boldsymbol{\mu}(\boldsymbol{\beta}, \widetilde{\mathbf{Z}})), \qquad (*) \\ &\equiv \Delta + R. \end{aligned}$$

Define $\boldsymbol{\Sigma} = (\widetilde{\mathbf{Z}}^\top \mathbf{V}(\boldsymbol{\beta}, \widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}})/n$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathbb{E}\boldsymbol{\Sigma} = \mathbb{E}(v(\boldsymbol{\beta}, \widetilde{Z}_1)\widetilde{Z}_1\widetilde{Z}_1^\top)$, and suppose $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = V_{\boldsymbol{\beta}}\Lambda_{\boldsymbol{\beta}}V_{\boldsymbol{\beta}}^\top$ is the eigenvalue decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. Since $(V_{\boldsymbol{\beta}}, C)$ is full rank and orthonormal, we have

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = (V_{\boldsymbol{\beta}}, C)\begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix}(V_{\boldsymbol{\beta}}, C)^\top,\ \Omega_{\boldsymbol{\beta}} = (V_{\boldsymbol{\beta}}, C)\begin{pmatrix} \Lambda_{\boldsymbol{\beta}}^{-1} & 0 \\ 0 & 0 \end{pmatrix}(V_{\boldsymbol{\beta}}, C)^\top,$$

which implies

$$\mathbf{\Sigma}_{\boldsymbol{\beta}}\Omega_{\boldsymbol{\beta}} = (V_{\boldsymbol{\beta}}, C)\begin{pmatrix} \mathbf{I}_{p-r} & 0 \\ 0 & 0 \end{pmatrix}(V_{\boldsymbol{\beta}}, C)^{\top} = V_{\boldsymbol{\beta}}V_{\boldsymbol{\beta}}^{\top} = \mathbf{I}_p - P_C.$$

So Step 4 of Algorithm 1 approximates $\Omega_{\boldsymbol{\beta}}$ by rows.

2.3.2. Asymptotic distribution

In order to derive the asymptotic distribution of the de-biased estimator $\hat{\boldsymbol{\beta}}^u$, several regularity conditions are required.

C1. $\|\mathbf{I}_p - P_C\|_\infty \leq k_0$ for a constant k_0 that is free of p.

C2. The diagonal elements of $\mathbf{I}_p - P_C$ are greater than zero.

Conditions C1 and C2 have been used in Shi, Zhang, and Li (2016) and naturally hold in our setting as well. In addition, define $\widetilde{\mathbf{Z}}^* = D\widetilde{\mathbf{Z}}$, where $D \in \widetilde{D}_{ab}$ is defined as:

$$\widetilde{D}_{ab} = \{D \in \mathbb{R}^{n\times n} : \mathrm{diag}(d_1, d_2, \ldots, d_n), a \leq d_i \leq b,\, 0 < a < b\}.$$

For any matrix $A \in \mathbb{R}^{n\times m}$, the upper and lower restricted isometry property (RIP) constant of order k, $\delta_k^+(A)$ and $\delta_k^-(A)$, are defined as:

$$\delta_k^+(A) = \sup\left(\frac{\|A\alpha\|_2^2}{\|\alpha\|_2^2} :\ \alpha \in \mathbb{R}^m \text{ is } k\text{-sparse vector}\right),$$
$$\delta_k^-(A) = \inf\left(\frac{\|A\alpha\|_2^2}{\|\alpha\|_2^2} :\ \alpha \in \mathbb{R}^m \text{ is } k\text{-sparse vector}\right).$$

We assume the following RIP condition:

C3. $\inf_{\widetilde{D}_{01}}\left\{(3\tau-1)\delta_{2s}^-(\widetilde{\mathbf{Z}}^*/\sqrt{n}) - (\tau+1)\delta_{2s}^+(\widetilde{\mathbf{Z}}^*/\sqrt{n})\right\} \geq 4\tau\phi_0$ for some constant ϕ_0.

Condition C3 is slightly stronger than the one used for linear regression, which here we require the inequality holds uniformly over a set of matrices. The following theorem quantifies the difference between $\hat{\boldsymbol{\beta}}^n$ and $\boldsymbol{\beta}$ in ℓ_1 norm.

Theorem 1. *Let $\hat{\boldsymbol{\beta}}^n$ be the solution for (2.5), where $\boldsymbol{\beta}$ is s-sparse. If Conditions C1-C3 hold, and*

the tuning parameter $\lambda = \tau\tilde{c}\sqrt{(\log p)/n}$*, then*

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1 \geq \frac{s\lambda(k_0 + 1/\tau)}{\phi_0}\right) \leq 2p^{-c'},$$

where $c' = \dfrac{\tilde{c}^2}{2K^2} - 1$ *and* $K = \max_i \sqrt{(\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}}/n)_{i,i}}$.

In order to establish the asymptotic distribution of the de-biased estimates, additional conditions are required:

C4. There exist uniform constants C_{min} and C_{max} such that $0 < C_{\text{min}} \leq \sigma_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \leq \sigma_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \leq C_{\text{max}} < \infty$.

C5 $|\Omega_{\boldsymbol{\beta}}\Theta|_\infty < \infty$.

C6 The variance function $v(\boldsymbol{\beta}, \mathbf{Z}_i)$ satisfies Lipschitz condition with constant C;

C7 There exists a uniform constant $\kappa > 0$ such that $\|\Omega^{1/2}\widetilde{Z}_k\|_{\psi_2} \leq \kappa$ for all $k = 1, \ldots, n$.

In Condition C7, the sub-Gaussian norm of a random vector $Z \in \mathbb{R}^n$ is defined as

$$\|Z\|_{\psi_2} = \sup\left(\|Z^\top x\|_{\psi_2} : \ x \in \mathbb{R}^n \text{ and } \|x\|_2 = 1\right),$$

and the sub-Gaussian norm for a random variable X, is defined as

$$\|X\|_{\psi_2} = \sup_{q \geq 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}.$$

Conditions C4 and C7 are bounded eigenvalue assumption and bounded sub-Gaussian norm that are widely used in the literature of inference with respect to Lasso type estimator (Javanmard and Montanari, 2014; Shi, Zhang, and Li, 2016). Condition C5 eliminates extreme situations on $|\Omega_{\boldsymbol{\beta}}\Theta|_\infty$, which actually can be relaxed to hold in probability. For logistic regression, similar conditions are used in Ning and Liu (2017). Condition C6 is a Lipschitz condition on the variance function, which holds for many of the GLMs including logistic regression.

The following Lemma shows that if the tuning parameter γ in the optimization problem (2.8) is chosen to be $c\sqrt{(\log p)/n}$, then $\Omega_{\boldsymbol{\beta}}$ is in the feasible set with a large probability.

Lemma 1. *Denote $\Theta = \mathbb{E}\widetilde{Z}_1\widetilde{Z}_1^\top$. Suppose Conditions C1-C7 hold, then for any constant $c > 0$, the following inequality holds:*

$$\mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}\right) \leq 2p^{-c_1^{''}} + 2p^{-c_2^{''}},$$

where $c_1^{''} = c^2 C_{min}/(24e^2 C_{max}\kappa^4) - 2$ and $c_2^{''} = \hat{c}^2/2K^2 - 1$, with $\hat{c} = c\phi_0/C|\mathbf{\Omega}_{\boldsymbol{\beta}}\Theta|_\infty s(k_0\tau + 1)$ and $K = \max_i \sqrt{(\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}/n)_{i,i}}$.

The following Theorem provides the bound on $\|\Delta\|_\infty$ and also the asymptotic distribution of the de-biased estimates.

Theorem 2. *For $\Delta = \sqrt{n}\left\{(\mathbf{I}_p - P_C) - \widetilde{M}\widehat{\mathbf{\Sigma}}^0\right\}(\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta})$, if conditions C1-C7 hold, then for n large enough,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^u - \boldsymbol{\beta}) = R + \Delta,$$

where $R|\mathbf{Z} \to N(0, \widetilde{M}\widehat{\mathbf{\Sigma}}\widetilde{M}^\top)$ in distribution and $\|\Delta\|_\infty$ converge to 0 as $n, p \to \infty$, i.e.,

$$\mathbb{P}\left(\|\Delta\|_\infty > \frac{c\tilde{c}k_0(k_0\tau + 1)}{\phi_0} \cdot \frac{s\log p}{\sqrt{n}}\right) \leq 2p^{-c'} + 2p^{-c_1^{''}} + 6p^{-c_2^{''}},$$

for some constants c', $c_1^{''}$ and $c_1^{''}$ defined in Theorem 1 and Lemma 1.

This theorem allows us to obtain the confidence intervals for the regression coefficients, which can be used to further select the variables based on their statistical significance. Proofs of Lemma 1, Theorem 1 and Theorem 2 will be included in the supplementary materials.

2.3.3. Selections of tuning parameters

The tuning parameter λ in (2.5) can be selected using extended Bayesian information criterion (EBIC) (Chen and Chen, 2008), which is an extension of the standard BIC in high dimensional cases. Specifically, denote $\hat{\boldsymbol{\beta}}_\lambda^n$ the solution of (2.5) using λ as the tuning parameter, the EBIC is defined as

$$\text{EBIC}(\hat{\boldsymbol{\beta}}_\lambda^n) = -2\ell(\hat{\boldsymbol{\beta}}_\lambda^n|y, \mathbf{Z}) + \nu(\hat{\boldsymbol{\beta}}_\lambda^n)\log n + 2\nu(\hat{\boldsymbol{\beta}}_\lambda^n)\xi\log p,$$

where $\nu(s)$ is the number of none zero components of s. The choice of ξ is to solve for $p = n^\delta$ and set $\xi = 1 - 1/(2\delta)$ as suggested by Chen and Chen (2008). The optimal λ_{opt} is to minimize the

EBIC

$$\lambda_{\text{opt}} = \underset{\lambda}{\operatorname{argmin}}\, \text{EBIC}(\hat{\boldsymbol{\beta}}^n_\lambda). \tag{2.11}$$

over $\lambda_1, \lambda_2, \ldots$, with $\nu(\hat{\boldsymbol{\beta}}^n_{\lambda_i}) = i$. Tunning parameter γ in (2.8) is chosen as $0.01\lambda_{\text{opt}}$. Chen and Chen, 2012 showed that EBIC is variable selection consistent under generalized linear models.

2.4. Applications to Gut Microbiome Studies

The proposed method was applied to a study aiming at exploring the association between pediatric inflammatory bowel disease (IBD) and the gut microbiome conducted at the University of Pennsylvania (Lewis et al., 2015). This study collected the fecal samples of 85 IBD cases and 26 normal controls and conducted a metagenomic sequencing for each sample, resulting a total of 97 bacterial species identified. Among these bacterial species, 77 had non-zero values in at least 20 percent of the samples and were used in our analysis. The zero values in the relative abundance matrix were replaced with 0.5 times the minimum abundance observed, which is commonly used in microbiome data analyses (Cao, Lin, and Li, 2018; Kurtz et al., 2015). The composition of species is then computed after replacing the zeros and used to fit the regression model.

2.4.1. Identifying bacterial species associated with IBD

The proposed method was applied to the logistic regression analysis between IBD and log-transformed compositions of the 77 species as covariates. To be specific, let y be the binary indicator of IBD and $\log(X_k)$ is the logarithm of the relative abundance of the k-th species. We consider the following model

$$\text{logit}\,\{Pr(y=1)\} = \boldsymbol{\beta}_0 + \sum_{k=1}^{77} \boldsymbol{\beta}_k \log(X_k), \quad \text{where } \sum_{k=1}^{77} \boldsymbol{\beta}_k = 0.$$

Our goal was to identify the bacteria species that are associated with IBD and to evaluate how well one can predict IBD based on the gut microbiome composition.

Figure 2.1(a) shows the Lasso estimates, de-biased estimates and 95% confidence intervals of the regression coefficients in the model. Five bacteria were selected using our methods with the 95% CI not including zero, including *Prevotella_copri*, *Ruminococcus_bromii*, *Clostridium_leptum*, *Escherichia_coli* and *Ruminococcus_gnavus*. The estimated coefficients and the corresponding 95% CIs are summarized in Table 2.1. Among them, *Prevotella_copri*, *Ruminococcus_bromii*, *Clostrid-*

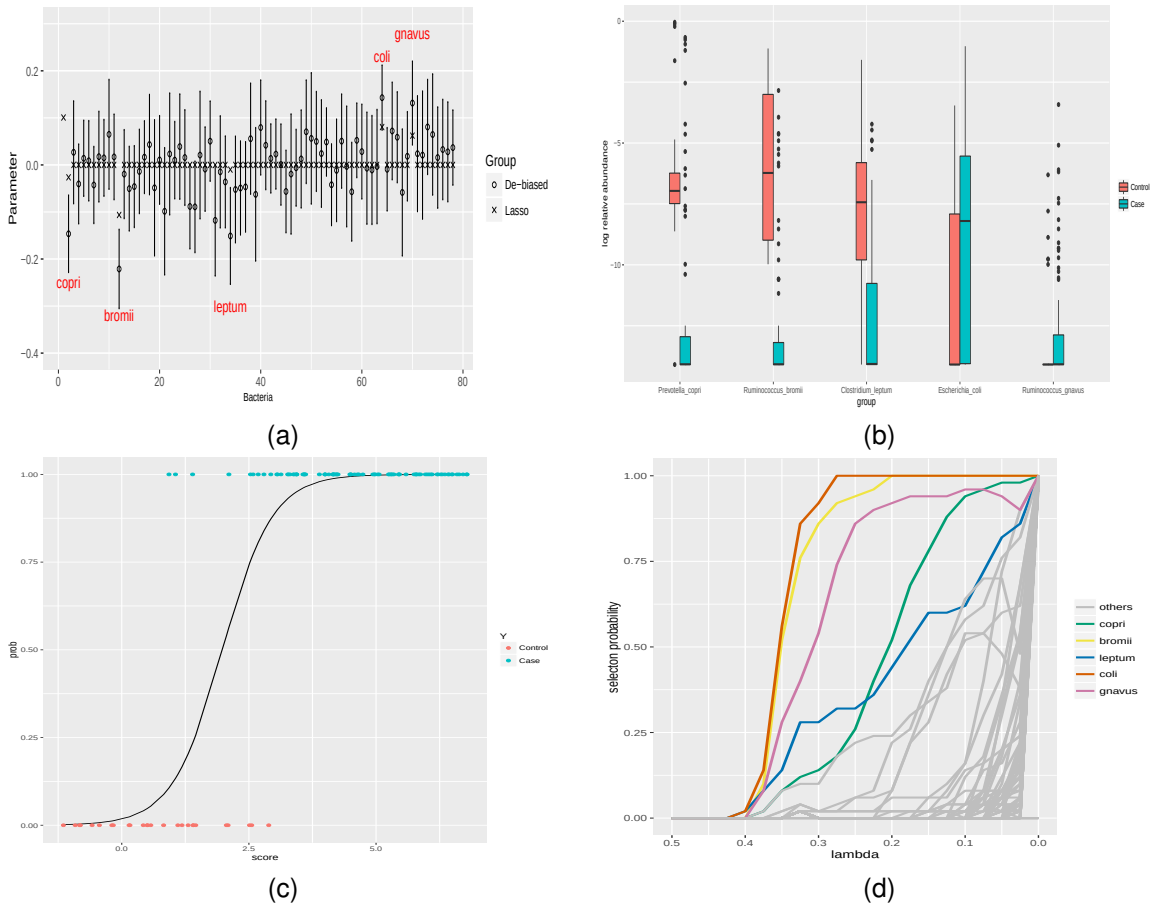


Figure 2.1: Analysis of the IBD microbiome data using a single constraint on regression coefficients. (a) Lasso estimates, de-biased estimates and 95% confidence intervals of the regression coefficients. Species selected based on the CIs are annotated. (b) Boxplots of log-relative abundances of the five identified species. The red and blue boxplots correspond to controls and case samples, respectively. (c) Fitted probability plot. (d) Selection stability plot.

ium_leptum are negatively associated with the risk of IBD, indicating possible beneficial effects on IBD. On the other hand, *Escherichia_coli* and *Ruminococcus_gnavus* are positively associated with IBD. Figure 2.1(b) plots the log-relative abundances of the five identified species in IBD children and in controls, indicating the identified bacterial species indeed showed differential abundances between IBD cases and controls. Figure 2.1(c) shows the fitted probability curve using the estimated regression coefficients of the identified species, indicating that the model fits the data well.

Our results were confirmed from other studies. Kaakoush et al. (2012) showed healthy people have high level of *Prevotella_copri* within their fecal microbial compared to Crohn's disease patients.

Ruminococcus_bromii and *Clostridium_leptum* (Kabeerdoss et al., 2013; Mondot et al., 2011; Sokol et al., 2009) were also shown to be negatively associated with the risk of IBD. Furthermore, Rhodes (2007) pointed out the association of an increase of *Escherichia_coli* and IBD. Matsuoka and Kanai (2015) also indicated the abundance of *Ruminococcus_gnavus* is higher in IBD patients.

To assess the sensitivity to zero replacement, we also performed the same analysis by replacing the zeros in the relative abundance matrix by 0.1 times the minimum non-zero abundance. The same set of species were identified and their estimated coefficients were almost unchanged (See Table A.1 in Appendix).

Table 2.1: Selected bacterial species and their corresponding phylum, estimated coefficients (standard errors in the parenthesis) and 95% confidence intervals. Model 1: regression analysis with the compositions of 77 bacterial species as covariates. Model 2: regression analysis with the subcompositions of bacterial species that belong to different genera as covariates.

Bacteria name	Phylum	β (se)	CI
Model 1: one constraint on regression coefficients			
Prevotella_copri	Bacteroidetes	$-0.15(0.042)$	$(-0.23, -0.064)$
Ruminococcus_bromii	Firmicutes	$-0.22(0.043)$	$(-0.31, -0.18)$
Clostridium_leptum	Firmicutes	$-0.15(0.052)$	$(-0.25, -0.048)$
Escherichia_coli	Proteobacteria	$0.14(0.035)$	$(0.074, 0.21)$
Ruminococcus_gnavus	Firmicutes	$0.13(0.045)$	$(0.043, 0.22)$
Model 2: multiple constraints on regression coefficients			
Prevotella_copri	Bacteroidetes	$-0.12(0.040)$	$(-0.20, -0.047)$
Ruminococcus_bromii	Firmicutes	$-0.20(0.038)$	$(-0.27, -0.12)$
Bacteroides_cellulosilyticus	Bacteroidetes	$0.087(0.044)$	$(0.0011, 0.17)$
Clostridium_leptum	Firmicutes	$-0.14(0.051)$	$(-0.24, -0.043)$
Clostridium_symbiosum	Firmicutes	$0.12(0.056)$	$(0.012, 0.23)$
Ruminococcus_gnavus	Firmicutes	$0.17(0.042)$	$(0.091, 0.26)$

2.4.2. Identifying bacterial species using subcompositions and multiple constraints

We also performed an analyses by considering multiple constraints. Particularly, we considered the subcomposition of bacterial species that belong to the same genus, for a total of 13 genera with multiple species. This led to fitting a logistic regression model with 13 constraints, where for each genus, the sum of the coefficients corresponding to the subcompositions of the bacteria classified under this genus is constrained to be zero. Our goal is to identify the species whose subcompositions are associated with IBD.

Figure 2.2(a) shows the Lasso estimates, de-biased estimates and the 95$\%$ confidence intervals of the regression coefficients using multiple constraints. The model identified six species

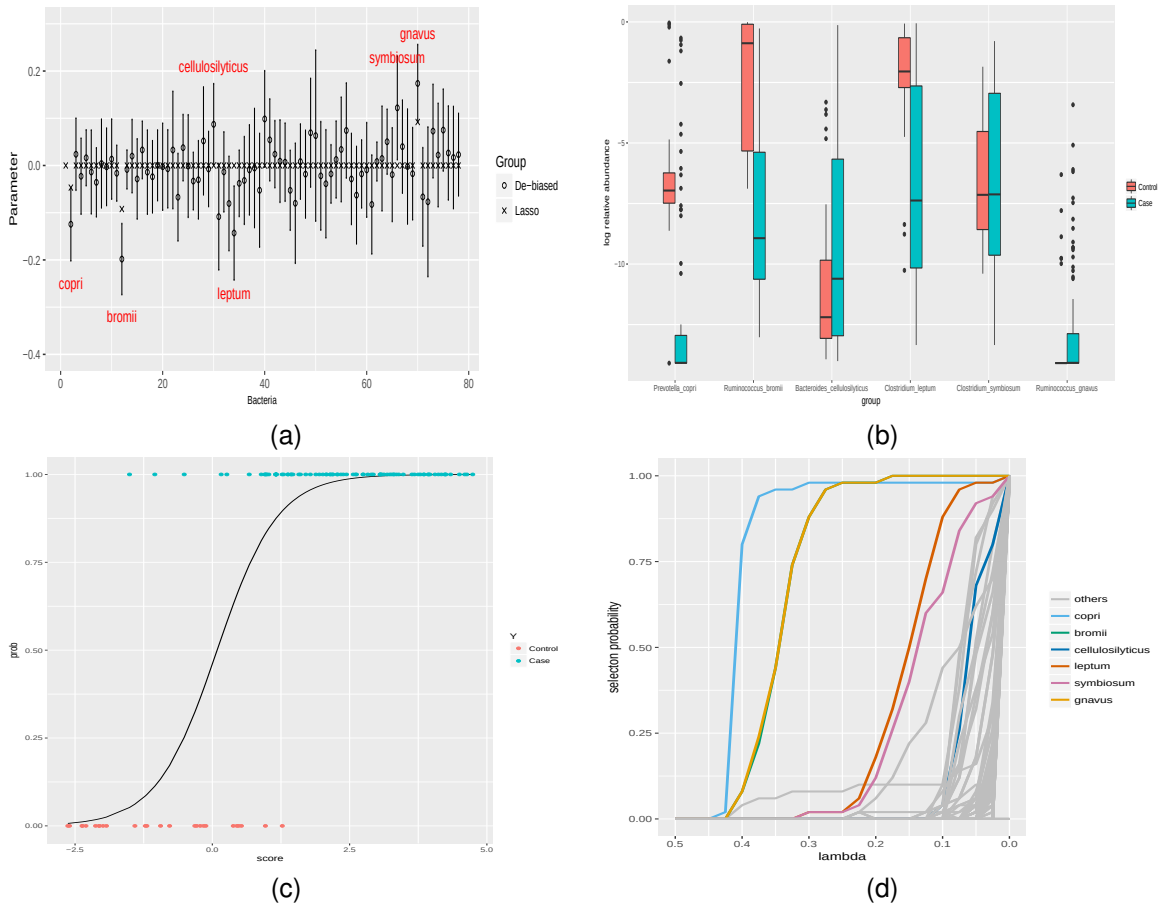


Figure 2.2: Analysis of the IBD microbiome data using multiple constraints. (a) Lasso estimates, de-biased estimates and 95% confidence intervals of the regression coefficients. Species selected based on the CIs are annotated. (b) Boxplots of log-relative abundances of the six identified species. The red and blue boxplots correspond to controls and cases samples respectively. (c) Fitted probability plot. (d) Selection stability plot.

with confidence intervals not covering zero, including *Prevotella_copri*, *Ruminococcus_bromii*, *Bacteroides_cellulosilyticus*, *Clostridium_leptum*, *Clostridium_symbiosum* and *Ruminococcus_gnavus*. Compared to the results with using a single constraint, *Bacteroides_cellulosilyticus* and *Clostridium_symbiosum* were identified to be positively associated with IBD while *Escherichia_coli* became less significant. The estimated regression coefficients and confidence intervals for the species identified by both models were only slightly different.

Figure 2.2(b) plots the log-relative abundances of the six selected species, showing differential abundance between IBD cases and normal controls. The positive association between *Clostrid-*

ium_symbiosum and IBD was also reported in Lozupone et al. (2012). Finally, Figure 2.2(c) shows the fitted probability curve using the estimated regression coefficients of the identified species, indicating that the model fits the data well.

2.4.3. Stability and prediction evaluation

To assess how stable the results are, we performed stability selection analysis (Meinshausen and Bühlmann, 2010) by sample splitting. Among the 50 replications, each time we randomly sampled two third of the data including 56 cases and 16 controls and fitted the model using different tuning parameters. Figure 2.1(d) and Figure 2.2(d) show the selection probability for each of the bacteria versus values of the tuning parameter for models with a single constraint and multiple constraints, respectively. The selected species from both models had the highest stability selection probabilities, indicating that the species selected were very stable.

To evaluate the performance of prediction of IBD based on bacterial composition, we randomly split the data into a training set of 56 cases and 16 controls to estimate the parameters and a testing set of 28 cases and 8 controls to evaluate the prediction performance. Models with a single constraint or multiple constraints were fitted on the training data sets and were used to predict the IBD status in the testing set. The prediction was evaluated using area under the ROC curve (AUCs and was repeated 50 times. The average AUC (se) for model with a single constraints were 0.92(0.049) , 0.93(0.043) and 0.93 (0.051) based on Lasso, debiased Lasso and de-biased Lasso using only the selected bacterial species. The corresponding average AUC (se) for model with multiple constraints were 0.94(0.036) , 0.94(0.038) and 0.94 (0.038). The result indicates that the model can predict IBD very well. Finally, as a comparison, the Random Forests using the same training/testing samples gave an AUC (se) of 0.97 (0.026), slightly better than those from the linear logistic regression models. This is not surprising given the non-linear nature of Random Forests.

2.5. Simulation Studies

We evaluate the performance of the proposed methods through a set of simulation studies. In order to simulate covariate Z and outcome Y, we simulate the true bacterial abundances W, where each row of W is generated from a log-normal distribution $\ln N(\mu, \mathbf{\Sigma})$, where $\mathbf{\Sigma}_{ij} = \zeta^{|i-j|}$ with $\zeta = 0.2$ is the covariance matrix to reflect the correlation between different taxa. Mean parameters are set

as $\mu_j = \frac{p}{2}$ for $j = 1, \ldots, 5$ and $\mu_j = 1$ for $j = 6, \ldots p$. The log-compositional covariate matrix $\mathbf{Z}$ is obtained by normalizing the true abundances

$$\mathbf{Z}_{ij} = \log\left(\frac{W_{ij}}{\sum_{k=1}^{p} W_{ik}}\right),$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$. The true parameter $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = (0.45, -0.4, 0.45, 0, -0.5, 0, 0, 0, 0, 0, -0.6, 0, 0.3, 0, 0, 0.3, 0, \ldots 0)$$

and $\boldsymbol{\beta}_0 = -1$. Based on these covariates, we simulate the binary outcome Y based on the logistic probability $p_i = \text{expit}(\mathbf{Z}_i^\top \boldsymbol{\beta} + \boldsymbol{\beta}_0)$ and obtained the number of cases and controls at a 2:3 ratio. Different dimensions and sample sizes are considered and simulations are repeated 100 times for each setting. The true regression coefficients $\boldsymbol{\beta}$ are assumed to satisfy the following linear constraints:

$$\sum_{i=1}^{10} \boldsymbol{\beta}_i = 0, \sum_{i=11}^{16} \boldsymbol{\beta}_i = 0, \sum_{i=17}^{20} \boldsymbol{\beta}_i = 0, \sum_{i=21}^{23} \boldsymbol{\beta}_i = 0,$$
$$\sum_{i=24}^{30} \boldsymbol{\beta}_i = 0, \sum_{i=31}^{32} \boldsymbol{\beta}_i = 0, \sum_{i=33}^{40} \boldsymbol{\beta}_i = 0, \sum_{i=41}^{p} \boldsymbol{\beta}_i = 0.$$

2.5.1. Simulation results

We evaluate the performance of the simulation by comparing the coverage probability, length of the confidence interval and the true positive and false positive of selecting variables based on the confidence interval. We compare the results of fitting the models with no constraint, one constraint, true constraint and misspecified constraints specified below,

$$\sum_{i=1}^{4} \boldsymbol{\beta}_i = 0, \sum_{i=5}^{12} \boldsymbol{\beta}_i = 0, \sum_{i=13}^{23} \boldsymbol{\beta}_i = 0, \sum_{i=24}^{30} \boldsymbol{\beta}_i = 0, \sum_{i=31}^{p} \boldsymbol{\beta}_i = 0.$$

Figure 2.3 shows that the coverage probabilities are closer to 95% and the length of CIs decrease as sample size becomes larger. In addition, the coverage probabilities under true constraints are closer to the correct coverage probability (95%) especially when n is relatively larger($n = 200, 500$). As for length of CIs, the CIs using the true constraints have the shortest CIs while the length of the CIs for single constraint and no constraints are relatively wider. We did not compare the length of

CI for using misspecified constraints because the coverage probability in this case is really poor. The figure also shows that the coverage probabilities are sensitive to the constraints when sample size becomes larger and the length is sensitive to the constraints for small sample size. This is expected as when the sample size is small, we are more likely to obtain wider CI, and using the correct constraints, which provide more information, would provide shorter CI. While for the coverage probability, since our algorithm provides an asymptotic CI, the sample size has bigger effects than the constraints. The coverage probability becomes really poor when the constraints are misspecified when $n = 500$.

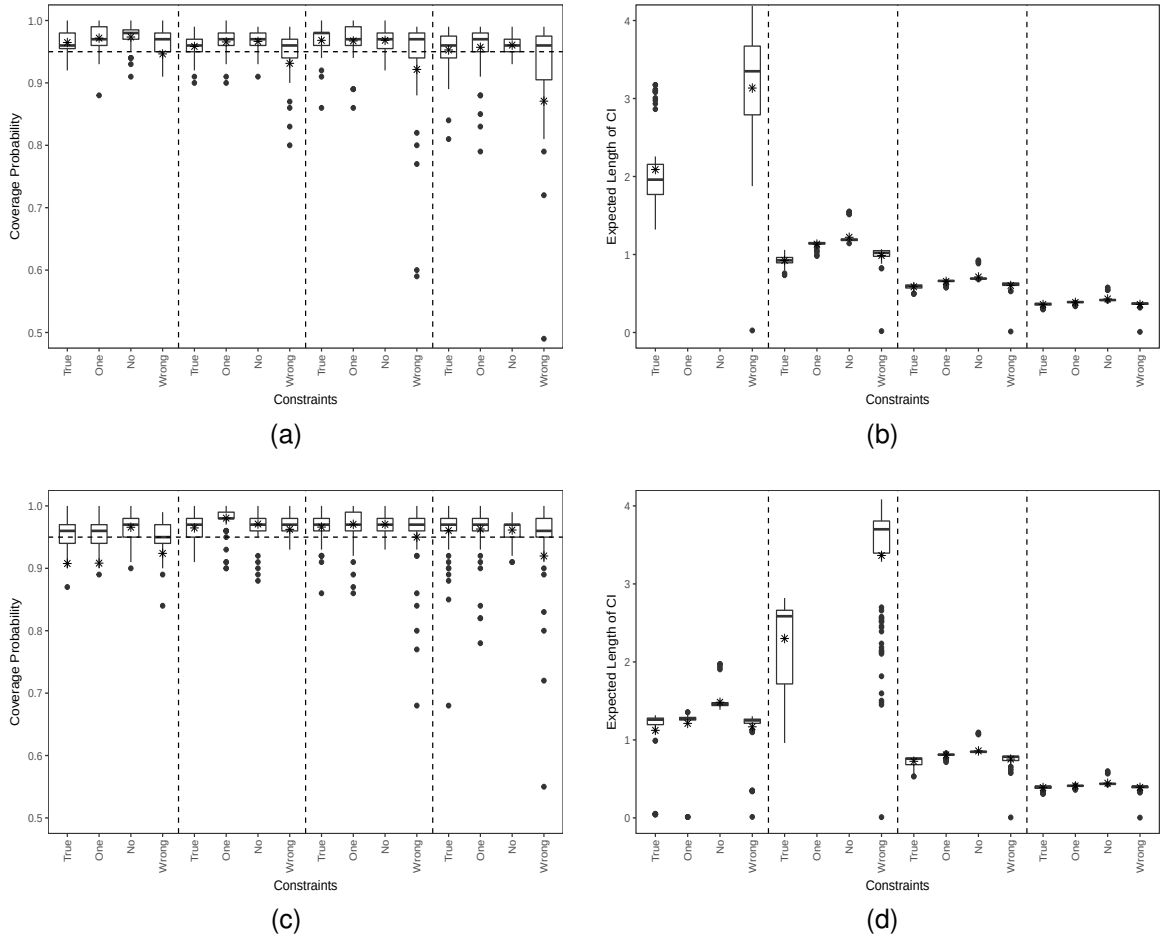


Figure 2.3: Coverage probabilities and length of confidence intervals based on 100 simulations for $p = 50$ ((a) and (b)) and $p = 100$ ((c) and (d)) and $n = 50, 100, 200, 500$ (separated by vertical dashed lines).

Table 2.2 shows the true positive and false positive rates of selecting the significant variables us-

ing the 95% confidence interval under multiple, one, no and misspecified constraints for various dimensions p and sample sizes n. The false positive rates are correctly controlled under 5% for all models, even when the constraints are misspecified. However, models with correctly specified linear constraints have higher true positive rates. When the sample size is 500, true positive rate is greater than 90%, which is the highest among all models considered.

Table 2.2: True /False positive rates of the significant variables selected by the 95% confidence interval using multiple, one, no and misspecified constraints. $p = 50, 100$ and $n = 50, 100, 200, 500$ are considered.

n	TP	FP	TP	FP	TP	FP	TP	FP
	Multi		One		No		Wrong	
				$p = 50$				
50	0.069	0.034	0.026	0.025	0.029	0.026	0.054	0.036
100	0.260	0.038	0.206	0.031	0.141	0.034	0.299	0.038
200	0.569	0.026	0.549	0.025	0.411	0.030	0.546	0.037
500	0.914	0.038	0.897	0.030	0.840	0.038	0.814	0.058
				$p = 100$				
50	0.220	0.045	0.071	0.044	0.109	0.034	0.134	0.046
100	0.103	0.035	0.023	0.016	0.107	0.026	0.154	0.027
200	0.431	0.030	0.389	0.025	0.283	0.029	0.481	0.032
500	0.907	0.032	0.873	0.029	0.801	0.037	0.804	0.042

2.6. Discussion

In this chapter we considered estimation and inference for the generalized linear models with high dimensional compositional covariates. In order to accounting for the nature of compositional data, a group of linear constraints were imposed on the regression coefficients to ensure subcompositional coherence. With these constraints, the standard GLM Lasso algorithm based on Taylor expansion and coordinate descent algorithm did not work due to the non-separable nature of the penalty function. Instead, a generalized accelerated proximal gradient algorithm was developed to estimate the regression coefficients. To make statistical inference, a de-biased procedure was proposed to construct valid confidence intervals of the regression coefficients. Application of the method to an analysis of IBD microbiome data identified five bacterial species that were associated with pediatric IBD with a high stability using a single constraint and six species when imposing multiple constraints. The identified model had also shown a great prediction performance based on cross-validation.

The proposed method could be extended to incorporate the phylogenetic tree information in order

to identify the taxa at different taxonomic levels that are associated with the outcome. At each of the internal node of the phylogenetic tree, we could create a subcomposition of all the taxa under this node. We can apply the proposed regression methods that include all these subcompositions as covariates with sum-zero constraint for the coefficients that correspond to each of the subcompositions.

CHAPTER 3

POST-SELECTION INFERENCE FOR REGRESSION MODELS WITH LINEAR CONSTRAINTS, WITH AN APPLICATION TO MICROBIOME DATA

3.1. Introduction

In many cases, certain constraints are imposed on the regression coefficients in order to enhance the interpretability and to reveal the true data generating processes. As we introduced in Chapter 1, Lin et al. (2014) considered linear regression model with microbiome compositional data as covariates, where a set of linear equality constraints are imposed on the regression coefficients. The necessity and importance of these constraints have been emphasized in many literatures. Estimation of the regression coefficients for linear models with constraints can be obtained in a straightforward way by constrained optimization algorithms, but the inference problem is not trivial in the presence of linear constraints. In classical settings, the equality constrained least-squares (ECLS) estimator is known to be an unbiased and normally distributed under certain assumptions.

With the emergence of high-dimensional data, using Lasso-type regularized estimators has become an effective method for estimating the regression coefficients under the sparsity assumption. In the setting where the number of covariates p, is potentially larger than n, Lasso (Tibshirani, 1996) was applied instead of ordinary least square (OLS) for estimation and variable selection. In many applications, a standard procedure of analyzing the data is to fit a Lasso-type estimator, then to refit the linear model using the variables selected by Lasso. The inference after this refitting procedure for the Lasso has been studied in Lee et al. (2016). The confidence intervals proposed in Lee et al., 2016 are shown to have the desired coverage probability conditioned on the model selected by Lasso. This is the major difference between this procedure and the refitted confidence interval based on OLS or the confidence intervals obtained via debiasing (Javanmard and Montanari, 2014; Zhang and Zhang, 2014).

For models with linear constraints in high-dimensional settings, the inference problem has not been fully addressed. The presence of the constraints complicates the statistical analysis and ignoring such constraints causes problem in variable selection and leads to inefficient estimators. In the

framework of regression models for microbiome data, Shi, Zhang, and Li (2016) and Lu, Shi, and Li (2019) provided inference for linear model and generalized linear model with a set of linear constraints. The interpretation of the results is not conditioned on the selected model, hence is different from that approach that we take. To the best of our knowledge, there has been no published work on the post-selection inference for models with linear constraints in high-dimensional setting.

In this chapter, we studied the post-selection inference problem for linear models with linear equality constraints. We established a method to obtain the confidence intervals for the target parameters conditioned on the selected model using a Lasso-type estimator. By exploring the Karush-Kuhn-Tucker (KKT) conditions, we obtained an equivalent form for the event of selecting a submodel, in terms of a group of linear inequalities of the response vector $\mathbf{y}$. Based on this fact, we were able to obtain the distribution of any linear functional of $\mathbf{y}$ conditioned on the selected model and hence to use it as a pivot for inference of the target parameters. By inverting the pivot we obtained the confidence intervals with desired coverage probabilities. We would like to emphasis that conditioned on the selected model, our method requires fewer assumptions compared to those debiased inference procedures.

3.2. Post-selection inference for high dimensional linear models with linear equality constraints

3.2.1. Linear model with constraints

In this section, we presented a procedure for constructing the post-selection confidence intervals and their theoretical properties for the linear model with linear equality constraints. One of the motivating examples includes the regression model for microbiome compositional data, in which the regression coefficients sum up to zero. As presented in Lin et al. (2014) and Shi, Zhang, and Li (2016), the zero-sum constraints on the regression coefficients ensure the so-called subcompositional coherence (Aitchison, 1982; Aitchison and Bacon-shone, 1984) of the model and lead to more interpretable results. A general version of the model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \text{ subject to } \mathbf{C}^\top \boldsymbol{\beta}_0 = \mathbf{0}, \tag{3.1}$$

where $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$, $\mathbf{y} \in \mathbb{R}^n$, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and

$$\mathbf{C}^\top = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 \\ & \vdots & & & \vdots & & \vdots & & \vdots & & & \vdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 1 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{r \times p}. \tag{3.2}$$

From mathematical and statistical points of view, the specific structure of $\mathbf{C}$ may not be essential. A general form is more desirable but we consider this particular type of $\mathbf{C}$ for two reasons: first, this type of $\mathbf{C}$ has a clear biological interpretation; secondly, using a general form requires imposing many conditions on $\mathbf{C}$ that naturally hold in this special case. This form of constraints has two desired properties:

(A1) $\mathbf{C}$ is a full rank matrix with $r < p$.

(A2) For any set $M \subseteq \{1, 2, \ldots, p\}$, sub-matrix $\mathbf{C}_M$ is a full rank matrix such that $\mathrm{rank}(\mathbf{C}_M) < |M|$.

For the second property, without abusive of notation, we used $\mathbf{C}_M$ to represent the active constraints, instead of the sub-matrix $\mathbf{C}_M$. For example, if the constraints are:

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \\ \boldsymbol{\beta}_4 \end{pmatrix} = 0$$

and $M = \{1, 2\}$, then the form $\mathbf{C}_M^\top a = \mathbf{0}$ is

$$\begin{pmatrix} 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = 0$$

rather than:

$$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = 0.$$

This definition, together with the structure of $\mathbf{C}$, guarantees the condition (A2) always holds. The choice of $\mathbf{C}$ does not limit the application of our method as for other types of $\mathbf{C}$, one only needs to verify if they satisfy the listed assumptions.

Under the classical setting where $p < n$ is fixed and $\mathbf{X}$ is full rank, it is natural to consider the ordinary least square method with the linear equality constraints. The resulting estimator is unbiased and normally distributed. The result is summarized in the following proposition.

Proposition 1. *Assuming $\mathbf{X}$ is full rank and $r < p$, then the linear equality constrained ordinary least square estimator of model* (3.1) *is given by:*

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}[\mathbf{C}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}]^{-1}[\mathbf{C}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}], \\ &= \left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}\left(\mathbf{C}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\mathbf{y}. \qquad (3.3)\end{aligned}$$

The distribution of $\widehat{\boldsymbol{\beta}}$ is $N(\boldsymbol{\beta}_0, \mathrm{Var}(\widehat{\boldsymbol{\beta}}))$, where $\mathrm{Var}(\widehat{\boldsymbol{\beta}})$ is given by:

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\left[\mathbf{I}_p - \mathbf{C}\left(\mathbf{C}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\right].$$

Furthermore, $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \succ \mathrm{Var}(\widehat{\boldsymbol{\beta}})$ where $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ corresponds to the variance of the ordinary least square estimator without using any constraints.

Based on Proposition 1, it is easy to obtain a confidence interval C_j for each $j = 1, 2, \ldots, p$ such that $P(\boldsymbol{\beta}_{0j} \in C_j) = 1 - \alpha$ for some pre-specified α. It also indicates that such confidence intervals would have shorter lengths compared to those based on the OLS estimator. This confirms that ignoring the constraints leads in inefficient estimators.

The idea of post-selection inference is different from this classical method such that the post-selection confidence intervals have the desired coverage probabilities conditioned on a model-selection procedure. That is, $P(\boldsymbol{\beta}_{0j} \in C_j \mid \widehat{M} = M) = 1 - \alpha$. For the models with constraints, we considered a similar problem. This type of inference that conditions on the selected model emphasizes the interpretation of the regression coefficient, which is the effect of a variable on the outcome, adjusting for all other selected variables.

3.2.2. Target parameter

One key property of the post-selection inference that distinguishes it from classical inference is that all the procedures are conditioned on the model selected. Hence, the target parameter depends on the model selection procedure. For any set $M \subseteq \{1, 2, \ldots, p\}$, the target parameter corresponding to the sub-model M is defined as:

$$\boldsymbol{\beta}_{oracle}^{M} = \underset{\boldsymbol{\beta} \in H(M,C)}{\operatorname{argmin}} \mathbb{E}\|\mathbf{y} - \mathbf{X}_M \boldsymbol{\beta}\|_2^2, \tag{3.4}$$

here $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$ with $\Sigma = \sigma^2 \mathbf{I}_p$ and $H(M, \mathbf{C}) = \{a \in \mathbb{R}^{|M|} : \mathbf{C}_M^\top a = \mathbf{0}\}$. When $H(M, \mathbf{C}) = \emptyset$, the target parameter does not exist, we therefore need to impose conditions so that $\boldsymbol{\beta}_{oracle}^{M}$ always exists. Without lose of generality, the columns of $\mathbf{C}$ are assumed to be orthonormal and we impose the following two conditions:

(B1) $\mathbf{X}_M$ is full rank with $\operatorname{rank}(\mathbf{X}_M) = |M| < \min(n, p)$.

(B2) The diagonal elements of $\mathbf{I}_p - \mathbf{C}\mathbf{C}^\top$ are greater than zero.

For assumption (B1), it not only relates to the existence of the target parameter, but also guarantees the uniqueness of the Lasso estimator. In practice, the dimension of the selected sub-model $|M|$ is smaller than $\min(n, p)$, the first part of the assumption is not hard to satisfy. Assumption (B2) is also used in related literature such as Shi, Zhang, and Li (2016) and Lu, Shi, and Li (2019). This assumption eliminates some trivial constraints such as $c_j \boldsymbol{\beta}_j = 0$ for some j.

Under the assumptions (A2), (B1) and (B2), the solution to (3.4) is given by:

$$\begin{aligned}\boldsymbol{\beta}_{oracle}^{M} = \Big[&(\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top - (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M \left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M\right)^{-1} \mathbf{C}_M^\top \\ &(\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top \Big] \boldsymbol{\mu}.\end{aligned} \tag{3.5}$$

Further if $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_0$ and $M = \operatorname{supp}(\boldsymbol{\beta}_0)$, then $\boldsymbol{\beta}_{oracle}^{M}$ could exactly recovers the non-zero elements of $\boldsymbol{\beta}_0$. This indicates that under a true linear model and if we could correctly select the subset (recovers the support of $\boldsymbol{\beta}_0$), then the target parameter could capture the information of the true parameter $\boldsymbol{\beta}_0$. However, for general choices of M, the target parameter of interests $\boldsymbol{\beta}_{oracle}^{M}$ has no direct relation with $\boldsymbol{\beta}_0$. So when studying the inference problem conditioned on the selected

model, one should notice that the target we are making inference on, which is defined by (3.4), is the best-linear estimator under the constraints. Under a true linear data generating process (the case when $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}_0$), it is still possible that the target has no relationship with the true $\boldsymbol{\beta}_0$.

3.2.3. Confidence intervals for target parameter

In the high dimensional setting, when p is potentially larger than n, an estimator of model (3.1) is given by:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in H(\mathbf{C})}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{3.6}$$

where $H(\mathbf{C}) = \{a \in \mathbb{R}^p : \mathbf{C}^\top a = \mathbf{0}\}$. To guarantee the uniqueness of $\widehat{\boldsymbol{\beta}}$, we also need an assumption (B3) that the columns of $\mathbf{X}$ are in general position (Tibshirani, 2013). We considered the post-selection inference for the target parameter $\boldsymbol{\beta}^M_{oracle}$, conditioned on $\{\widehat{M} = \operatorname{supp}(\widehat{\boldsymbol{\beta}}) = M\}$. That is, we would like to find a confidence interval C_j such that for each $j \in M$,

$$P(\boldsymbol{\beta}^M_{oracle,j} \in C_j \mid \widehat{M} = M) = 1 - \alpha.$$

The key part of post-selection inference is to study the event of selecting certain sub-model M. For technical reasons, we focus on that event $\{\widehat{M} = M, \widehat{s} = s\}$ instead of $\{\widehat{M} = M\}$, where $\widehat{s} = \operatorname{sign}(\widehat{\boldsymbol{\beta}})$. The following lemma indicates that the event $\{\widehat{M} = M, \widehat{s} = s\}$ can be quantified by a set of inequalities of $\mathbf{y}$.

Lemma 2. *Suppose $\widehat{\beta}$ is defined in* (3.6)*, assumptions (A1)-(A2), (B1)-(B3) hold, then the event $\{\widehat{M} = M, \widehat{s} = s\}$ is equivalent to $\{A\mathbf{y} < b\}$, where A and b is defined by the following:*

$$A = \begin{pmatrix} A_0 \\ A_1 \end{pmatrix}, b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix},$$

with:

$$
\begin{aligned}
A_1 = &- \textit{diag}(s)(\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \times \\
&\left(\mathbf{X}_M^\top - \mathbf{C}_M \left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M \right)^{-1} \mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top \right), \\
b_1 = &- \textit{diag}(s)(\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \times \\
&\left(\lambda s - \mathbf{C}_M \left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M \right)^{-1} \left[\lambda \mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} s \right] \right),
\end{aligned}
$$

and

$$
A_0 = \begin{pmatrix} A_{01} \\ A_{02} \end{pmatrix}, b_0 = \begin{pmatrix} b_{01} \\ b_{02} \end{pmatrix},
$$

which A_{01}, A_{02}, b_{01} *and* b_{02} *are given as following:*

$$
\begin{aligned}
A_{01} = &-\frac{1}{\lambda} \mathbf{X}_{-M}^\top \left(I - \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top \right) - \left[\mathbf{X}_{-M}^\top \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \cdot \frac{1}{\lambda} \mathbf{C}_M - \right. \\
&\left. \frac{1}{\lambda} \mathbf{C}_{-M} \right] \cdot \left[\left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M \right)^{-1} \mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top \right], \\
A_{02} = &\frac{1}{\lambda} \mathbf{X}_{-M}^\top \left(I - \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top \right) + \left[\mathbf{X}_{-M}^\top \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \cdot \frac{1}{\lambda} \mathbf{C}_M - \right. \\
&\left. \frac{1}{\lambda} \mathbf{C}_{-M} \right] \cdot \left[\left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M \right)^{-1} \mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top \right],
\end{aligned}
$$

and

$$
\begin{aligned}
b_{01} = &\mathbf{1} + \mathbf{X}_{-M}^\top \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} s - \left[\mathbf{X}_{-M}^\top \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \cdot \frac{1}{\lambda} \mathbf{C}_M - \frac{1}{\lambda} \mathbf{C}_{-M} \right] \cdot \\
&\left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M \right)^{-1} \left[\lambda \mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} s \right], \\
b_{02} = &\mathbf{1} - \mathbf{X}_{-M}^\top \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} s + \left[\mathbf{X}_{-M}^\top \mathbf{X}_M (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \cdot \frac{1}{\lambda} \mathbf{C}_M - \frac{1}{\lambda} \mathbf{C}_{-M} \right] \cdot \\
&\left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{C}_M \right)^{-1} \left[\lambda \mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} s \right].
\end{aligned}
$$

Lemma 2 indicates that the event we conditioned on $\{\widehat{M} = M, \widehat{s} = s\}$, is actually a system of linear inequalities on $\mathbf{y}$. To utilize this fact, we first provided an equivalent form of the event $\{A\mathbf{y} < b\}$. Notice that $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$, then for any $\boldsymbol{\xi} \in \mathbb{R}^n$, define $\mathbf{z} = (I_n - c\boldsymbol{\xi}^\top)\mathbf{y}$, with $c = \Sigma\boldsymbol{\xi}(\boldsymbol{\xi}^\top \Sigma \boldsymbol{\xi})^{-1}$, it is easy to verify that $\mathbf{z}$ is independent of $\xi^\top \mathbf{y}$. Hence, based on Lemma 5.1 in Lee et al. (2016), we

know that:

$$\{A\mathbf{y} < b\} = \{\nu^-(\mathbf{z}) \leq \boldsymbol{\xi}^\top \mathbf{y} \leq \nu^+(\mathbf{z}), \nu^0(\mathbf{z}) \geq 0\},$$

where

$$\nu^-(\mathbf{z}) = \max_{j:(Ac)_j<0} \frac{b_j - (A\mathbf{z})_j}{(Ac_j)} \tag{3.7}$$

$$\nu^+(\mathbf{z}) = \min_{j:(Ac)_j>0} \frac{b_j - (A\mathbf{z})_j}{(Ac_j)} \tag{3.8}$$

$$\nu^0(\mathbf{z}) = \min_{j:(Ac)_j=0} b_j - (A\mathbf{z})_j$$

Based on this key fact, the following theorem provides the post-selection confidence intervals of $\boldsymbol{\beta}^M_{oracle}$.

Theorem 3. *Suppose $\boldsymbol{\beta}^M_{oracle}$ is defined in* (3.5) *and for each $j \in M$, let:*

$$\begin{aligned}\boldsymbol{\xi} =& e_j^\top \big[(\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{X}_M^\top - (\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{C}_M \left(\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{C}_M\right)^{-1} \\ & \mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{X}_M^\top\big],\end{aligned}$$

and let U and L are the unique values defined by:

$$F^{\nu^-,\nu^+}_{L,\sigma^2\|\boldsymbol{\xi}\|^2}(\boldsymbol{\xi}^\top \mathbf{y}) = 1 - \frac{\alpha}{2},\ F^{\nu^-,\nu^+}_{U,\sigma^2\|\boldsymbol{\xi}\|^2}(\boldsymbol{\xi}^\top \mathbf{y}) = \frac{\alpha}{2}, \tag{3.9}$$

where ν^- and ν^+ are defined in (3.7) *and* (3.8)*, and $F^{a,b}_{\mu,\sigma^2}$ is the CDF of a normal distribution $N(\mu, \sigma^2)$ truncated to the interval $[a, b]$, then*

$$P(\boldsymbol{\beta}^M_{oracle,j} \in [L, U] \mid \widehat{M} = M, \widehat{s} = s) = 1 - \alpha$$

That is, $[L, U]$ is a $(1 - \alpha) \times 100\%$ confidence interval for $\boldsymbol{\beta}^M_{oracle,j}$ conditional on the event $\{\widehat{M} = M, \widehat{s} = s\}$. Furthermore,

$$P(\boldsymbol{\beta}^M_{oracle,j} \in [L, U] \mid \widehat{M} = M) \geq 1 - \alpha.$$

This indicates that the resulting confidence interval has the coverage probability above $1 - \alpha$.

This theorem provides a way of constructing the post-selection confidence intervals conditioned on the model and sign. The following corollary provides the confidence interval that only conditioned on the selected model.

Corollary 1. *Suppose $\boldsymbol{\beta}^M_{oracle}$ is defined in* (3.5) *and for each $j \in M$, $\boldsymbol{\xi}$ is defined as in Theorem 3,and let $\widetilde{L}$ and $\widetilde{U}$ are the unique values defined by:*

$$F^{\bigcup_s(\nu_s^-,\nu_s^+)}_{\widetilde{L},\sigma^2\|\boldsymbol{\xi}\|^2}(\boldsymbol{\xi}^\top \mathbf{y}) = 1-\frac{\alpha}{2},\ F^{\bigcup_s(\nu_s^-,\nu_s^+)}_{\widetilde{U},\sigma^2\|\boldsymbol{\xi}\|^2}(\boldsymbol{\xi}^\top \mathbf{y}) = \frac{\alpha}{2},$$

where ν_s^- and ν_s^+ are defined in (3.7) *and* (3.8) *with given sign s, and F^S_{μ,σ^2} is the CDF of a normal distribution $N(\mu,\sigma^2)$ truncated to a set S, then*

$$P(\boldsymbol{\beta}^M_{oracle,j} \in [\widetilde{L},\widetilde{U}] \mid \widehat{M} = M) = 1-\alpha$$

That is, $[\widetilde{L},\widetilde{U}]$ is a $(1-\alpha)\times 100\%$ confidence interval for $\boldsymbol{\beta}^M_{oracle,j}$ conditioned on the selected model $\{\widehat{M} = M\}$.

Comparing the results from Theorem 3 and Corollary 1, there is a trade-off between accuracy and efficiency. Conditioned only on the selected model is our desired result, which provides a confidence interval with exact coverage probability, but computing $\bigcup_s(\nu_s^-,\nu_s^+)$ can be time-consuming. With s runs through all possible $2^{|M|}$ sign combinations, this will not be feasible with large $|\widehat{M}|$. In contrast, confidence intervals conditioned on both the model and the sign is computationally efficient, but the confidence intervals do not have exact coverage probability of $1-\alpha$. In Section 3.5, we presented simulations to compare these two types of confidence intervals.

Despite this issue, the post-selection confidence interval for linear model with linear constraints still has many advantages. First, under the assumption of Gaussian error, the confidence interval has an exact coverage probability that requires no further assumption on n and p and $\boldsymbol{\beta}_0$. This is the key difference between our method and the de-biased estimator. In addition, this approach has its own benefits in interpretation, particularly in applications to the microbiome regression analysis. Due to the normalization step, studying the sub-model (refitted model) is important after a variable selection step. When focusing on the sub-models, one should renormalize the data into compositions and refit the model for further analysis. The advantage of using constraints is that the renormalization step is not necessary and the post-selection inference provides a natural interpretation of the

confidence intervals by emphasizing the model selection procedure.

In Theorem 3 and Corollary 1, the obtained confidence intervals involve the potentially unknown parameter σ^2. In theory, any consistent estimator of σ^2 can be used. In Section 3.5, we provided two different methods of estimating σ^2 under different scenarios.

3.3. Optimization algorithm and computational details

3.3.1. Optimization algorithm

The optimization algorithms for (3.6) have been well studied in literature. This type of optimization problems belong to the class of convex optimization problems with constraints. We used a coordinate descent algorithm (Shi, Zhang, and Li, 2016) to estimate the parameters in the model.

In Theorem 3 there is a key step in obtaining the confidence interval that requires to find the unique value satisfying (3.9). Since $F^{a,b}_{\mu,\sigma^2}(x)$ is monotone-decreasing in μ, we used a grid searching method to find the unique value that satisfies the equalities.

3.3.2. Estimation of σ^2 and choice of tuning parameter

As we discussed in Section 3.2, the unknown parameter σ^2 need to be estimated. When n is much larger than p, σ^2 could be well-estimated using the residual sum of square:

$$\widehat{\sigma}^2 = \frac{1}{n-p-1}\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{full}}\|_2^2.$$

Here, $\widehat{\boldsymbol{\beta}}_{\text{full}}$ is the regression coefficient obtained by fitting a model using all candidate covariates. When n is smaller than p, the above estimator is not valid. We suggested using scaled Lasso (Sun and Zhang, 2012) to get an estimate of σ^2.

For the choice of tuning parameter, there are several approaches that are applicable. For data driven method, one can use K-fold cross validation and Bayesian information criterion (BIC) to select the tuning parameter. These are standard procedures for selecting the tuning parameter in the penalized regression literature. This parameter can also be chosen manually through analyzing the piecewise-linear solution path or via the stability selection plot (Meinshausen and Bühlmann, 2010). Since cross-validation or BIC tends to select too many variables, for the purpose of interpretability,

one may select the variables by plotting the solution path or the stability selection plot and then select a model of relatively small size.

3.4. Applications: UK twins data

We applied our methods to a UK twins dataset (Goodrich et al., 2016) to associate gut microbiome with age. The UK twin study includes 13500 twins registered in the database since 1992, of which over 9000 are actively participating. In our application, data on 1110 pairs of twins with gut microbiome information are available. We analyzed this dataset aiming at exploring the association between gut microbiome composition and age. The analysis aims to address the questions of whether microbiome can serve as a biological marker for true age (Woodmansey, 2007).

We randomly chose one individual from each twin pair and obtained the relatively abundance of 55 bacterial genera after removing the bacterial genera that only appeared in a few samples. We renormalized the data at genus level and fitted the model with proper constraints. Specifically, we considered the model

$$\text{age}_i = \beta_0 + \sum_{k=1}^{55} \beta_k \log X_{ik} + \epsilon_i, \text{ subject to } \textstyle\sum_k \beta_k = 0,$$

where X_{ik} is the relative abundance of the kth bacterial in individual i, and β_k is the regression coefficient. Since the scaled Lasso method selects too many variables and hence makes post-selection inference unfeasible, we presented results based on the tuning parameter selected by stability selection (Meinshausen and Bühlmann, 2010), as shown in Figure 3.1. Specifically, we randomly selected 800 subjects and fitted the model with different tuning parameters and recorded the variables that were selected. This procedure was replicated 500 times. In the stability selection plots, we showed the probability of each variable being selected under different tuning parameters and chose the tuning parameter that results in stable variable selection.

Using the stability plot, we chose 6 bacterial genera with the tuning parameter $\lambda = 0.094$. In Figure 3.2 and Table 3.1, we provided the post-selection confidence intervals together with the de-biased Lasso estimates, refitted estimates and their corresponding confidence intervals (the computation for the post-selection confidence interval for *Blautia* fails to converge and hence is not shown). The results show that the relative abundances of *Bifidobacterium* and *Blautia Faecalibacterium*

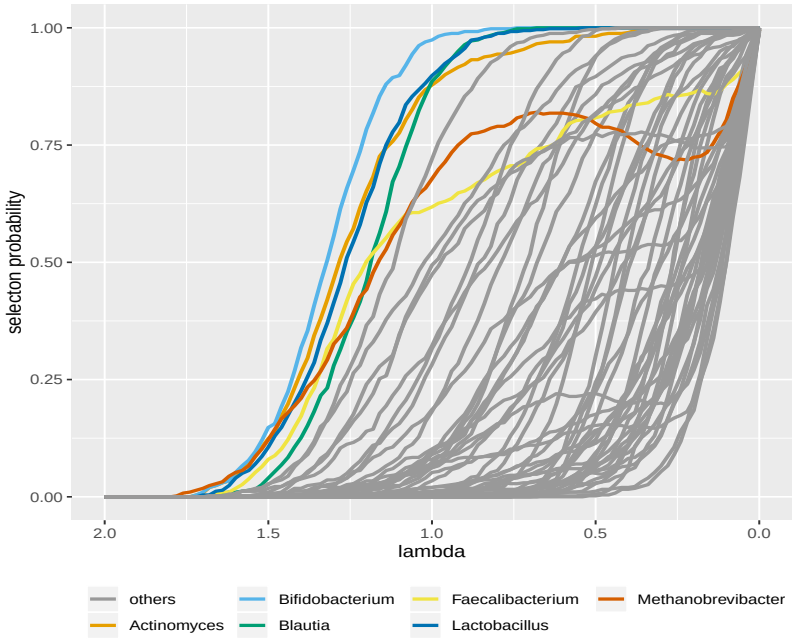


Figure 3.1: Stability selection plot for UK twin data based on Lasso with a zero sum constraint of the regression coefficients.

decrease in the elderly, but *Actinomyces*, *Lactobacillus* and *Methanobrevibacter* increase in the elderly. These results largely agree with the consensus is that the elderly gut has lower counts of short chain fatty acid producers such as *Faecalibacterium* and and an increased number of aero-tolerant and pathogenic bacteria such as *Actinomyces* and *Methanobrevibacter*. As expected, the lengths of the confidence intervals for these 6 regression coefficients are wider than those based on refitted regression using the selected variables. The post-selection inference identified two bacterial genera that are statistically significant based on their 95% post-selection confidence intervals, including *Actinomyces* and *Bifidobacterium*. Both genera appear the top of the stability plot of Figure 3.1. They both belong to phylum *Actinobacteria*. *Bifidobacterium* is the most predominant genus of the breast-fed infant gut microbiota. It has been show that the numbers of this genus substantially decrease after weaning and continue to decrease with age (Kato et al., 2017; Woodmansey, 2007).

We compared the results with the model without imposing the linear constraint on coefficients. For a direct comparison, we manually selected the tuning parameter based on the solution path for Lasso so that it also selects 6 genera. The 6 selected genera are listed in Table 3.1 and presented in Figure 3.2, only *Actinomyces* and *Lactobacillus* are identified by both methods. The

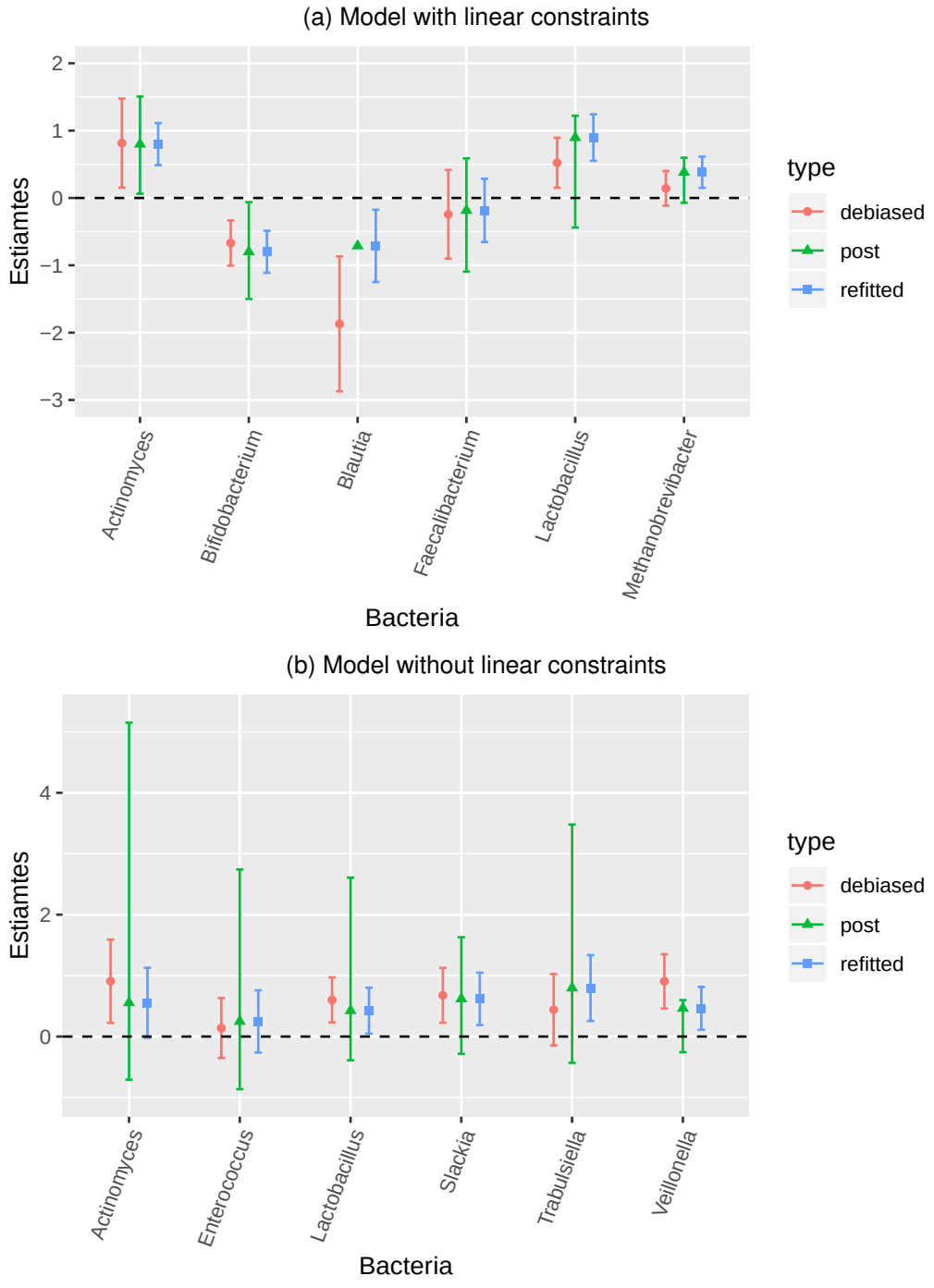


Figure 3.2: Estimates and confidence intervals of the regression coefficients for UK twins dataset. (a): Model with linear constraints; (b): Model without linear constraints

resulting point estimates are also different such that the all the regression coefficients are positive for the model without constraints. In addition, the post-selection confidence intervals from the model

Table 3.1: Estimates and confidence intervals of the regression coefficients using different methods applying to the UK twins dataset. The computation for the post-selection confidence interval for *Blautia* fails to converge and hence is not shown.

Genus	Post-selection	Refitted
Model with a constraint on β: $\sum_k \beta_k = 0$		
Actinomyces	$0.80(0.063, 1.51)$	$0.80(0.49, 1.11)$
Bifidobacterium	$-0.80(-1.50, -0.062)$	$-0.80(-1.11, -0.49)$
Blautia	-0.71	$-0.71(-1.25, -0.17)$
Faecalibacterium	$-0.19(-1.09, 0.59)$	$-0.19(-0.66, 0.28)$
Lactobacillus	$0.90(-0.44, 1.22)$	$0.90(0.55, 1.25)$
Methanobrevibacter	$0.38(-0.072, 0.60)$	$0.38(0.15, 0.61)$
Model without a constraint on β		
Actinomyces	$0.55(-0.71, 5.15)$	$0.55(-0.02, 1.13)$
Enterococcus	$0.25(-0.86, 2.74)$	$0.25(-0.26, 0.76)$
Lactobacillus	$0.42(-0.39, 2.61)$	$0.42(0.05, 0.80)$
Slackia	$0.62(-0.28, 1.63)$	$0.62(0.19, 1.05)$
Trabulsiella	$0.80(-0.43, 3.48)$	$0.80(0.26, 1.34)$
Veillonella	$0.46(-0.26, 0.60)$	$0.46(0.11, 0.82)$

with out constraints all include zero. The results indicate the importance of imposing constraint for compositional covariates, both in term of biological interpretability and in term of identifying biologically important bacterial genera.

3.5. Simulation studies

3.5.1. Simulation setup

We performed a set of simulations to examine the validity of our methods under different settings. We consider different sample sizes with $n = 100, 200$ and 500 and $p = 50$, 500 (for moderate and high dimensional settings). For given n and p, the data $\{\mathbf{y}_i, \mathbf{X}_i\}_{i=1}^n$ is generated as following: $\mathbf{X} \sim N(0, \mathbf{I}_p)$. With given $\mathbf{X}$, $\mathbf{y}_i$ is generated by $\mathbf{y}_i = \mathbf{X}^\top \boldsymbol{\beta}_0 + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$. In this setting, the covariate matrix is fixed across different replications. The methodology we provided does not require a random design $\mathbf{X}$, hence we fixed it so that the model selection step would be stable. The true parameter $\boldsymbol{\beta}_0$ is chosen as:

$$\boldsymbol{\beta}_0 = (2, -2, 2, 0, 0, -2, 0, 0, 0, 0, -4, 0, 2, 0, 0, 2, 0, \ldots, 0)\,.$$

This true parameter satisfies the linear equality constraints:

$$\sum_{i=1}^{6} \boldsymbol{\beta}_i = 0, \sum_{i=7}^{p} \boldsymbol{\beta}_i = 0.$$

For any selected model M such that $M \neq \text{supp}(\boldsymbol{\beta}_0)$, the $\boldsymbol{\beta}^M_{oracle}$ is computed based on (3.5). Throughout the simulations, we chose the tuning parameter in Lasso problem (3.6) by setting the tuning parameter to a fix value. The reason for manually choosing the tuning parameter is that we need to evaluate the inference conditioned on a selected model. Among all the simulation runs, the models selected cannot be guaranteed to be the same and we could only use those that are same so that the inference is made conditioned on the same model. With this requirement, setting the tuning parameter to certain fixed value allows us to determine which submodel is selected. It should be emphasized that our method is targeted to the inference after model selection and is applicable to almost any reasonably chosen tuning parameters. The choice of tuning parameter determines which model we conditioned on, but has no impact on the inference procedure.

To evaluate the performances of the post-selection confidence intervals, we measured the empirical coverage probability and average length of the CI for each coefficient. The empirical coverage probability for each coefficient is defined as, for each $j \in \{1, \ldots, M\}$,

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\boldsymbol{\beta}^M_{oracle,j} \in [L_i, U_i] \mid \widehat{M} = M\},$$

where N is the number of replications. The reported value is averaged over the selected variables. Since the number of selected variables is small, we do not report the standard errors. The simulation is replicated until a specific submodel is selected 500 times.

We compared the simulation results for models with different constraints and the performance of the confidence intervals conditioned on the selected model and signs and those conditioned on the selected model only. We also examined the impact of estimating the unknown parameter σ^2. Specifically, , when $p = 50$, we estimated σ^2 using the residual sum of square of fitting a full model, and when $p = 500$ we used scaled Lasso.

3.5.2. Simulation results

We considered models with three different constraints on regression coefficients: no constraint, one single constraint with $\sum_{i=1}^{p} \boldsymbol{\beta}_i = 0$ and true multiple constraints ($\sum_{i=1}^{6} \boldsymbol{\beta}_i = 0, \sum_{i=7}^{p} \boldsymbol{\beta}_i = 0$). For the case of $p = 50$, we chose the tuning parameter as $\lambda = 0.5n$ and $2n$, $\lambda = 0.5n$ and n and $\lambda = n$ and $2n$ for $n = 500$, 200 and 100 respectively. When $p = 500$, we chose different tuning parameters for different n and models. To be specific, when $n = 500$, we considered $\lambda_1 = 0.5n$ and $\lambda_2 = 1.5n$ for all three models. When $n = 200$, we chose λ_1 and λ_2 to be $0.5n$, n for models with no and single constraint, and λ_1 and λ_2 to be n and $1.5n$ for model with multiple constraints. For $n = 100$, we chose $\lambda_1 = n$, $\lambda_2 = 1.5n$ for models with no constraint, $1.5n$, $3n$ for model with a single constraint, and n, $2n$ for model with multiple constraints. For each n and p setting and each model, there is at least one tuning parameter that the corresponding selected variables contain the support of $\boldsymbol{\beta}_0$.

Table 3.2 presents average coverage probabilities of the post-selection confidence intervals under different (n, p) settings with different parameter constraints using the true σ^2 and the estimated σ^2. As we previously suggested, we estimated σ^2 using the residual sum of square obtaining from the full model for $p = 50$ and using scaled Lasso for $p = 500$. Results in Table 3.2 indicate that using the estimated σ^2 does not affect the results too much. Overall, we observed that the coverage probabilities are averaged above the pre-specified 0.95 $(\alpha = 0.05)$ level for all models and methods, showing the validity of our proposed methods. We observed that the proposed methods of estimating the variance σ^2 work very well for $p = 50$. When $p = 500$, there is an inflation in the coverage probabilities due to the finite-sample estimation bias of the variance parameter σ^2. With the increase of the sample sizes, the bias is reduced and the coverage probabilities are closer to 0.95. This demonstrates the impact of estimating the unknown variance parameter in both moderate and high dimensional settings using two different methods. Table 3.2 also shows that the coverage probabilities are close to 0.95 when conditioned on both the selected model and signs of the coefficients, which implies that conditioned on selected model and signs can provide a computationally efficient alternative to the method that only conditions on the selected model.

When comparing the average length of the confidence intervals, we need to consider the target parameter. Since different model selection leads to distinct target parameter, we can only compare across the settings where the selected models and the target parameters are the same. When

$n = 500$, 200 and 100 and $p = 50$ with $\lambda = 0.5n$, $0.5n$ and n respectively, the corresponding estimator recovers the support of β_0 and hence their targets are the same. Table 3.3 shows the lengths of post-selection confidence intervals for the same settings as in Table 3.2. We observed that the lengths of the confidence intervals for these settings decrease when multiple constraints are imposed. This is consistent with Proposition 1. Similar to the coverage probabilities, when $p = 500$, a increased length of the confidence intervals is observed due to finite sample bias of the estimated variance. With the increase of the sample sizes, the bias is reduced and the lengths of the confidence intervals decrease. These results show that using correct constraints leads to reduction of the length of the confidence intervals.

Table 3.2: Average coverage probabilities of the post-selection confidence intervals obtained by conditioned on selected model and signs or conditioned only on the selected model. Three different contraints on coefficients and two different tuning parameters are considered. For each setting, the first row represents the confidence intervals calculated assuming the variance parameter σ^2 is known and the second row represents the confidence intervals calculated when the variance parameter σ^2 is estimated. For $p = 500$, see text for selection of the tuning parameters λ_1 and λ_2.

		Model and sign			Model only		
		No	Single	True	No	Single	True
			$p = 50$				
$n = 500$	$\lambda = 0.5n$	0.95	0.95	0.95	0.95	0.95	0.95
		0.95	0.95	0.95	0.95	0.95	0.95
	$\lambda = 2n$	0.96	0.95	0.96	0.96	0.95	0.96
		0.97	0.95	0.96	0.96	0.95	0.96
$n = 200$	$\lambda = 0.5n$	0.95	0.95	0.95	0.95	0.95	0.95
		0.95	0.95	0.95	0.95	0.95	0.95
	$\lambda = n$	0.95	0.95	0.95	0.95	0.95	0.95
		0.95	0.94	0.95	0.95	0.94	0.95
$n = 100$	$\lambda = n$	0.95	0.95	0.95	0.95	0.96	0.95
		0.95	0.95	0.95	0.95	0.95	0.95
	$\lambda = 2n$	0.96	0.96	0.97	0.97	0.96	0.97
		0.97	0.95	0.96	0.97	0.95	0.96
			$p = 500$				
$n = 500$	$\lambda = \lambda_1$	0.95	0.95	0.95	0.95	0.95	0.95
		0.96	0.96	0.96	0.96	0.96	0.96
	$\lambda = \lambda_2$	0.95	0.95	0.95	0.95	0.95	0.95
		0.96	0.96	0.96	0.96	0.96	0.96
$n = 200$	$\lambda = \lambda_1$	0.95	0.95	0.95	0.95	0.95	0.95
		0.98	0.97	0.97	0.98	0.98	0.97
	$\lambda = \lambda_2$	0.95	0.95	0.95	0.96	0.95	0.95
		0.97	0.98	0.97	0.98	0.98	0.97
$n = 100$	$\lambda = \lambda_1$	0.95	0.96	0.96	0.96	0.96	0.96
		1.00	1.00	1.00	1.00	1.00	1.00
	$\lambda = \lambda_2$	0.95	0.97	0.95	0.96	0.97	0.95
		1.00	1.00	1.00	1.00	1.00	1.00

Table 3.3: Average length of the post-selection confidence intervals obtained by conditioned on the selected model and signs or conditioned only on the selected model. Three different contraints on coefficients and two different tuning parameters are considered. For each setting, the first row represents the confidence intervals calculated assuming the variance parameter σ^2 is known and the second row represents the confidence intervals calculated when the variance parameter σ^2 is estimated. For $p = 500$, see text for selection of the tuning parameters λ_1 and λ_2.

		Model and sign			Model only		
		No	Single	True	No	Single	True
			$p = 50$				
$n = 500$	$\lambda = 0.5n$	0.17	0.16	0.15	0.17	0.16	0.15
		0.17	0.16	0.15	0.17	0.16	0.15
	$\lambda = 2n$	0.17	0.16	0.18	0.17	0.16	0.18
		0.17	0.15	0.18	0.17	0.15	0.18
$n = 200$	$\lambda = 0.5n$	0.27	0.25	0.23	0.28	0.25	0.23
		0.27	0.25	0.23	0.27	0.25	0.23
	$\lambda = 2n$	0.27	0.25	0.23	0.27	0.25	0.23
		0.27	0.25	0.23	0.27	0.25	0.23
$n = 100$	$\lambda = n$	0.38	0.35	0.33	0.38	0.35	0.33
		0.38	0.36	0.33	0.39	0.36	0.33
	$\lambda = 2n$	0.45	0.33	0.30	0.45	0.33	0.30
		0.45	0.33	0.30	0.46	0.33	0.31
			$p = 500$				
$n = 500$	$\lambda = \lambda_1$	0.17	0.16	0.15	0.17	0.16	0.15
		0.18	0.17	0.15	0.18	0.17	0.15
	$\lambda = \lambda_2$	0.17	0.16	0.15	0.17	0.16	0.15
		0.18	0.17	0.15	0.18	0.17	0.15
$n = 200$	$\lambda = \lambda_1$	0.27	0.25	0.23	0.28	0.25	0.23
		0.24	0.22	0.20	0.24	0.22	0.20
	$\lambda = \lambda_2$	0.27	0.25	0.27	0.27	0.25	0.23
		0.24	0.22	0.20	0.24	0.22	0.20
$n = 100$	$\lambda = \lambda_1$	0.49	0.50	0.52	0.49	0.49	0.52
		0.72	0.74	0.75	0.73	0.71	0.73
	$\lambda = \lambda_2$	0.47	0.64	0.33	0.45	0.64	0.34
		0.64	1.03	0.39	0.61	1.02	0.40

3.6. Discussion

Regression models with constraints raise naturally in field of microbiome research. Imposing constraints on the model brings benefits in capturing the true data generating process, which leads to more efficient estimators. When the number of covariates is potentially larger than the sample size, Lasso-type estimators are often used for estimation and model selection. In this chapter we considered the problem of post-selection inference for high-dimensional linear models with linear constraints and developed a method for obtaining the confidence intervals that have desired coverage probability conditioned on the selected models using Lasso. We carefully explored the

statistical properties of the linear-constrained model and used these properties to propose post-selection confidence intervals. Using the KKT conditions of the constrained Lasso, we found the equivalent form of the model selection procedure and utilize this form to construct a pivot. Finally, by inverting the pivotal quantity, we obtained the confidence intervals for the target parameter.

Due to the linear functional form of the target parameter, the confidence intervals we obtained have an exact coverage probability that does not require any asymptotic assumptions on n and p. This method emphasizes the refitting procedure that is widely used in the microbiome applications. Through the simulations we also compared the results of using or not using the constraints, which indicates that using the constraints would provide more efficient confidence intervals. Lastly, as commended in Lee et al. (2016), when the signal is weak, the algorithm may not converge and can fail to provide valid confidence intervals.

A natural extension of our method is to consider the post-selection inference for generalized linear models (GLM) with constraints. Using the results introduced in this chapter and those of Taylor and Tibshirani (2018), it is straight forward to make inference on the target parameter for GLMs with constraints on regression coefficients by applying the iterated weighted least square method.

CHAPTER 4

HYPOTHESIS TESTING IN HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES

4.1. Introduction

Many genomic studies collect both germline genetic variants and tissue-specific gene expression data on the same set of individuals in order to understand how genetic variants perturb gene expressions that lead to clinical phenotypes. Among various methods, association analysis between gene expression and phenotype such as differential gene expression analysis has been widely reported. Such studies have shown that gene expressions are associated with many common human diseases, such as liver disease (Romeo et al., 2008; Speliotes et al., 2011) and heart failure (Liu et al., 2015). However, there are possibly many unmeasured factors that affect both gene expressions and phenotypes of interest (Hoggart et al., 2003; Leek and Storey, 2007). The existence of such unmeasured confounding variables can cause correlation between the error term and one or some of the independent variables and lead to identifying false associations. Particularly, the independence assumption between gene expressions and errors are required in linear regression in order to obtain valid statistical inference of the effects of gene expressions on phenotype. If this assumption is violated, standard methods can lead to biased estimates (Fan and Liao, 2014; Lin, Feng, and Li, 2015).

One way to deal with unmeasured confounding is to apply instrumental variables (IV) regression, which has been studied extensively in low dimensional settings (Imbens, 2014). In the context of our applications, we treat genetic variants as instrumental variables in studying the association between gene expressions and phenotypes. Standard method to fit the IV models is to apply two-stage regressions to obtain valid estimation of the true parameters. However, in genetical genomics studies, the dimensions of both genetic variants and gene expressions are much larger than the sample sizes, making the classic two-stage regression methods of fitting the IV models infeasible. To account for high dimensionality, penalized regression methods have been developed to select the instruments in the first stage and then to select gene expressions in the second stage (Lin, Feng, and Li, 2015). Lin, Feng, and Li, 2015 provided the estimation error bounds of proposed two-stage estimators but did not study the related problem of statistical inference.

For linear regression models in high-dimensional setting, Javanmard and Montanari, 2014 developed a de-biased procedure to construct an asymptotically normally distributed estimator based on the original biased Lasso estimator. The asymptotic results can be used for hypothesis testing. Zhang and Zhang (2014) proposed a low-dimensional projection estimator to correct the bias, sharing a similar idea as Javanmard and Montanari (2014). In a more general framework, Ning and Liu (2017) considered the hypothesis testing problem for general penalized M-estimator, where they constructed a decorrelated score statistic in high-dimensional setting. All these methods for high dimensional linear regression inference require the critical assumption that the error terms are independent of the covariates, and therefore cannot be applied to the IV models directly.

This chapter presented methods for hypothesis testing for high dimensional IV models, including statistical test of a single regression coefficient and a multiple testing procedure for variable selection. The methods build on the work of Lin, Feng, and Li (2015) to obtain a consistent estimator of the regression coefficients, and the work of Liu (2013) to perform inverse regressions to construct the bias-corrected test statistics. The idea of inverse regression is first used to study the Gaussian graphical model, and has been extended to hypothesis testing problem in high dimensional linear regression (Liu and Luo, 2014). The procedure uses information from the precision matrix so that the correlations between test statistics become quantifiable. We combine this inverse regression procedure with the estimation methods in Lin, Feng, and Li, 2015 to propose a test statistic with desired properties. In addition, in high dimensional setting, the sparsity assumption on the true regression coefficient results in a small number of alternatives, which leads to conservative false discovery rate (FDR) control. A less conservative approach is to control the number of falsely discovered variables (FDV) (Liu and Luo, 2014). The proposed test statistic for single regression coefficient in IV models is shown to be asymptotically normal and the proposed multiple testing procedure is shown to control the FDR or FDV.

4.2. IV Models and Proposed Methodology

4.2.1. Sparse Instrumental Variable Model

Denote $Y \in \mathbb{R}^n$ as the n-dimension phenotype vector, $\mathbf{X} \in \mathbb{R}^{n \times p}$ as the gene expression matrix of p genes and $\mathbf{Z} \in \mathbb{R}^{n \times q}$ as the matrix of q possible instrumental variables such as the genotypes of q genetic variants. Lin, Feng, and Li (2015) considered the following high dimensional IV regression

model:

$$Y = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\eta}, \tag{4.1}$$

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}_0 + \mathbf{E}, \tag{4.2}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the vector of regression coefficients that reflects the association between phenotype Y and gene expression $\mathbf{X}$, while $\boldsymbol{\Gamma}_0$ reveals the relationships between the gene expressions $\mathbf{X}$ and the genetic variants $\mathbf{Z}$. Without lose of generality, we assume $\mathbf{Z}$ is centered and standardized. The error terms $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$ and $\mathbf{E} = (\boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_n})^\top$ are n-dimensional vector and n by p matrix, respectively. The joint distribution of $\left(\boldsymbol{\varepsilon_i}^\top, \eta_i\right)$ is a multivariate normal distribution with mean 0, covariance matrix $\boldsymbol{\Sigma}_e$ and is independent with $\mathbf{Z}$. To emphasize the correlation between Y and $\mathbf{X}$, we assume that the correlation between $\boldsymbol{\varepsilon_i}$ and η_i is not zero. In this chapter we are interested in the high-dimensional setting where the dimension of the covariates p and the dimension of potential instrumental variables q can both be larger than n.

As suggested by Lin, Feng, and Li (2015), estimation of $\boldsymbol{\beta}_0$ in sparse setting can be performed by a two-stage penalized least squares method. To be specific, we first estimate the coefficients matrix $\boldsymbol{\Gamma}_0$ in (4.2) column by column as the following:

$$\widehat{\boldsymbol{\Gamma}}_{\cdot,j} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^q}{\operatorname{argmin}} \left(\frac{1}{2n} \|\mathbf{X}_{\cdot,j} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda_{2j}\|\boldsymbol{\gamma}\|_1 \right), \quad j = 1, 2, \ldots, p, \tag{4.3}$$

where λ_{2j} is a tuning parameter. After obtaining an estimate of $\boldsymbol{\Gamma}_0$, we plug in the predicted value of $\mathbf{X}$, which is $\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}}$, to the second stage model (4.1) and obtain an estimator of $\boldsymbol{\beta}_0$:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{2n} \|Y - \widehat{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \right), \tag{4.4}$$

where λ_1 is a tuning parameter.

The focus of this chapter is to develop statistical test of $\mathcal{H}_0 : \beta_{0i} = 0$ for a given i and to develop a procedure for the multiple hypothesis testing problem:

$$\mathcal{H}_{0i} : \ \beta_{0i} = 0 \quad \text{vs.} \quad \mathcal{H}_{1i} : \ \beta_{0i} \neq 0, \quad i = 1, 2, \ldots, p,$$

with a correct control of FDR or FDV.

4.2.2. Hypothesis Testing for a Single Hypothesis Using Inverse Regression

Denote $\mathbf{D} = \mathbf{Z}\boldsymbol{\Gamma}_0$, from models (4.1) and (4.2),

$$Y = \mu + \mathbf{D}\boldsymbol{\beta}_0 + \boldsymbol{\xi}, \tag{4.5}$$

where $\boldsymbol{\xi} = \boldsymbol{\eta} + \mathbf{E}\beta_0$. When $\mathbf{Z}$ consists of all the valid instruments, $\mathbf{D}$ and $\boldsymbol{\xi}$ are independent by the causal assumptions for a valid instrument and (4.5) can be treated as a standard linear regression. Using the idea of inverse regression (Liu, 2013; Liu and Luo, 2014), for each $i = 1, 2, \ldots, p$, $\mathbf{D}_i$ is regressed on $(Y, \mathbf{D}_{\cdot,-i})$ as:

$$\mathbf{D}_{\cdot,i} = a_i + (Y, \mathbf{D}_{\cdot,-i})\,\boldsymbol{\theta}_i + \boldsymbol{\zeta}_i, \tag{4.6}$$

where $\boldsymbol{\zeta}_i$ satisfies $\mathbb{E}\boldsymbol{\zeta}_i = 0$ and is uncorrelated with $(Y, \mathbf{D}_{\cdot,-i})$. Based on the properties of multivariate normal distribution (Anderson, 2003), the regression coefficient $\boldsymbol{\theta}_i$ is related to the target parameter $\boldsymbol{\beta}_0$ by the following equality:

$$\boldsymbol{\theta}_i = -\sigma^2_{\zeta_i}\left(-\frac{\beta_{0i}}{\sigma^2_\xi}, \frac{\beta_{0i}\boldsymbol{\beta}^\top_{-0i}}{\sigma^2_\xi} + \boldsymbol{\Omega}^{\mathbf{D}}_{-i,i}\right), \tag{4.7}$$

where $\sigma^2_{\zeta_i}$ and σ^2_ξ denote the variance of ζ_i and ξ, respectively, and $\boldsymbol{\Omega}^{\mathbf{D}} = \boldsymbol{\Sigma}^{-1}_{\mathbf{D}}$ is the precision matrix for $\mathbf{D}$. Since $\mathrm{Cov}(\mathbf{D}, \xi) = 0$, we have $\sigma^2_{\zeta_i}\beta_{0i} = \sigma^2_\xi\boldsymbol{\theta}_{i1} = \boldsymbol{\theta}_{i1}\mathrm{Cov}(\xi, y) = -\mathrm{Cov}(\xi, \zeta_i)$, therefore, the null hypothesis $\mathcal{H}_{0i} : \beta_{0i} = 0$ is equivalent to

$$\mathcal{H}_{0i} :\ \mathrm{Cov}(\xi, \zeta_i) = 0 \quad \text{vs.} \quad \mathcal{H}_{1i} :\ \mathrm{Cov}(\xi, \zeta_i) \neq 0, \quad i = 1, 2, \ldots, p.$$

Since the data observed are $\{y_k, \mathbf{X}_k, \mathbf{Z}_k, k = 1, 2, \cdots n\}$, the vector $\mathbf{D}_i$ in (4.6) is not observed for any $i = 1, 2, \ldots, p$. One can estimate $\boldsymbol{\theta}_i$ via regularization by replacing $\mathbf{D}$ with its estimated value $\widehat{\mathbf{D}} = \widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\Gamma}}$,

$$\widehat{\boldsymbol{\theta}_i} = \underset{\boldsymbol{\theta}}{\mathrm{argmin}}\left\{\frac{1}{2n}\|\widehat{\mathbf{D}}_{\cdot,i} - \left(Y, \widehat{\mathbf{D}}_{\cdot,-i}\right)\boldsymbol{\theta}_i\|^2_2 + \mu_i\|\boldsymbol{\theta}\|_1\right\}, \quad i = 1, 2, \ldots, p, \tag{4.8}$$

where μ_i is a tuning parameter.

The sample correlation between ξ and ζ_i is then used to construct the test statistic for $\mathcal{H}_{0i}$ (Liu, 2013). Using the estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{D}}$ and $\widehat{\boldsymbol{\theta}}_i$, the estimated residuals are

$$\begin{aligned}\widehat{\xi}_k &= y_k - \overline{Y} - \left(\widehat{\mathbf{D}}_k - \overline{\widehat{\mathbf{D}}}\right)^\top \widehat{\boldsymbol{\beta}},\\ \widehat{\zeta}_{k,i} &= \widehat{\mathbf{D}}_{k,i} - \overline{\widehat{\mathbf{D}}_i} - \left(y_k - \overline{Y}, \left(\widehat{\mathbf{D}}_{k,-i} - \overline{\widehat{\mathbf{D}}}_{-i}\right)^\top\right)\widehat{\boldsymbol{\theta}}_i,\end{aligned}$$

for $k = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, p$, where

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} y_k,\ \overline{\widehat{\mathbf{D}}} = \frac{1}{n}\sum_{k=1}^{n}\widehat{\mathbf{D}}_k,\ \overline{\widehat{\mathbf{D}}_i} = \frac{1}{n}\sum_{k=1}^{n}\widehat{\mathbf{D}}_{k,i},\ \overline{\widehat{\mathbf{D}}}_{-i} = \frac{1}{n}\sum_{k=1}^{n}\widehat{\mathbf{D}}_{k,-i}.$$

Using the bias correction formula in Liu (2013), for each i, define the test statistic as

$$T_i = \sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}\widehat{\xi}_k\widehat{\zeta}_{k,i} + \frac{1}{n}\sum_{k=1}^{n}\widehat{\xi}_k^2\widehat{\boldsymbol{\theta}}_{1,i} + \frac{1}{n}\sum_{k=1}^{n}\widehat{\zeta}_{k,i}^2\widehat{\boldsymbol{\beta}}_i\right)\Big/\widehat{\sigma}_\xi\widehat{\sigma}_{\zeta_i},$$

where

$$\widehat{\sigma}_\xi^2 = \frac{1}{n}\sum_{k=1}^{n}\widehat{\xi}_k^2, \quad \widehat{\sigma}_{\zeta_i}^2 = \frac{1}{n}\sum_{k=1}^{n}\widehat{\zeta}_{k,i}^2.$$

The bias correction formula adds two extra terms to the original sample correlation in order to eliminate the higher order bias resulting from the bias of the Lasso-type estimator. Using the transformation theorem in Anderson (2003), the final test statistic for testing $\mathcal{H}_{0i}: \operatorname{Cov}(\xi, \zeta_i) = 0$ is defined as

$$\widehat{T}_i = \frac{T_i}{1 - \frac{T_i^2}{n}\mathbf{1}\left(\frac{T_i^2}{n} < 1\right)},$$

which has an asymptotic $N(0,1)$ distribution under the null (see Theorem 4).

4.2.3. Rejection Regions for Multiple Testing Procedure with FDR and FDV control

After obtaining the test statistic $\widehat{T}_i$ for $\mathcal{H}_{0i}$, we determine the rejection region for simultaneous tests of $\widehat{T}_i$ for $\mathcal{H}_{0i}$ for $i = 1, \cdots, p$. Recall that the definitions of FDR and FDV are:

$$FDR = \mathbb{E}\left\{\frac{\sum_{i\in\mathcal{H}_0}\mathbf{1}\left(|\widehat{T}_i| \geq t\right)}{\sum_{i=1}^{p}\mathbf{1}\left(|\widehat{T}_i| \geq t\right)\vee 1}\right\}, FDV = \mathbb{E}\left\{\sum_{i\in\mathcal{H}_0}\mathbf{1}\left(|\widehat{T}_i| \geq t\right)\right\}.$$

Suppose the rejection region for each $\mathcal{H}_{0i}$ is $\{|\widehat{T}_i| \geq t\}$, by the definition of false discovery proportion and false discovery rate, an ideal choice of t that controls the FDR below a certain level α is

$$t_0 = \inf\left\{0 \leq t \leq \sqrt{2\log p} : \frac{\sum_{i\in\mathcal{H}_0}\mathbf{1}\left(|\widehat{T}_i| \geq t\right)}{\sum_{i=1}^{p}\mathbf{1}\left(|\widehat{T}_i| \geq t\right)\vee 1} \leq \alpha\right\}.$$

In practice the quantity $\sum_{i\in\mathcal{H}_0}\mathbf{1}\left(|\widehat{T}_i| \geq t\right)$ can be estimated by $2p\left(1-\Phi(t)\right)$, where $\Phi(t)$ is the cumulative distribution function of the standard normal distribution. Based on this approximation, the quantity t_0 in the multiple testing procedure can be estimated by

$$\widehat{t}_0 = \inf\left\{0 \leq t \leq \sqrt{2\log p} : \frac{2p\left(1-\Phi(t)\right)}{\sum_{i=1}^{p}\mathbf{1}\left(|\widehat{T}_i| \geq t\right)\vee 1} \leq \alpha\right\}. \tag{4.9}$$

We reject the hypothesis $\mathcal{H}_{0i}$ if $|\widehat{T}_i| \geq \widehat{t}_0$ for $i = 1, 2, \ldots, p$.

Similarly, to control the FDV, the rejection region $|\widehat{T}_i| \geq \widehat{t}_0$ is given by

$$\widehat{t}_0 = G^{-1}\left(\frac{k}{p}\right), \tag{4.10}$$

where $G(t) = 2(1-\Phi(t))$.

4.2.4. Implementation

The construction of the test statistics involves a set of convex optimizations and selection of the tuning parameters in order to solve the Lasso regressions (4.3), (4.4) and (4.8). The optimizations can be efficiently implemented using the coordinate descent (CD) algorithm (Friedman, Hastie, and Tibshirani, 2010; Lin, Feng, and Li, 2015). The CD algorithm is a well-known and widely used

convex optimization algorithm for penalized regressions so we omitted the details here.

For tuning parameter selection, we have separate strategies for the two groups of tuning parameters λ and μ. For the optimization problems (4.3) and (4.4), the tuning parameters λ_1 and λ_{2j}, $j = 1, 2, \ldots, p$ can be chosen by a K-fold cross-validation (CV) for $K = 5$ or 10, where λ_1^{opt} and $\lambda_{2j}^{\text{opt}}$, $j = 1, 2, \ldots, p$ are determined by minimizing the CV errors of the corresponding optimization problem. When both p and q are very large, performing CV can be time-consuming. So in our simulations and real data applications, we applied an alternative method for selecting these two groups of tuning parameters that relies on scaled Lasso (Sun and Zhang, 2012), which is computationally more efficient.

Selection of the tuning parameters for the inverse regression (4.8) is done by a data-driven procedure as suggested by Liu (2013) and Liu and Luo (2014). To be specific, let $\delta_j = j$ for $j = 1, 2, \ldots, 100$ and $\mu_j = 0.02\delta_j\sqrt{\widehat{\mathbf{\Sigma}}^D_{i,i}\log p/n}$, where $\widehat{\mathbf{\Sigma}}^D$ is the sample covariance matrix of $\widehat{\mathbf{D}}$. The choice of the δ is determined by:

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}} \sum_{k=30}^{90} \left\{ \frac{\sum_{i=1}^{p} \mathbf{1}\left(|\widehat{T}_i| \geq \Phi^{-1}\left(1 - k/200\right)\right)}{kp/100} - 1 \right\}^2.$$

The tuning parameter μ_i in (4.8) is chosen as $\hat{\mu}_i = 0.02\hat{\delta}\sqrt{\widehat{\mathbf{\Sigma}}^D_{i,i}\log p/n}$.

4.3. Theoretical Results

We provide in this section some theoretical results of the proposed methods. We first restate the estimation error bounds of $\mathbf{\Gamma}_0$ and $\boldsymbol{\beta}_0$ in models (4.1) and (4.2) derived in Lin, Feng, and Li (2015), which are needed in constructing the test statistics. Before stating the results, we first introduce some assumptions. For any matrix $\mathbf{X}$, we say it satisfies the restricted eigenvalue (RE) condition if its restricted eigenvalue is strictly bounded away from 0. That is, for some $1 \leq s \leq p$, the following condition holds:

$$\kappa(s, \mathbf{X}) \triangleq \min_{\substack{J \subseteq \{1,\ldots,p\} \\ |J| \leq s}} \min_{\substack{\boldsymbol{\delta} \neq 0 \\ \|\boldsymbol{\delta}_{J^c}\|_1 \leq 3\|\boldsymbol{\delta}_J\|_1}} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{n}\|\boldsymbol{\delta}_J\|_2} > 0.$$

Denote $s_1 = \|\boldsymbol{\beta}_0\|_0$, $s_2 = \max_j \|\boldsymbol{\Gamma}_{\cdot,j}\|_0$, $r = \max_j \|\boldsymbol{\theta}_j\|_0$ and κ is the restricted eigenvalue defined above. The following assumptions are needed:

(A1) The instrumental variable matrix $\mathbf{Z}$ and matrix $\mathbf{D} = \mathbf{Z}\boldsymbol{\Gamma}_0$ satisfies the restricted eigenvalue condition with some constants $\kappa(s_2, \mathbf{Z}), \kappa(s_1, \mathbf{D}) > 0$, respectively.

(A2) There exists a positive constant C such that $\max\{\|\boldsymbol{\beta}_0\|_1, \|\boldsymbol{\Gamma}_0\|_1, \{\|\boldsymbol{\theta}_i\|_1\}_{i=1,\ldots,p}\} \leq C$.

(A3) There exists a positive constant C such that $\max_{1\leq j\leq p} \left(\boldsymbol{\Sigma}^e_{j,j}\right) \leq C^2$.

(B1) In the inverse regression model (4.6), denote $\mathbf{M}_i = (Y, \mathbf{D}_{\cdot,-i})$, for $i = 1, \ldots, p$, then $\mathbf{M}_i$ satisfies the restricted eigenvalue condition with some constant $\kappa(r, \mathbf{M}_i)$. In addition, assume that there exists a positive constant $\kappa(Y, \mathbf{D})$ such that $\min_i \kappa(r, \mathbf{M}_i) \geq \kappa(Y, \mathbf{D})$.

(C1) The precision matrix $\boldsymbol{\Omega}^{\mathbf{D}}$ and covariance matrix $\Sigma_{\mathbf{D}}$ satisfies $\max_{1\leq j\leq p} \left(\boldsymbol{\Omega}^{\mathbf{D}}_{j,j}, \Sigma^{\mathbf{D}}_{j,j}\right) \leq C$ for some constant C and $\mathrm{Var}(Y_i) \leq C$.

(C2) The dimensional parameters n, p, q, s_1, s_2, r satisfy the following asymptotic scaling condition as $n \to \infty$:

$$\max\{r\sqrt{s_2}, s_1, s_2\}\sqrt{\frac{\log p\,(\log p + \log q)}{n}} = o(1).$$

(C3) The precision matrix $\boldsymbol{\Omega}^{\mathbf{D}}$ satisfies the following condition: for some $\varepsilon > 0$ and $\delta > 0$,

$$\sum_{(i,j)\in\mathcal{A}(\varepsilon)} p^{\frac{2|\rho_{ij,\omega_{\mathbf{D}}}|}{1+|\rho_{ij,\omega_{\mathbf{D}}}|}+\delta} = \mathcal{O}(p^2/(\log p)^2),$$

where $\rho_{ij,\omega_{\mathbf{D}}} = \Omega^{\mathbf{D}}_{ij}/(\Omega^{\mathbf{D}}_{ii}\Omega^{\mathbf{D}}_{jj})^{1/2}$ and $\mathcal{A}(\varepsilon) = \mathcal{B}((\log p)^{-2-\varepsilon})$ with $\mathcal{B}(\delta) = \{(i,j) : |\rho_{ij,\omega_{\mathbf{D}}}| \geq \delta, i \neq j\}$.

These assumptions play different roles in establishing the asymptotic results. To be specific, assumptions (A1) to (A3) are required to obtain the estimation error bounds for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Gamma}}_{\cdot,j}$. These assumptions are similar to those in Bickel, Ritov, and Tsybakov (2009) and are used in Lin, Feng, and Li (2015). They require that matrix $\mathbf{Z}$ and $\mathbf{D}$ are well-behaved and ℓ_1 norms of the true parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\Gamma}_0$ are bounded away from infinity. Assumption (B1) guarantees that $\boldsymbol{\theta}_i$ can be well estimated. This assumption is implicitly assumed, though not stated, in Liu and Luo (2014). Assumptions (C1) and (C2) are needed to obtain the asymptotic distribution of $\widehat{T}_i$. Particularly, assumption (C1)

bounds the entries of the covariance matrix $\Sigma_{\mathbf{D}}$ and precision matrix $\Omega^{\mathbf{D}}$ and assumption (C2) provides the relation among the dimension and sparsity parameters n, p, q, s_1, s_2 and r, where s_1, s_2 and r control the sparsity of $\boldsymbol{\beta}_0$, $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\theta}_i$ respectively. Assumption (C3) is used for controlling the FDR, which imposes some conditions on the precision matrix (Liu and Luo, 2014). In addition, if we fix q, which is the number of instruments, then assumption (C2) is equivalent to $\log p = o(\sqrt{n})$. This assumption is often made in the inference results related with Lasso and other high dimensional models (Gold, Lederer, and Tao, 2017; Javanmard and Montanari, 2014; Ning and Liu, 2017).

4.3.1. Asymptotic distribution of test statistic for single null hypothesis

Since our test statistics rely on the estimation of the parameters in models (4.1) and (4.2), we first provide a lemma on the estimation errors of $\boldsymbol{\Gamma}_{\cdot,j}$ and $\boldsymbol{\beta}$.

Lemma 3 (Estimation error bounds of $\boldsymbol{\Gamma}_{\cdot,j}$ and $\boldsymbol{\beta}_0$ (Lin, Feng, and Li, 2015))**.** *Under assumptions (A1)-(A3), for each $j = 1, 2, \ldots, p$, if the tuning parameter λ_{2j} is chosen as*

$$\lambda_{2j} = \widetilde{C}\sqrt{\frac{\boldsymbol{\Sigma}^e_{j,j}\left(\log p + \log q\right)}{n}},$$

for some $\widetilde{C} \geq 2\sqrt{2}$, then with probability at least $1 - (pq)^{1-\widetilde{C}^2/8}$, $\widehat{\boldsymbol{\Gamma}}$ defined in (4.3) *satisfies*

$$\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\|_1 \leq \frac{16\widetilde{C}C}{\kappa^2(s_2, \mathbf{Z})} s_2 \sqrt{\frac{\log p + \log q}{n}},$$

and

$$\|\mathbf{Z}\left(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}_0\right)\|_F^2 \leq \frac{16\widetilde{C}^2C^2}{\kappa^2(s_2, \mathbf{Z})} s_2 p \left(\log p + \log q\right).$$

Furthermore, if the set of tuning parameters $\{\lambda_{2j} : j = 1, \ldots, p\}$ satisfy

$$\lambda_{\max}(2C + \lambda_{\max}) \leq \frac{\kappa^2(s_2, \mathbf{Z})\kappa^2(s_1, \mathbf{D})}{1024 s_1 s_2},$$

where $\lambda_{\max} = \max_{1 \leq j \leq p} \lambda_{2j}$, if λ_1 is chosen as:

$$\lambda_1 = C_0\sqrt{\frac{s_2\left(\log p + \log q\right)}{n}},$$

then with probability at least $1 - C_1(pq)^{-C_2}$*,* $\widehat{\boldsymbol{\beta}}$ *defined in* (4.4) *satisfies*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 \leq C_3 s_1 \sqrt{\frac{s_2\left(\log p + \log q\right)}{n}},$$

for some positive constants $C_0 - C_3$*.*

In addition, we have the following lemma on the estimation error bound of $\boldsymbol{\theta}_i$.

Lemma 4 (Estimation error bounds of $\boldsymbol{\theta}_i$)**.** *Under assumptions (A1)-(A3) and (B1), for each* $i = 1, 2, \ldots, p$*, there exists some positive constants* C_4, C_5, C_5^**, if the tuning parameter* μ_i *is chosen as*

$$\mu_i = \frac{C_4^*}{\kappa(s_2, \mathbf{Z})} \sqrt{\frac{s_2(\log p + \log q)}{n}},$$

with $C_4^* = C_5^* \max(C, \sigma_{\zeta_i})$*, then with probability at least* $1 - C_4\left(pq\right)^{-C_5}$*,* $\widehat{\boldsymbol{\theta}}_i$ *in* (4.8) *satisfies*

$$\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 \leq \frac{64 C_4^*}{\kappa^2(Y, \mathbf{D})\kappa(s_2, \mathbf{Z})} r \sqrt{\frac{s_2(\log p + \log q)}{n}}.$$

Based on Lemmas 3 and 4, the following theorem provides the asymptotic distribution of the test statistic $\widehat{T}_i$ under the null $\mathcal{H}_{0i}$.

Theorem 4 (Asymptotic distribution of $\widehat{T}_i$)**.** *Under assumptions (A1)-(A3), (B1) and (C1)-(C2), with the proper choices of the tuning parameters* λ_1*,* λ_{2j} *and* μ *as stated in Lemma 3 and 4, for each* $i = 1, 2. \ldots, p$*, under the null* $\mathcal{H}_{0i} : \beta_{0i} = 0$*,*

$$\widehat{T}_i \rightsquigarrow N(0, 1).$$

This null distribution can be used to test the individual null hypothesis $\mathcal{H}_{0i} : \beta_{0i} = 0$.

4.3.2. Theoretical results on FDR and FDV

The next theorem shows that the proposed multiple testing procedure controls the FDR.

Theorem 5 (Asymptotic result for multiple testing procedure)**.** *Denote FDR*$= FDR(\widehat{t}_0)$*, assuming (A1)-(A3), (B1) and (C1), (C3) hold,* $p \leq n^c$ *for some* $c > 0$*. We further assume a condition stronger than* $C2$ *such as the quantities in the left of assumption* $C2$ *are of order* $o((\log p)^{-\frac{1}{2}})$ *instead of* $o(1)$*,*

and for some $\widetilde{c} > 2$*,*

$$\sum_{i \in \mathcal{H}_1} \mathbf{1}\left(\frac{\boldsymbol{\beta}_i}{\sqrt{\sigma_\xi^2 \Omega^{\mathbf{D}}_{i,i}}} \geq \sqrt{\widetilde{c} \log p / n} \right) \to \infty, \tag{4.11}$$

as $(n, p) \to \infty$*. Then with the proper choice of all tuning parameters and the threshold* $\widehat{t}_0$*, with a pre-specified level* α*, we have*

$$\lim_{n,p \to \infty} \frac{FDR}{\alpha p_0 / p} = 1.$$

This theorem indicates that under proper conditions, the empirical FDR is controlled under a pre-specified level. Notice that in addition to the assumptions previously mentioned, we require a stronger condition (4.11). This condition indicates that the number of true alternatives needs to tend to infinity, which is also required in Liu and Luo (2014).

Similar to the result of the FDR but with weaker assumptions, for the FDV control, we have the following result:

Theorem 6 (Asymptotic results for multiple testing procedure)**.** *Assuming (A1)-(A3), (B1) and (C1) hold,* $p \leq n^c$ *for some* $c > 0$ *and we further assume a condition stronger than* $C2$ *such as the quantities in the left of assumption* $C2$ *are of order* $o((\log p)^{-\frac{1}{2}})$ *instead of* $o(1)$*. Then with the proper choice of all tuning parameters and the threshold* $\widehat{t}_0$*, with a pre-specified level* k*, we have:*

$$\lim_{n,p \to \infty} \frac{FDV}{k p_0 / p} = 1. \tag{4.12}$$

Here to control the FDV, we do not need assumption (C3) on the precision matrix and condition (4.11).

4.4. Simulations

We evaluate the performance of the proposed methods through a set of simulations. Following models (4.1) and (4.2), we first generate the instruments matrix $\mathbf{Z}$ where $\mathbf{Z}_i \sim N(0, \mathbf{\Sigma}_z)$. The covariance matrix $\mathbf{\Sigma}_z$ satisfies $(\mathbf{\Sigma}_z)_{ij} = 0.5^{|i-j|}$. For each $\mathbf{\Gamma}_{\cdot,j}$, we first randomly pick s_2 out of q nonzero entries and then each entry is generated randomly from a uniform distribution

$U([-b,-a]\cup[a,b])$ with $a=0.75, b=1$. Parameter $\boldsymbol{\beta}_0$ is generated similarly where we pick s_1 out of p nonzero entries and each entry is generated randomly from $U([-0.3,0.1]\cup[0.1,0.3])$. As for the joint distribution of $\left(\boldsymbol{\varepsilon_i}^\top,\eta_i\right)$, its covariance matrix $\boldsymbol{\Sigma}_e$ is generated by: $(\boldsymbol{\Sigma}_e)_{ij}=0.2^{|i-j|}$ for $1\le i,j\le p$, $(\boldsymbol{\Sigma}_e)_{p+1,p+1}=1$ and among $(\boldsymbol{\Sigma}_e)_{i,p+1}$ where $i=1,\ldots,p$, 10 entires are picked randomly and set to be 0.3. We impose this structure so that η_i is correlated with $\boldsymbol{\varepsilon_i}$. Covariates $\mathbf{X}$ and response Y are generated based on our model. We consider different values of (n,p,q) with $(n,p,q)=(200,100,100),(400,200,200),(200,500,500)$ and $(s_1,s_2)=(10,10)$. We compare our methods with the test developed in Liu and Luo (2014) for high dimensional regression analysis linking Y to $\mathbf{X}$ ignoring the fact that $\mathbf{X}$ and $\boldsymbol{\eta}$ are correlated. It should be noted that the independent error assumption is necessary for the method in Liu and Luo (2014) to work. We evaluated the performances of hypothesis testing procedures by calculating the empirical type-I errors for testing single regression coefficients and eFDR, eFDV for multiple testing procedures. We also evaluated the estimation performances and included the results in the Appendix.

4.4.1. Test of Single Hypothesis

First, to show the validity of the asymptotic distribution of the proposed test statistic $\widehat{T}_i$ for single null hypothesis, we present in Figure A.1 of the Appendix the QQ-plots of the test statistics $\widehat{T}_i$ for several randomly selected covariates over in 500 replications, showing that when using the correct two-stage IV model, the test statistic proposed follows a normal distribution under the null hypothesis (panels (a)-(f)). However, for the covariates with non-zero distribution, the test statistic has a distribution that clearly deviates from the standard normal distribution (panels (g)-(i)).

To demonstrate the importance of applying the IV model when the covariates and the error terms are dependent, Figure A.2 of the Appendix shows the QQ-plots of the same set of variables as in the previous figure for the test statistic of Liu and Luo (2014). For the variables with zero coefficients (panels (a)-(f)), the null distribution of the test statistic clearly deviates from the standard normal distribution for some variables, indicating greater chance of identifying wrong variables.

Figure 4.1 shows the box plots of the empirical type I errors for testing the single null hypothesis for the variables with zero coefficient based on IV models and the standard Lasso regression. When the errors and covariates are correlated due to unobserved confounding, the naive Lasso regression may fail to control the type I error for some null coefficients, leading to inflated type I

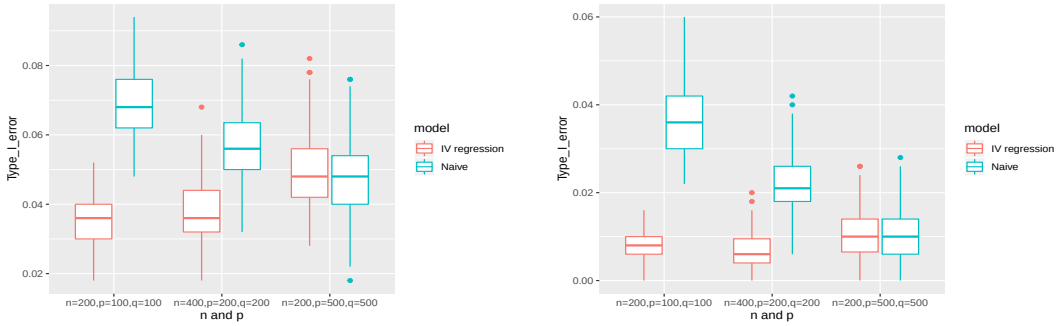


Figure 4.1: Box plots of the empirical type I errors for single hypothesis testing based on IV regression and naive Lasso regression under different settings for α-level of 0.05 (left) and 0.01 (right).

errors. This indicates that the naive method may falsely select some unrelated variables. As a comparison, the test based on the IV regression controls the type-I errors below the specified level.

4.4.2. FDR Controlling for Multiple Testing

To exam the performance of the proposed multiple testing procedure, the empirical FDR, defined as

$$\text{eFDR} = \text{average(FDR)} \quad \text{where FDR} = \frac{\sum_{i\in\mathcal{H}_0} \mathbf{1}\left(|\widehat{T}_i| \geq \widehat{t}_0\right)}{\sum_{i=1}^{p} \mathbf{1}\left(|\widehat{T}_i| \geq \widehat{t}_0\right) \vee 1}, \tag{4.13}$$

is calculated. Similarly, the mean and standard deviation of the power defined as

$$\text{power} = \frac{\sum_{i\in\mathcal{H}_1} \mathbf{1}\left(|\widehat{T}_i| \geq \widehat{t}_0\right)}{|\mathcal{H}_1|}. \tag{4.14}$$

The α-level is chosen to be $\alpha = 0.05, 0.1, 0.2$. Table 4.1 shows the empirical FDR for the proposed procedure using IV regression and the method of Liu and Luo (2014) using naive high dimensional regression models. The proposed multiple test procedure can indeed control the FDR at the correct level. In contrast, test based on naive high dimensional regression fails to control the FDR.

We similarly evaluated the procedure for controlling the number of falsely discovered variables. The empirical FDV is defined as

$$\text{eFDV} = \text{average(FDV)} \quad \text{where FDV} = \sum_{i\in\mathcal{H}_0} \mathbf{1}\left(|\widehat{T}_i| \geq \widehat{t}_{FDV}\right),$$

Table 4.1: Simulation results based on 500 replications. The eFDR and power for multiple testing procedure based on IV regression and naive high dimensional linear regression for different combinations of (n, p, q) and different α levels.

(n, p, q)	α-level	eFDR	power (sd)	eFDR (naive)
$(n, p, q) = (200, 100, 100)$	0.05	0.044	0.547 (0.15)	0.198
	0.10	0.075	0.58 (0.15)	0.239
	0.20	0.134	0.622 (0.15)	0.296
$(n, p, q) = (400, 200, 200)$	0.05	0.026	0.752 (0.13)	0.153
	0.10	0.060	0.781 (0.12)	0.197
	0.20	0.124	0.814 (0.12)	0.268
$(n, p, q) = (200, 500, 500)$	0.05	0.074	0.390 (0.12)	0.055
	0.10	0.129	0.427 (0.13)	0.103
	0.20	0.224	0.472 (0.14)	0.197

Table 4.2: Simulation results based on 500 replications. The eFDV and power for multiple testing procedures based on IV regression and naive high dimensional linear regression for different combinations of (n, p, q) and and different k levels.

(n, p, q)	k-level	eFDV	power (sd)	eFDV (naive)
$(n, p, q) = (200, 100, 100)$	2	1.35	6.35 (1.5)	4.11
	3	1.94	6.57 (1.4)	4.87
	4	2.49	6.71 (1.4)	5.55
$(n, p, q) = (400, 200, 200)$	2	1.27	8.16 (1.1)	4.18
	3	1.94	8.31 (1.1)	5.13
	4	2.59	8.42 (1.1)	5.96
$(n, p, q) = (200, 500, 500)$	2	2.21	4.93 (1.3)	2.04
	3	3.19	5.17 (1.4)	3.01
	4	4.13	5.39 (1.4)	3.98

and its power is given by

$$\text{power} = \sum_{i \in \mathcal{H}_1} \mathbf{1}\left(|\widehat{T}_i| \geq \widehat{t}_{FDV}\right).$$

We consider the k-level of 2,3 and 4. Table 4.2 shows that the proposed procedure also controls the FDV at the specified level. However, naive test that ignoring the covariate-error dependence can result in failing to control the FDV.

It is worth noting that for $p = 500$, the performance of our proposed method is very similar to the naive test. The reason is that by our construction of the covariance matrix of the error terms, the dependency between covariates and errors becomes very week for large p, in which case the two methods are expected to perform similarly.

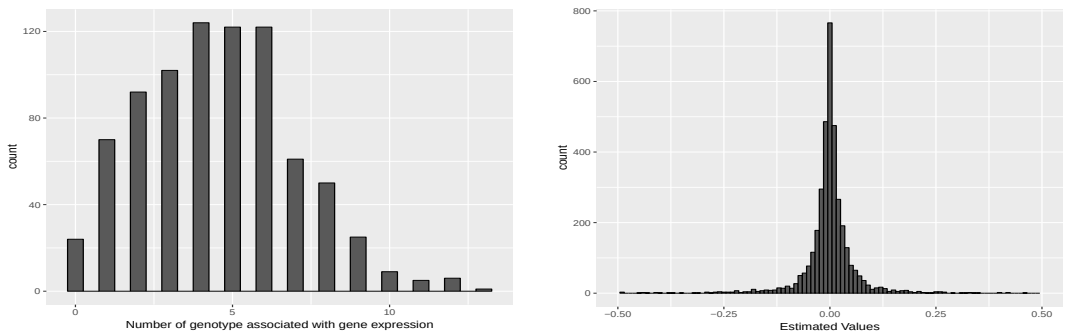


Figure 4.2: Analysis of yeast eQTL data sets, showing the histogram of the number of genotypes associated with each gene expression (left plot) and the histogram of the estimated regression coefficients in the first stage ($\widehat{\mathbf{\Gamma}}$) based on Lasso regressions (right plot).

4.5. Application to a Yeast Data Set

We demonstrate our method using a data set collected on 102 yeast segregants created by crossing of two genetically diverse strains (Brem and Kruglyak, 2005). The data set includes the growth yields of each segregant grown in the presence of different chemicals or small molecule drugs (Perlstein et al., 2007). These segregants have different genotypes represented by 585 markers after removing the markers that are in almost complete linkage disequilibrium. The genotype differences in these strains contribute to rich phenotypic diversity in the segregants. In addition, 6189 yeast genes were profiled in rich media and in the absence of any chemical or drug using expression arrays (Brem and Kruglyak, 2005). Using the same data preprocessing steps as Chen et al., 2009, we compiled a list of candidate gene expression features based on their potential regulatory effects, including transcription factors, signaling molecules, chromatin factors and RNA factors and genes involved in vacuolar transport, endosome, endosome transport and vesicle-mediated transport. We further filtered out the genes with $s.d \leq 0.2$ in expression level, resulting a total of 813 genes in our analysis.

We are interested in identifying the genes whose expression levels are associated with yeast growth yield after being treated with hydrogen peroxide by fitting the proposed two-stage sparse IV model. Figure 4.2 shows the histogram of the number of SNPs selected for each gene expression and the histogram of the estimated regression coefficients ($\mathbf{\Gamma}_0$) from Lasso. These results show that genetic variants are strongly associated with gene expressions and therefore can be used as instrument variables for gene expressions.

Using these selected genotypes as the instrumental variables for each of the gene expressions, we obtained the fitted expression values and applied Lasso with these fitted expressions as predictors and yeast growth yield as the response. For each gene j, we tested the null of $\beta_j = 0$ and obtained its p-value. The 15 significant genes at a nominal $p < 0.05$ are presented in Table 4.3. At FDR< 0.10, three genes were selected. These genes are related with resistance to chemicals, competitive fitness and cell growth, partially explaining their association with the yeast growth in the presence of hydrogen peroxide. For example, among the genes with negative coefficient, over-expression indicates decreased yeast growth. RRM3 gene is involved in DNA replication, and over-expression of the gene leads to abnormal budding and decreased resistance to chemicals. Over-expression of POP5 and FUN26 genes causes decreased vegetative growth rate of yeast (`https://www.yeastgenome.org`).

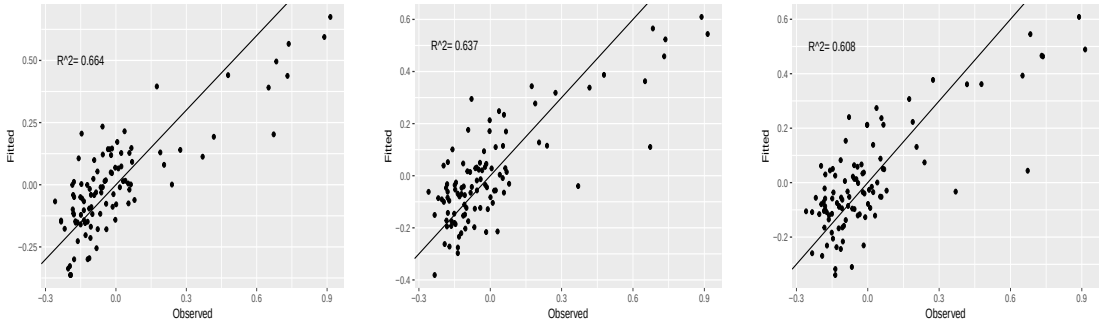
The three selected genes using FDR< 0.10 all had positive coefficients, indicating over-expression of these genes led to increased yeast growth in the presence of hydrogen peroxide. Among these, BDP1 is a general activator of RNA polymerase III transcription and is required for transcription from all three types of polymerase III promoters (Ishiguro, Kassavetis, and Geiduschek, 2002), and over-expression of this gene is expected to increase the yeast viability and growth. PET494 is a mitochondrial translational activator specific for mitochondrial mRNA encoding cytochrome c oxidase subunit III (coxIII) (Marykwas and Fox, 1989). Finally, null mutant of ARG4 gene shows decreased resistance to chemicals (`https://www.yeastgenome.org`) and therefore segregants with higher expression of this gene are expected to have increased resistance to chemicals and increased growth yield.

As a comparison, we also applied Lasso regression with 813 gene expressions as the predictors without using the genotype data. The same statistical test was applied to each of the genes. At a nominal p-value of 0.05, 34 genes were selected by Lasso. However, no gene was selected after adjusting for multiple comparisons with FDR< 0.10. This suggests that by effectively using the genotype data, we were able to identify biologically meaningful genes that are associated with yeast growth in the presence of hydrogen peroxide.

We further compared the model fits by calculating the R^2 statistics in three different scenarios. The first scenario is to use the 15 genes selected using our proposed multiple testing method and refit a linear model with the estimated $\widehat{\mathbf{X}}$. The second scenario is use the 34 genes identified by naive test

Table 4.3: Results from analysis of yeast growth yield data. Table shows the selected genes using single test statistics ($p < 0.05$) and multiple testing procedure with FDR< 0.10 and FDV< 2(marked by $*$). The gene names and estimated regression coefficients and refitted values are listed.

Gene id	Gene name	$\hat{\boldsymbol{\beta}}$	Refitted $\hat{\boldsymbol{\beta}}$
	Negative coefficient		
YHR031C	RRM3	-3.82	-5.00
YAL033W	POP5	-0.22	-0.69
YLR275W	SMD2	-0.20	-0.31
YNL236W	SIN4	-4.67	-5.63
YNL138W	SRV2	-0.63	-1.68
YNL146W	YNL146W	-0.24	-0.12
YAR035W	YAT1	-1.74	-2.79
YAL022C	FUN26	-2.89	-4.79
YHL018W	YHL018W	-0.79	-2.29
	Positive coefficient		
YNL331C	AAD14	0.07	0.17
YHR014W	SPO13	0.47	2.20
YHR018C*	ARG4	0.22	0.34
YHR097C	YHR097C	0.06	0.15
YNL039W*	BDP1	1.82	3.96
YNR045W*	PET494	0.70	0.86



(a) (b) (c)

Figure 4.3: Scatter-plots of the fitted versus the observed yeast growth yield. (a): refitted model using the estimated expression levels of the 15 genes selected by our proposed method; (b): refitted model using expression levels of 34 genes selected using naive test; (c): the refitted model using expression levels of genes selected based on Lasso.

and refit a linear model using the original $\mathbf{X}$. The last scenario is use the genes selected by Lasso using $\mathbf{X}$ and refit a linear model with the original $\mathbf{X}$. Figure 4.3 shows that our method provides the highest R^2 value among the three, with a value of 0.664, indicating that using refitted $\mathbf{X}$ can lead to better fit of the data.

4.6. Discussion

We have developed methods for exploring the association between gene expression and phenotype in the framework IV regression when there are possible unmeasured confounders. Here the genetic variants are used as possible instrumental variables. We have constructed a test statistic using the idea of inverse regression and derived its asymptotic null distribution. We have further developed a multiple testing procedure for the high-dimensional two stage least square methods and provided the rejection region of multiple testing that controls the false discovery rate or number of falsely discovered variables. Both theoretical results and simulations have shown the correctness of our procedure and improved performance over the Lasso regression.

For the yeast genotype and gene expression data, our two-stage regression method was able to identify three yeast genes whose expressions were associated growth in the presence of hydrogen peroxide. In contrast, using gene expression data alone and Lasso regression did not identify any growth associated genes. Since growth yield is highly inheritable (Perlstein et al., 2007), using genotype-predicted gene expressions in our two-stage estimation can help to identify the gene expressions that might be causal to the phenotype. For model organisms such as yeast, the conditional independence assumption between the genotypes and the outcome given gene expression levels is expected to hold. However, for human studies, one should be cautious of such an assumption since genetic variants can affect phenotype via other mechanisms such as changing protein structures.

One possible application of the proposed two-stage regression is to identify gene expressions that cause diseases by jointly analyzing genotype and gene expression data. This is similar in spirit to PredXscan (Gamazon et al., 2015) that aims to identify the molecular mechanisms through which genetic variation affects phenotype. PredXscan builds gene expression prediction models using reference eQTL data. In contrast, our method requires that the genotype and gene expression data are measured on the same set of individuals.

Potential extensions of this method include detecting and accounting for the existence of weak instrumental variables and developing methods that are robust to the residual distributions. Recent papers such as Chatterjee and Lahiri (2010) and Dezeure, Bühlmann, and Zhang (2017) developed bootstrapping inference methods for Lasso estimator. It is possible to apply such ideas to the high

dimensional IV model considered in this chapter. Besides the two-stage least square method we developed here, an alternative to estimating the parameters in IV model is by estimating equations. The two-stage least square methods provides optimal estimator under proper model assumptions while the estimating equation is expected to be robust. The problem of testing a single parameter using estimating equation under high-dimensional setting has been explored by Neykov et al. (2018). It is interesting to consider the multiple testing procedure when estimating equations are used for estimating the parameters in high-dimensional IV models. Another potential direction of extension is to consider the existence of invalid instruments. As far as we knew, there are existing paper considering the problem of having potential invalid instrument Kang et al. (2016). But in their paper the number of covariates is fixed and small, which is different from our setup. It would be interesting to extend their methods to high-dimensional covariates and multiple testing problem.

CHAPTER 5

DISCUSSION

In this thesis, we considered several research problems related to genomics and microbiome. Statistical inference methods were developed for high-dimensional models and were applied to large-scale and complex-structured datasets. In Chapter 2 and Chapter 3, we focused on the inference problem for regression models with linear constraints on the regression coefficients. A de-biased and post-selection inference procedure were introduced respectively. These methods were applied to the PLEASE study and UK twins study. And in Chapter 4, we proposed a statistical testing procedure for the high dimensional IV model. This model is often applied to explore the association between gene expression and phenotype using genetic variants as instruments.

For further research projects, I have two direction of interests. The first one is the covariance matrix estimation problem for multi-omics data. With the observed compositional data, naive sample covariance matrix is biased towards to true covariance matrix obtained via the true unobserved abundance. Cao, Lin, and Li, 2019 introduced a composition-adjusted thresholding method to estimate the true covariance matrix. There is a similar problem when estimating the joint covariance matrix of compositional data and data from other sources (for example, gene expression data or metabolic data). The existing method, however, is not applicable to estimate the partial covariance matrix (off-diagonal part). By applying centered log-ratio transformation, there is a connection between the true partial covariance matrix and its sample version obtained via the transformed data. This connection could be used to develop new methods. The second is the modeling of microbiome data. Currently the most common way of studying the microbiome is through the compositional data. Certain transformation is applied to the compositional data (such as the central-log-transformation) and statistical analysis may need special care (such as imposing constraints). There are some known problems with these approaches including handling zero-values and existing methods haven't fully characterized the geometry structures of microbiome data. The compositional nature of the data links it to statistical literature such as spherical data analysis, non-Euclidian data, phylogenetic-tree-based models and topological data analysis. This motivates me to consider alternative methods of modeling the microbiome data which might gain extra benefits by imposing more complex structures.

APPENDIX

PROOFS AND ADDITIONAL SIMULATIONS

A.1. Proofs for Chapter 2

We provided proofs for the main theorems in this Chapter. Before that, we first introduced a lemma.

Lemma 5. *If Conditions C1 and C2 hold, then for any matrix A,*

$$|(\mathbf{I}_p - P_C)A|_\infty \leq k_0|A|_\infty.$$

The proof for this lemma is in the appendix of Shi, Zhang, and Li (2016).

Proof of Lemma 1. We first provided a bound for $\mathbf{\Sigma}$. Notice that:

$$\begin{aligned}\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Sigma} - (\mathbf{I}_p - P_C) &= \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{\Omega}_{\boldsymbol{\beta}} v(\boldsymbol{\beta}, \widetilde{Z}_k)\widetilde{Z}_k\widetilde{Z}_k^\top - (\mathbf{I}_p - P_C)\right), \\ &= \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2} v(\boldsymbol{\beta}, \widetilde{Z}_k)\widetilde{Z}_k\widetilde{Z}_k^\top\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}\mathbf{\Sigma}_{\boldsymbol{\beta}}^{1/2} - (\mathbf{I}_p - P_C)\right).\end{aligned}$$

The last equality is true as $\mathbf{\Sigma}_{\boldsymbol{\beta}}^{1/2}\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}\widetilde{Z}_k = (\mathbf{I}_p - P_C)\widetilde{Z}_k = \widetilde{Z}_k$ for $k = 1, 2\ldots, n$. Then notice that $\mathbb{E}\mathbf{\Omega}_{\boldsymbol{\beta}} v(\boldsymbol{\beta}, \widetilde{Z}_k)\widetilde{Z}_k\widetilde{Z}_k^\top = \mathbb{E}\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Sigma}_{\boldsymbol{\beta}} = \mathbf{I}_p - P_C$, so define:

$$v_k^{(ij)} = \mathbf{\Omega}_{i,\cdot}^{1/2}\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2} v(\boldsymbol{\beta}, \widetilde{Z}_k)\widetilde{Z}_k\widetilde{Z}_k^\top\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}(\mathbf{\Sigma}_{\boldsymbol{\beta}})_{\cdot,j}^{1/2} - (\mathbf{I}_p - P_C)_{i,j},$$

we know that $\mathbb{E}v_k^{(ij)} = 0$ for $k = 1, 2\ldots, n$ and any i, j. Then by the proof of Lemma 6.2 in Javanmard and Montanari (2014), we have:

$$\begin{aligned}\|v_k^{(ij)}\|_{\psi_1} &\leq 2\|\mathbf{\Omega}_{i,\cdot}^{1/2}\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2} v(\boldsymbol{\beta}, \widetilde{Z}_k)\widetilde{Z}_k\widetilde{Z}_k^\top\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}(\mathbf{\Sigma}_{\boldsymbol{\beta}})_{\cdot,j}^{1/2}\|_{\psi_1}, \\ &\leq 2v(\boldsymbol{\beta}, \widetilde{Z}_k)\|\mathbf{\Omega}_{i,\cdot}^{1/2}\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}\widetilde{Z}_k\|_{\psi_2}\|(\mathbf{\Sigma}_{\boldsymbol{\beta}})_{\cdot,j}\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}\widetilde{Z}_k\|_{\psi_2}, \\ &\leq 2\|(\mathbf{\Sigma}_{\boldsymbol{\beta}})_{\cdot,j}\|_2\|\mathbf{\Omega}_{i,\cdot}^{1/2}\|_2 \cdot \|\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}\widetilde{Z}_k\|_{\psi_2}\|\mathbf{\Omega}_{\boldsymbol{\beta}}^{1/2}\widetilde{Z}_k\|_{\psi_2}, \\ &\leq 2\sqrt{C_{\mathsf{max}}/C_{\mathsf{min}}}\kappa^2 \equiv \kappa_1'.\end{aligned}$$

Then by inequality for centered sub-exponential random variables from Bühlmann and Van De Geer

(2011), we have:

$$\mathbb{P}\left(\frac{1}{n}|\sum_{k=1}^{n} v_k^{(ij)}| \geq \gamma\right) \leq \exp\left[-\frac{n}{6}\min\left\{\left(\frac{\gamma}{e\kappa'}\right)^2, \left(\frac{\gamma}{e\kappa'}\right)\right\}\right].$$

Pick $\gamma = c\sqrt{(\log p)/n}$ with $c \leq e\kappa_1'\sqrt{n/(\log p)}$, we have:

$$\mathbb{P}\left(\frac{1}{n}|\sum_{k=1}^{n} v_k^{(ij)}| \geq c\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-c^2/(6e^2\kappa_1'^2)} = 2p^{-c^2 C_{\min}/(24e^2 C_{\max}\kappa^4)}. \tag{A.1}$$

Since (A.1) is true for all i, j, we have:

$$\mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Sigma} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}\right) \leq 2p^{-c^2 C_{\min}/(24e^2 C_{\max}\kappa^4)+2} = 2p^{-c_1''}.$$

Then by the following inequality:

$$\begin{aligned}
&\mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}\right) \\
&\leq \mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Sigma} - (\mathbf{I}_p - P_C)|_\infty + |\mathbf{\Omega}_{\boldsymbol{\beta}}(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})| \geq c\sqrt{(\log p)/n}\right) \\
&\leq \mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Sigma} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}\right) + \mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})| \geq c\sqrt{(\log p)/n}\right)
\end{aligned}$$

Notice that:

$$\begin{aligned}
\left|\mathbf{\Omega}_{\boldsymbol{\beta}}(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})\right|_\infty &= \frac{1}{n}\left|\sum_{k=1}^{n}\left(\mathbf{\Omega}_{\boldsymbol{\beta}}\left(v(\boldsymbol{\beta}, \widetilde{Z}_k) - v(\hat{\boldsymbol{\beta}}^n, \widetilde{Z}_k)\right)\widetilde{Z}_k\widetilde{Z}_k^\top\right)\right|_\infty \\
&\leq \frac{1}{n}\left|\sum_{k=1}^{n}\left(C\|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1\mathbf{\Omega}_{\boldsymbol{\beta}}\widetilde{Z}_k\widetilde{Z}_k^\top\right)\right|_\infty
\end{aligned}$$

As

$$\frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{\Omega}_{\boldsymbol{\beta}}\widetilde{Z}_k\widetilde{Z}_k^\top\right) \to \mathbb{E}\mathbf{\Omega}_{\boldsymbol{\beta}}\widetilde{Z}_1\widetilde{Z}_1^\top = \mathbb{E}\mathbf{\Omega}_{\boldsymbol{\beta}}\Theta,$$

together with the result we obtain from theorem 1,

$$\begin{aligned}
\mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})|_\infty \geq c\sqrt{(\log p)/n}\right) &\leq \mathbb{P}\left(\frac{1}{n}\left|\sum_{k=1}^{n}\left(C\|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1\mathbf{\Omega}_{\boldsymbol{\beta}}\widetilde{Z}_k\widetilde{Z}_k^\top\right)\right|_\infty \geq c\sqrt{(\log p)/n}\right) \\
&\leq 2p^{1-\hat{c}^2/(2K^2)} = 2p^{-c_2''}
\end{aligned}$$

where $\hat{c} = \frac{c\phi_0}{C|\mathbf{\Omega}_{\boldsymbol{\beta}}\Theta|_\infty s(k_0\tau+1)}$. So finally:

$$\begin{aligned}
&\mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}\right) \\
&\leq \mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}\mathbf{\Sigma} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}\right) + \mathbb{P}\left(|\mathbf{\Omega}_{\boldsymbol{\beta}}(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})|_\infty \geq c\sqrt{(\log p)/n}\right) \\
&\leq 2p^{-c_1^{''}} + 2p^{-c_2^{''}}
\end{aligned}$$

□

Proof of Theorem 1. By the definition of $\hat{\boldsymbol{\beta}}^n$ and (2.5), we have:

$$-\frac{1}{n}\left\{Y^\top \widetilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}^n - \sum_{i=1}^n A(\widetilde{Z}_i^\top \hat{\boldsymbol{\beta}}^n)\right\} + \lambda||\hat{\boldsymbol{\beta}}^n||_1 \leq -\frac{1}{n}\left\{Y^\top \widetilde{\mathbf{Z}}\boldsymbol{\beta} - \sum_{i=1}^n A(\widetilde{Z}_i^\top \boldsymbol{\beta})\right\} + \lambda||\boldsymbol{\beta}||_1. \tag{A.2}$$

Denote $h = \hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}$, and S_h be the set of index of the s largest absolute values of h. Then rearrange (A.2), we get:

$$\lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}^n\|_1) \geq -\frac{1}{n}\left[Y^\top \widetilde{\mathbf{Z}}h - \sum_{i=1}^n \left\{A(\widetilde{Z}_i^\top \hat{\boldsymbol{\beta}}^n) - A(\widetilde{Z}_i^\top \boldsymbol{\beta})\right\}\right]. \tag{A.3}$$

Notice that,

$$\begin{aligned}
\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}^n\|_1 =& \|\boldsymbol{\beta}_{supp(\boldsymbol{\beta})}\|_1 - \|\hat{\boldsymbol{\beta}}^n_{supp(\boldsymbol{\beta})}\|_1 - \|\hat{\boldsymbol{\beta}}^n_{supp(\boldsymbol{\beta})^c}\|_1, \\
\leq& \|\boldsymbol{\beta}_{supp(\boldsymbol{\beta})} - \hat{\boldsymbol{\beta}}^n_{supp(\boldsymbol{\beta})}\|_1 - \|h_{supp(\boldsymbol{\beta})^c}\|_1, \\
\leq& \|h_{S_h}\|_1 - \|h_{S_h^c}\|_1.
\end{aligned} \tag{A.4}$$

Furthermore, for each i applied the mean value theorem to A defined in 2.3, there exists $\widetilde{\beta}_i^{\,0}$ such that $A(\widetilde{Z}_i^\top \hat{\beta}^n) - A(\widetilde{Z}_i^\top \beta) = \mu(\widetilde{\beta}, \widetilde{Z}_i)\widetilde{Z}_i^\top h + \frac{1}{2}v(\widetilde{\beta}_i^{\,0}, \widetilde{Z}_i)\left(\widetilde{Z}_i^\top h\right)^2$. Then we have:

$$\begin{aligned}
&-\frac{1}{n}\left[Y^\top \widetilde{\mathbf{Z}}h - \sum_{i=1}^n \left\{A(\widetilde{Z}_i^\top \hat{\beta}^n) - A(\widetilde{Z}_i^\top \beta)\right\}\right] \qquad (A.5) \\
&\geq -\frac{1}{n}\left\{Y^\top \widetilde{\mathbf{Z}}h - \mu(\beta, \widetilde{\mathbf{Z}})^\top \widetilde{\mathbf{Z}}h\right\}, \\
&\geq -\frac{1}{n}(Y - \mu(\beta, \widetilde{\mathbf{Z}}))^\top \widetilde{\mathbf{Z}}h, \\
&\geq -\frac{1}{n}\|Y - \mu(\beta, \widetilde{\mathbf{Z}})^\top \widetilde{\mathbf{Z}}\|_\infty \cdot \|h\|_1 = -\frac{1}{n}\|(Y - \mu(\beta, \widetilde{\mathbf{Z}}))^\top \widetilde{\mathbf{Z}}\|_\infty \cdot (\|h_{S_h}\|_1 + \|h_{S_h^c}\|_1).
\end{aligned}$$

When the event $\|(Y-\mu(\beta,\widetilde{\mathbf{Z}}))^\top\widetilde{\mathbf{Z}}\|_\infty \le \frac{n\lambda}{\tau}$ holds, we have:

$$\lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}^n\|_1) \ge -\frac{1}{n}\cdot\frac{n\lambda}{\tau}\cdot(\|h_{S_h}\|_1 + \|h_{S_h^c}\|_1). \tag{A.6}$$

So by (A.3), (A.4) and (A.6) we have:

$$\lambda(\|h_{S_h}\|_1 - \|h_{S_h^c}\|_1) \ge \lambda(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}^n\|_1) \ge -\frac{\lambda}{\tau}\cdot(\|h_{S_h}\|_1 + \|h_{S_h^c}\|_1).$$

That is,

$$\|h_{S_h^c}\|_1 \le \frac{\tau+1}{\tau-1}\|h_{S_h}\|_1. \tag{A.7}$$

Then by the KKT condition of optimization problem (2.5), we have:

$$\|\widetilde{\mathbf{Z}}^\top(Y-\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}})) + C\boldsymbol{\eta}\|_\infty \le n\lambda, \tag{A.8}$$

for some $\boldsymbol{\eta}\in\mathbb{R}^r$. Then by Lemma 1,

$$\|(\mathbf{I}_p - P_C)\left(\widetilde{\mathbf{Z}}^\top(Y-\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}})) + C\boldsymbol{\mu}\right)\|_\infty \le k_0\|\widetilde{\mathbf{Z}}^\top(Y-\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}})) + C\boldsymbol{\mu}\|_\infty \le k_0 n\lambda. \tag{A.9}$$

Then as

$$\begin{aligned}(\mathbf{I}_p - P_C)(\widetilde{\mathbf{Z}}^\top(Y-\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}})) + C\boldsymbol{\mu}) &= (\mathbf{I}_p - P_C)\widetilde{\mathbf{Z}}^\top(Y-\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}})) + (\mathbf{I}_p - P_C)C\boldsymbol{\mu},\\ &= \widetilde{\mathbf{Z}}^\top(Y-\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}})).\end{aligned}$$

with the the assumption that $\|(Y-\mu(\boldsymbol{\beta},\widetilde{\mathbf{Z}}))^\top\widetilde{\mathbf{Z}}\|_\infty \le \frac{n\lambda}{\tau}$, we have:

$$\|\widetilde{\mathbf{Z}}^\top(\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}}) - \mu(\boldsymbol{\beta},\widetilde{\mathbf{Z}}))\| \le \|\widetilde{\mathbf{Z}}^\top(Y-\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}}))\|_\infty + \|\widetilde{\mathbf{Z}}^\top(Y-\mu(\boldsymbol{\beta},\widetilde{\mathbf{Z}}))\|_\infty \le k_0 n\lambda + \frac{n\lambda}{\tau}.$$

As $\|\widetilde{\mathbf{Z}}^\top(\mu(\hat{\boldsymbol{\beta}}^n,\widetilde{\mathbf{Z}}) - \mu(\boldsymbol{\beta},\widetilde{\mathbf{Z}}))\| = \|\widetilde{\mathbf{Z}}^\top\mathbf{V}(\boldsymbol{\beta}^0,\widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}}h\|_\infty$, we get

$$\|\widetilde{\mathbf{Z}}^\top\mathbf{V}(\boldsymbol{\beta}^0,\widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}}h\|_\infty \le k_0 n\lambda + \frac{n\lambda}{\tau}.$$

Since $\mathbf{V}(\boldsymbol{\beta}^0,\widetilde{\mathbf{Z}})$ is a diagonal matrix with all its nonzero elements greater than zero, define $\widetilde{\mathbf{Z}}_v =$

$\mathbf{V}^{\frac{1}{2}}(\boldsymbol{\beta}^0, \widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}}$, where $\mathbf{V}^{\frac{1}{2}}(\boldsymbol{\beta}^0, \widetilde{\mathbf{Z}}) = \mathrm{diag}\{(v(\boldsymbol{\beta}^0, Z_1))^{\frac{1}{2}}, \ldots, (v(\boldsymbol{\beta}^0, Z_n))^{\frac{1}{2}}\}$. So $\widetilde{\mathbf{Z}}_v^\top \widetilde{\mathbf{Z}}_v = \widetilde{\mathbf{Z}}^\top \mathbf{V}(\boldsymbol{\beta}^0, \widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}}$. Using Lemma 5.1 in Cai and Zhang (2013), we have:

$$\begin{aligned}
|\langle \widetilde{\mathbf{Z}}_v h_{S_h}, \widetilde{\mathbf{Z}}_v h_{S_h^c} \rangle| &\leq \theta_{s,s}(\widetilde{\mathbf{Z}}_v)\|h_{S_h}\|_2 \cdot \max(\|h_{S_h^c}\|_\infty, \|h_{S_h^c}\|_1/s)\sqrt{s}, \\
&\leq \sqrt{s}\theta_{s,s}(\widetilde{\mathbf{Z}}_v)\|h_{S_h}\|_2 \cdot \frac{\tau+1}{\tau-1}\|h_{S_h}\|_1/s, \\
&\leq \frac{\tau+1}{\tau-1}\theta_{s,s}(\widetilde{\mathbf{Z}}_v)\|h_{S_h}\|_2^2.
\end{aligned}$$

Then,

$$\begin{aligned}
(k_0 n\lambda + \frac{n\lambda}{\tau})\|h_{S_h}\|_1 &\geq \|\widetilde{\mathbf{Z}}^\top \mathbf{V}(\boldsymbol{\beta}^0, \widetilde{\mathbf{Z}})\widetilde{\mathbf{Z}}h\|_\infty \|h_{S_h}\|_1 \geq \langle \widetilde{\mathbf{Z}}_v^\top \widetilde{\mathbf{Z}}_v h, h_{S_h} \rangle, \\
&= \langle \widetilde{\mathbf{Z}}_v h_{S_h}, \widetilde{\mathbf{Z}}_v h_{S_h} \rangle + \langle \widetilde{\mathbf{Z}}_v h_{S_h}, \widetilde{\mathbf{Z}}_v h_{S_h^c} \rangle, \\
&\geq \|\widetilde{\mathbf{Z}}_v h_{S_h}\|_2^2 - \frac{\tau+1}{\tau-1}\theta_{s,s}(\widetilde{\mathbf{Z}}_v)\|h_{S_h}\|_2^2, \\
&\geq \left(\delta_{2s}^-(\widetilde{\mathbf{Z}}_v) - \frac{\tau+1}{\tau-1}\theta_{s,s}(\widetilde{\mathbf{Z}}_v)\right)\|h_{S_h}\|_2^2, \\
&\geq \left(\frac{3\tau-1}{2(\tau-1)}\delta_{2s}^-(\widetilde{\mathbf{Z}}_v) - \frac{\tau+1}{2(\tau-1)}\delta_{2s}^+(\widetilde{\mathbf{Z}}_v)\right)\|h_{S_h}\|_1^2/s. \qquad \text{(A.10)}
\end{aligned}$$

So from (A.10) we have:

$$\begin{aligned}
\|h_{S_h}\|_1 &\leq \frac{s\left(k_0 n\lambda + \frac{n\lambda}{\tau}\right)}{\left(\frac{3\tau-1}{2(\tau-1)}\delta_{2s}^-(\widetilde{\mathbf{Z}}_v) - \frac{\tau+1}{2(\tau-1)}\delta_{2s}^+(\widetilde{\mathbf{Z}}_v)\right)}, \\
&\leq s\frac{k_0 n\lambda + \frac{n\lambda}{\tau}}{2n\tau\phi_0/(\tau-1)}. \qquad \text{(A.11)}
\end{aligned}$$

So combine (A.7) and (A.11), we have:

$$\|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1 = \|h_{S_h}\|_1 + \|h_{S_h^c}\|_1 \leq \frac{2\tau}{\tau-1}\|h_{S_h}\|_1 \leq \frac{s\lambda(k_0 + 1/\tau)}{\phi_0}.$$

Take $\lambda = \tau\tilde{c}\sqrt{(\log p)/n}$, so we have:

$$\begin{aligned}
\mathbb{P}\left(\|\hat{\beta}^n - \beta\|_1 \leq \frac{s\lambda(k_0 + 1/\tau)}{\phi_0}\right) &\geq 1 - \mathbb{P}\left(\|(Y - \mu(\beta, \widetilde{\mathbf{Z}}))^\top \widetilde{\mathbf{Z}}\|_\infty > \frac{n\lambda}{\tau}\right) \\
&\geq 1 - \sum_{i=1}^{p} \mathbb{P}\left(|((Y - \mu(\beta, \widetilde{\mathbf{Z}}))^\top \widetilde{\mathbf{Z}})_i| > \frac{n\lambda}{\tau}\right) \\
&\geq 1 - 2\sum_{i=1}^{p} \exp\left(-\frac{(\sqrt{n}\lambda/\tau}{2K^2}\right) \geq 1 - 2p^{1-\tilde{c}^2/(2K^2)}
\end{aligned}$$

□

Proof of Theorem 2. As we obtained in lemma 1, Ω_β is in the feasible set with a large probability. That is, event $|M\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}$ happens with large probability. Further more,

$$\begin{aligned}
\mathbb{P}\left(|(\mathbf{I}_p - P_C) - M\widehat{\mathbf{\Sigma}}^0|_\infty \geq c\sqrt{(\log p)/n}\right) &\leq \mathbb{P}\left(|M\widehat{\mathbf{\Sigma}} - (\mathbf{I}_p - P_C)|_\infty \geq c\sqrt{(\log p)/n}\right) \\
&\quad + \mathbb{P}\left(|M(\widehat{\mathbf{\Sigma}}^0 - \widehat{\mathbf{\Sigma}})|_\infty \geq c\sqrt{(\log p)/n}\right).
\end{aligned}$$

The bound for the first term on the RHS is the result from lemma 1. Applying the similar method to the second term, notice that $\|\hat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\|_1 \leq \|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1$, hence, $\mathbb{P}(|M(\widehat{\mathbf{\Sigma}}^0 - \widehat{\mathbf{\Sigma}})|_\infty \geq c\sqrt{(\log p)/n}) \leq 4p^{-c_2''}$. So,

$$\mathbb{P}\left(|(\mathbf{I}_p - P_C) - M\widehat{\mathbf{\Sigma}}^0|_\infty \geq c\sqrt{(\log p)/n}\right) \leq 2p^{-c_1''} + 6p^{-c_2''}$$

Finally,

$$\begin{aligned}
\|\Delta\|_\infty &\leq \sqrt{n}\left|(\mathbf{I}_p - P_C) - \widetilde{M}\widehat{\mathbf{\Sigma}}^0\right|_\infty \|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1 \\
&= \sqrt{n}\left|(\mathbf{I}_p - P_C)\left((\mathbf{I}_p - P_C) - \widetilde{M}\widehat{\mathbf{\Sigma}}^0\right)\right|_\infty \|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1 \\
&\leq k_0\sqrt{n}|(\mathbf{I}_p - P_C) - M\widehat{\mathbf{\Sigma}}^0|_\infty \|\hat{\boldsymbol{\beta}}^n - \boldsymbol{\beta}\|_1
\end{aligned}$$

We have:

$$
\begin{aligned}
&\mathbb{P}\left(\|\Delta\|_\infty > \frac{c\tilde{c}k_0(k_0\tau+1)}{\phi_0}\cdot\frac{s\log p}{\sqrt{n}}\right)\\
&\leq \mathbb{P}\left(\|\hat{\boldsymbol{\beta}}^n-\boldsymbol{\beta}\|_1 \geq \frac{s\lambda(k_0+1/\tau)}{\phi_0} = \frac{s\tilde{c}(k_0\tau+1)\sqrt{(\log p)/n}}{\phi_0}\right)\\
&+\mathbb{P}\left(|(\mathbf{I}_p-P_C)-M\widehat{\boldsymbol{\Sigma}}^0|_\infty \geq \gamma = c\sqrt{(\log p)/n}\right)\\
&\leq 2p^{-c'}+2p^{-c_1^{''}}+6p^{-c_2^{''}}
\end{aligned}
$$

So we have finished the proof. □

A.2. Additional simulation results for Chapter 2

A.2.1. Sensitivity analysis to zero replacement

Table A.1 compares the parameter estimates by replacing zeros with 0.5 or 0.1 times the minimum nonzero abundance.

Table A.1: Comparisons of parameter estimates and CIs using two different methods of replacing zeros. Selected bacteria and their estimated coefficients (standard errors in the parenthesis) and 95% confidence intervals.

	0.5		0.1	
Bacteria name	β(se)	CI	β(se)	CI
Prevotella_copri	$-0.15(0.042)$	$(-0.23,-0.064)$	$-0.13(0.036)$	$(-0.20,-0.061)$
Ruminococcus_bromii	$-0.22(0.043)$	$(-0.31,-0.18)$	$-0.20(0.038)$	$(-0.27,-0.12)$
Clostridium_leptum	$-0.15(0.052)$	$(-0.25,-0.048)$	$-0.12(0.043)$	$(-0.20,-0.033)$
Escherichia_coli	$0.14(0.035)$	$(0.074,0.21)$	$0.13(0.029)$	$(0.066,0.18)$
Ruminococcus_gnavus	$0.13(0.045)$	$(0.043,0.22)$	$0.11(0.039)$	$(0.036,0.19)$

From the result we could see that the two methods of replacing zeros are consistent in variable selection. And their estimates together with the confidence intervals are very closed to each other. This indicates that the estimating procedure is robust to the replacing method.

A.3. Proofs for Chapter 3

A.3.1. Proof of Lemma 2

Consider the KKT conditions of the optimization problem (3.6). Any solution of (3.6) satisfies the following:

$$\begin{aligned}
&\mathbf{X}^\top(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{y}) + \lambda\nu + C\eta = 0,\\
&\mathbf{C}^\top\widehat{\boldsymbol{\beta}} = 0,\\
&\nu_j = \mathsf{sign}(\boldsymbol{\beta}_j), \text{ if } \boldsymbol{\beta}_j \neq 0,\\
&\nu_j \in [-1, 1], \text{ if } \boldsymbol{\beta}_j = 0.
\end{aligned}$$

Hence, there exists w and u such that:

$$\mathbf{X}_M^\top(\mathbf{X}_M w - \mathbf{y}) + \lambda s + \mathbf{C}_M\eta = 0, \tag{A.12}$$

$$\mathbf{X}_{-M}^\top(\mathbf{X}_M w - \mathbf{y}) + \lambda u + \mathbf{C}_{-M}\eta = 0, \tag{A.13}$$

$$\mathbf{C}_M^\top w = 0, \tag{A.14}$$

$$\mathsf{sign}(w) = s, \tag{A.15}$$

$$\|u\|_\infty < 1. \tag{A.16}$$

From (A.12) we solved for w as:

$$w = (\mathbf{X}_M^\top\mathbf{X}_M)^{-1}(\mathbf{X}_M^\top\mathbf{y} - \lambda s - \mathbf{C}_M\eta). \tag{A.17}$$

Plugging (A.17) into (A.13) we could solve for u:

$$\begin{aligned}
u &= \frac{1}{\lambda}\left(-\mathbf{X}_{-M}^\top\mathbf{X}_M w + \mathbf{X}_{-M}^\top\mathbf{y} - \mathbf{C}_{-M}\eta\right),\\
&= \frac{1}{\lambda}\big(-\mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}(\mathbf{X}_M^\top\mathbf{y} - \lambda s - \mathbf{C}_M\eta) + \mathbf{X}_{-M}^\top\mathbf{y} - \mathbf{C}_{-M}\eta\big),\\
&= \frac{1}{\lambda}\mathbf{X}_{-M}^\top\left(I - \mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\right)\mathbf{y} + \mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\left(s + \frac{1}{\lambda}\mathbf{C}_M\eta\right) - \frac{1}{\lambda}\mathbf{C}_{-M}\eta.
\end{aligned} \tag{A.18}$$

Also plugging (A.17) into (A.14) we could solve for η:

$$\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}(\mathbf{X}_M^\top\mathbf{y} - \lambda s - \mathbf{C}_M\eta) = 0,$$

which implies

$$\eta = \left[\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right]^{-1}\left[\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\mathbf{y} - \lambda\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}s\right]. \tag{A.19}$$

So the KKT conditions reduced to (A.15) and (A.16). For condition (A.15),

$$\begin{aligned}\{\mathsf{sign}(w) = s\} &= \{\mathsf{diag}(s)w > 0\} \\ &= \{\mathsf{diag}(s)(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}(\mathbf{X}_M^\top\mathbf{y} - \lambda s - \mathbf{C}_M\eta) > 0\}.\end{aligned}$$

Replacing η with (A.19):

$$\begin{aligned}&\{\mathsf{diag}(s)(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}(\mathbf{X}_M^\top\mathbf{y} - \lambda s - \mathbf{C}_M\eta) > 0\} \\ &= \{\mathsf{diag}(s)(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\Big(\mathbf{X}_M^\top\mathbf{y} - \lambda s - \mathbf{C}_M\left[\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right]^{-1}\left[\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\mathbf{y}\right. \\ &\left.-\lambda\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}s\right]\Big) > 0\}\end{aligned}$$

Reorganizing the above inequality gives:

$$\{\mathsf{diag}(s)w > 0\} = \{A_1\mathbf{y} < b_1\}, \tag{A.20}$$

where A_1 and b_1 is given as following:

$$A_1 = -\mathsf{diag}(s)(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\left(\mathbf{X}_M^\top - \mathbf{C}_M\left[\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right]^{-1}\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\right) \tag{A.21}$$

$$b_1 = -\mathsf{diag}(s)(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\left(\lambda s - \mathbf{C}_M\left[\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right]^{-1}\left[\lambda\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}s\right]\right) \tag{A.22}$$

For condition (A.16),

$$\{\|u\|_\infty < 1\} = \{-\mathbf{1} < u < \mathbf{1}\}$$

Replacing u with (A.18) and reorganizing the inequality gives:

$$\{-\mathbf{1} < u < \mathbf{1}\} = \{A_0\mathbf{y} < b_0\},$$

where

$$A_0 = \begin{pmatrix} A_{01} \\ A_{02} \end{pmatrix}, b_0 = \begin{pmatrix} b_{01} \\ b_{02} \end{pmatrix}$$

and A_{01}, A_{02}, b_{01} and b_{02} are given as following:

$$\begin{aligned}
A_{01} = &-\frac{1}{\lambda}\mathbf{X}_{-M}^\top \left(I - \mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\right) - \left[\mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\cdot\frac{1}{\lambda}\mathbf{C}_M - \frac{1}{\lambda}\mathbf{C}_{-M}\right]\cdot \\
&\left[\left(\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right)^{-1}\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\right], \\
A_{02} = &\frac{1}{\lambda}\mathbf{X}_{-M}^\top \left(I - \mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\right) + \left[\mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\cdot\frac{1}{\lambda}\mathbf{C}_M - \frac{1}{\lambda}\mathbf{C}_{-M}\right]\cdot \\
&\left[\left(\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right)^{-1}\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{X}_M^\top\right],
\end{aligned}$$

and

$$\begin{aligned}
b_{01} = &\mathbf{1} + \mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}s - \left[\mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\cdot\frac{1}{\lambda}\mathbf{C}_M - \frac{1}{\lambda}\mathbf{C}_{-M}\right]\cdot \\
&\left(\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right)^{-1}\left[\lambda\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}s\right], \\
b_{02} = &\mathbf{1} - \mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}s + \left[\mathbf{X}_{-M}^\top\mathbf{X}_M(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\cdot\frac{1}{\lambda}\mathbf{C}_M - \frac{1}{\lambda}\mathbf{C}_{-M}\right]\cdot \\
&\left(\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}\mathbf{C}_M\right)^{-1}\left[\lambda\mathbf{C}_M^\top(\mathbf{X}_M^\top\mathbf{X}_M)^{-1}s\right].
\end{aligned}$$

So we have finished the proof.

A.3.2. Proof for Theorem 3

We know that:

$$\{A\mathbf{y} < b\} = \{\nu^-(\mathbf{z}) \le \boldsymbol{\xi}^\top\mathbf{y} \le \nu^+(\mathbf{z}), \nu^0(\mathbf{z}) \ge 0\},$$

where ν^- and ν^+ are defined in (3.7) and (3.8). Since $\mathbf{z}$ is independent of $\boldsymbol{\xi}^\top \mathbf{y}$, we know that $\nu^-(\mathbf{z}), \nu^+(\mathbf{z})$ and $\nu^0(\mathbf{z})$ are all independent of $\boldsymbol{\xi}^\top \mathbf{y}$. So for any $\boldsymbol{\xi} \in \mathbb{R}^n$, we have:

$$\begin{aligned}\left[\boldsymbol{\xi}^\top \mathbf{y} \mid A\mathbf{y} \leq b, \mathbf{z}\right] &= \left[\boldsymbol{\xi}^\top \mathbf{y} \mid \nu^-(\mathbf{z}) \leq \boldsymbol{\xi}^\top \mathbf{y} \leq \nu^+(\mathbf{z}), \nu^0(\mathbf{z}) \geq 0\right], \\ &= \left[\boldsymbol{\xi}^\top \mathbf{y} \mid \nu^-(\mathbf{z}) \leq \boldsymbol{\xi}^\top \mathbf{y} \leq \nu^+(\mathbf{z})\right], \\ &\sim \mathrm{TN}(\boldsymbol{\xi}^\top \mu, \sigma^2 \|\boldsymbol{\xi}\|^2, \nu^-(\mathbf{z}), \nu^+(\mathbf{z})). \end{aligned} \tag{A.23}$$

So applying probability integral transformation theorem (Casella and Berger, 2002) to (A.23), we know that for $F = F^{\nu^-(\mathbf{z}),\nu^+(\mathbf{z})}_{\boldsymbol{\xi}^\top \mu, \sigma^2\|\boldsymbol{\xi}\|^2}$ defined in Theorem 3,

$$\left[F(\boldsymbol{\xi}^\top \mathbf{y}) \mid A\mathbf{y} \leq b, \mathbf{z}\right] \sim \mathsf{unif}(0,1).$$

Further integrating over $\mathbf{z}$, we know that

$$\left[F(\boldsymbol{\xi}^\top \mathbf{y}) \mid A\mathbf{y} \leq b\right] \sim \mathsf{unif}(0,1). \tag{A.24}$$

As our target parameter is given in (3.5), let

$$\boldsymbol{\xi} = e_j^\top \left[(\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{X}_M^\top - (\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{C}_M[\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{C}_M]^{-1}\mathbf{C}_M^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1}\mathbf{X}_M^\top\right], \tag{A.25}$$

from (A.24) we have:

$$\left[F^{\nu^-(\mathbf{z}),\nu^+(\mathbf{z})}_{\boldsymbol{\beta}^M_{oracle,j}, \sigma^2\|\boldsymbol{\xi}\|^2}(\boldsymbol{\xi}^\top \mathbf{y}) \mid \widehat{M} = M, \widehat{s} = s\right] \sim \mathsf{unif}(0,1). \tag{A.26}$$

This indicates that $F^{\nu^-(\mathbf{z}),\nu^+(\mathbf{z})}_{\boldsymbol{\beta}^M_{oracle,j}, \sigma^2\|\boldsymbol{\xi}\|^2}(\boldsymbol{\xi}^\top \mathbf{y})$ is actually a pivot. So by inverting the pivot, based on the fact that the truncated Gaussian distribution has monotone likelihood ratio in the mean parameter (see proofs in Lee et al. (2016)), we know

$$P(\boldsymbol{\beta}^M_{oracle,j} \in [L, U] \mid \widehat{M} = M, \widehat{s} = s) = 1 - \alpha,$$

with L and U defined in (3.9). Furthermore, notice that:

$$P(\boldsymbol{\beta}^M_{oracle,j} \in [L,U] \mid \widehat{M} = M) = \sum_s P(\boldsymbol{\beta}^M_{oracle,j} \in [L,U] \mid \widehat{M} = M, \widehat{s} = s) P(\widehat{M} = M \mid \widehat{s} = s)$$
$$\geq \sum_s (1-\alpha) P(\widehat{M} = M \mid \widehat{s} = s) = 1 - \alpha.$$

So we have finished the proof.

A.4. Proofs for Chapter 4

In the section we provided the proofs for the lemmas and theorems in Chapter 4. We refer the proof of Lemma 1 to Lin, Feng, and Li (2015). Before proving lemma 4, we first state a useful proposition.

Proposition 2. *Denote $\widehat{\mathbf{D}}_i$ as defined previously, $\widehat{\mathbf{M}}_i$ as $\left(Y, \widehat{\mathbf{D}}_{\cdot,-i}\right)$ for $i = 1, 2 \ldots p$. Further for each $l = 1, 2, \ldots p$, we use $\mathbf{M}_{i,l}$ and $\widehat{\mathbf{M}}_{i,l}$ to be the l-th column of the matrix $\mathbf{M}_i$ and $\widehat{\mathbf{M}}_i$ respectively (Notice that $\mathbf{M}_i$ is a matrix so $\mathbf{M}_{i,l}$ is a column vector, not the $(i,l)-$th element of matrix $\mathbf{M}$). Then under the assumptions stated in lemma 3, with the same choice of the tuning parameters λ_{2i}, with probability at least $1 - (pq)^{1-C^2/8}$ for some $C \geq 2\sqrt{2}$, we have:*

$$\|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2 \leq \frac{4\sqrt{n}\sqrt{s_2}\lambda_{2i}}{\kappa^2(s_2, \mathbf{Z})}, \quad i = 1, 2, \ldots, p,$$
$$\|\widehat{\mathbf{M}}_{i,l} - \mathbf{M}_{i,l}\|_2 \leq \frac{4\sqrt{n}\sqrt{s_2}\lambda_{2i}}{\kappa^2(s_2, \mathbf{Z})}, \quad i = 1, 2, \ldots, p,$$

In addition, the estimated $\widehat{\mathbf{M}}_i$ satisfies the RE condition with some constant $\kappa(r_i, \widehat{\mathbf{M}}_i)$ which satisfies $\kappa(r_i, \widehat{\mathbf{M}}_i) \geq \frac{1}{2}\kappa(r_i, \mathbf{M}_i)$.

Proof of proposition 2. Notice that for $i = 1, 2, \ldots, p$,

$$\|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2 = \left\|\mathbf{Z}\left(\widehat{\mathbf{\Gamma}}_{0,i} - \mathbf{\Gamma}_{0,i}\right)\right\|_2 \leq \frac{4\sqrt{n}\sqrt{s_2}\lambda_{2i}}{\kappa(s_2, \mathbf{Z})},$$

where the last inequality follows from Bickel, Ritov, and Tsybakov (2009).

For the second inequality, notice that $\widehat{\mathbf{M}}_i = \left(Y, \widehat{\mathbf{D}}_{\cdot,-i}\right)$, so there exists some i_0 such that:

$$\|\widehat{\mathbf{M}}_{i,l} - \mathbf{M}_{i,l}\|_2 = \begin{cases} 0 \quad \text{if } l = 1, \\ \left\|\mathbf{Z}\left(\widehat{\mathbf{\Gamma}}_{0,i_0} - \mathbf{\Gamma}_{0,i_0}\right)\right\|_2. \end{cases}$$

So similarly we have:

$$\|\widehat{\mathbf{M}}_{i,l} - \mathbf{M}_{i,l}\|_2 \leq \frac{4\sqrt{n}\sqrt{s_2}\lambda_{2i}}{\kappa(s_2, \mathbf{Z})}.$$

Furthermore, according to Lin, Feng, and Li (2015), they proved that $Z\widehat{\mathbf{\Gamma}}$ satisfies the RE condition with $\kappa(s_1, Z\widehat{\mathbf{\Gamma}}) \geq \frac{1}{2}\kappa(s_1, \mathbf{D})$. Using the relationship between $\mathbf{M}_i$ and $\mathbf{D}$, it is straightforward that $\kappa(r_i, \widehat{\mathbf{M}}_i) \geq \frac{1}{2}\kappa(r_i, \mathbf{M}_i)$. □

Then we provided the proof of lemma 4.

Proof of lemma 4. Without lose of generality, we assume $a_i = 0$. For each $i = 1, 2, \ldots, p$, by the definition of $\widehat{\boldsymbol{\theta}}_i$ in (4.8), we have:

$$\frac{1}{2n}\|\widehat{\mathbf{D}}_i - \widehat{\mathbf{M}}_i\widehat{\boldsymbol{\theta}}_i\|_2^2 + \mu_i\|\widehat{\boldsymbol{\theta}}_i\|_1 \leq \frac{1}{2n}\|\widehat{\mathbf{D}}_i - \widehat{\mathbf{M}}_i\boldsymbol{\theta}_i\|_2^2 + \mu_i\|\boldsymbol{\theta}_i\|_1. \tag{A.27}$$

For the left hand side (LHS), notice that:

$$\begin{aligned} \frac{1}{2n}\|\widehat{\mathbf{D}}_i - \widehat{\mathbf{M}}_i\widehat{\boldsymbol{\theta}}_i\|_2^2 &= \frac{1}{2n}\|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2^2 + \frac{1}{2n}\|\mathbf{D}_i - \widehat{\mathbf{M}}_i\widehat{\boldsymbol{\theta}}_i\|_2^2 - \frac{1}{n}\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right)^\top\left(\mathbf{D}_i - \widehat{\mathbf{M}}_i\widehat{\boldsymbol{\theta}}_i\right), \\ &= \frac{1}{2n}\|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2^2 + \frac{1}{2n}\|\boldsymbol{\zeta}_i\|_2^2 + \frac{1}{2n}\|\widehat{\mathbf{M}}_i(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\|_2^2 + \frac{1}{2n}\|(\widehat{\mathbf{M}}_i - \mathbf{M}_i)\boldsymbol{\theta}_i\|_2^2, \\ &\quad - \frac{1}{n}\boldsymbol{\zeta}^\top\left(\widehat{\mathbf{M}}_i\widehat{\boldsymbol{\theta}}_i - \mathbf{M}_i\boldsymbol{\theta}\right) + \frac{1}{n}\boldsymbol{\theta}_i\left(\widehat{\mathbf{M}}_i - \mathbf{M}_i\right)^\top\widehat{\mathbf{M}}_i\left(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right), \\ &\quad - \frac{1}{n}\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right)^\top\left(\mathbf{D}_i - \widehat{\mathbf{M}}_i\widehat{\boldsymbol{\theta}}_i\right). \end{aligned} \tag{A.28}$$

While for the right hand side(RHS), similarly,

$$\begin{aligned}\frac{1}{2n}\|\widehat{\mathbf{D}}_i - \widehat{\mathbf{M}}_i\boldsymbol{\theta}_i\|_2^2 &= \frac{1}{2n}\|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2^2 + \frac{1}{2n}\|\mathbf{D}_i - \widehat{\mathbf{M}}_i\boldsymbol{\theta}_i\|_2^2 - \frac{1}{n}\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right)^\top\left(\mathbf{D}_i - \widehat{\mathbf{M}}_i\boldsymbol{\theta}_i\right), \\ &= \frac{1}{2n}\|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2^2 + \frac{1}{2n}\|\boldsymbol{\zeta}_i\|_2^2 + \frac{1}{2n}\|(\widehat{\mathbf{M}}_i - \mathbf{M}_i)\boldsymbol{\theta}_i\|_2^2 - \frac{1}{n}\boldsymbol{\zeta}_i^\top\left(\widehat{\mathbf{M}}_i - \mathbf{M}_i\right)\boldsymbol{\theta}_i, \\ &- \frac{1}{n}\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right)^\top\left(\mathbf{D}_i - \widehat{\mathbf{M}}_i\widehat{\boldsymbol{\theta}}_i\right). \end{aligned} \tag{A.29}$$

Combining (A.28), (A.29) and (A.27) we have:

$$\begin{aligned}\frac{1}{2n}\|\widehat{\mathbf{M}}_i(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\|_2^2 &\leq \frac{1}{n}\boldsymbol{\zeta}_i^\top\widehat{\mathbf{M}}_i\left(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}\right) - \frac{1}{n}\boldsymbol{\theta}_i\left(\widehat{\mathbf{M}}_i - \mathbf{M}_i\right)^\top\widehat{\mathbf{M}}_i\left(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right), \\ &+ \frac{1}{n}\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right)^\top\widehat{\mathbf{M}}_i\left(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right) + \mu_i\left(\|\boldsymbol{\theta}_i\|_1 - \|\widehat{\boldsymbol{\theta}}_i\|_1\right), \\ &\leq \left\|\frac{1}{n}\widehat{\mathbf{M}}_i^\top\boldsymbol{\zeta}_i - \frac{1}{n}\widehat{\mathbf{M}}_i^\top\left(\widehat{\mathbf{M}}_i - \mathbf{M}_i\right)\boldsymbol{\theta}_i + \frac{1}{n}\widehat{\mathbf{M}}_i^\top\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right)\right\|_\infty\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1, \\ &+ \mu_i\left(\|\boldsymbol{\theta}_i\|_1 - \|\widehat{\boldsymbol{\theta}}_i\|_1\right).\end{aligned}$$

We first show that the event $\left\|\frac{1}{n}\widehat{\mathbf{M}}_i^\top\boldsymbol{\zeta}_i - \frac{1}{n}\widehat{\mathbf{M}}_i^\top\left(\widehat{\mathbf{M}}_i - \mathbf{M}_i\right)\boldsymbol{\theta}_i + \frac{1}{n}\widehat{\mathbf{M}}_i^\top\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right)\right\|_\infty \leq \frac{\mu_i}{2}$ happens with large probability.

As

$$\begin{aligned}&\frac{1}{n}\widehat{\mathbf{M}}_i^\top\boldsymbol{\zeta}_i - \frac{1}{n}\widehat{\mathbf{M}}_i^\top\left(\widehat{\mathbf{M}}_i - \mathbf{M}_i\right)\boldsymbol{\theta}_i + \frac{1}{n}\widehat{\mathbf{M}}_i^\top\left(\widehat{\mathbf{D}}_i - \mathbf{D}_i\right), \\ &= \underbrace{\frac{1}{n}\mathbf{M}_i^\top\boldsymbol{\zeta}_i}_{T_1} + \underbrace{\frac{1}{n}(\widehat{\mathbf{M}}_i - \mathbf{M}_i)^\top\boldsymbol{\zeta}_i}_{T_2} - \underbrace{\frac{1}{n}\mathbf{M}_i^\top(\widehat{\mathbf{M}}_i - \mathbf{M}_i)\boldsymbol{\theta}_i}_{T_3} - \underbrace{\frac{1}{n}(\widehat{\mathbf{M}}_i - \mathbf{M}_i)^\top(\widehat{\mathbf{M}}_i - \mathbf{M}_i)\boldsymbol{\theta}_i}_{T_4}, \\ &+ \underbrace{\frac{1}{n}\mathbf{M}_i^\top(\widehat{\mathbf{D}}_i - \mathbf{D}_i)}_{T_5} + \underbrace{\frac{1}{n}(\widehat{\mathbf{M}}_i - \mathbf{M}_i)^\top(\widehat{\mathbf{D}}_i - \mathbf{D}_i)}_{T_6},\end{aligned}$$

we label these terms from T_1 to T_6. To bound term T_1, it follows from the union bound and the Gaussian tail bound:

$$\mathbb{P}(\|T_1\|_\infty \geq \frac{\mu_i}{12}) = \mathbb{P}\left(\left\|\frac{1}{n}\mathbf{M}_i^\top\boldsymbol{\zeta}\right\|_\infty \geq \frac{\mu_i}{12}\right) \leq p\exp\left\{-\frac{n}{2\sigma_{\boldsymbol{\zeta}_i}^2}\cdot\left(\frac{\mu_i}{12}\right)^2\right\}. \tag{A.30}$$

To bound term T_2, noticing that $\|\widehat{\mathbf{\Gamma}}_{0,i} - \widehat{\mathbf{\Gamma}}_{0,i}\|_1 \leq \dfrac{16 s_2 \lambda_{2i}}{\kappa^2(s_2, \mathbf{Z})}$,

$$\begin{aligned}
\mathbb{P}(\|T_2\|_\infty \geq \frac{\mu_i}{12}) &\leq \mathbb{P}\left(\left\| \frac{1}{n} \mathbf{Z}^\top \boldsymbol{\zeta}_i \right\|_\infty \geq \frac{\mu_i}{12} \cdot \frac{\kappa^2(s_2, \mathbf{Z})}{16 s_2 \lambda_{2i}} \right), \\
&\leq q C^* \exp\left\{ -\frac{n}{2\sigma^2_{\boldsymbol{\zeta}_i}} \cdot \left(\frac{\mu_i}{12} \cdot \frac{\kappa^2(s_2, \mathbf{Z})}{16 s_2 \lambda_{\max}} \right)^2 \right\},
\end{aligned} \tag{A.31}$$

for some positive constant C^*. As for term T_3, as $\|\boldsymbol{\theta}_i\|_\infty \leq C$ and by proposition 2,

$$\begin{aligned}
\|T_3\|_\infty = \left\| \frac{1}{n} \mathbf{M}_i^\top (\widehat{\mathbf{M}}_i - \mathbf{M}_i) \boldsymbol{\theta}_i \right\|_\infty &\leq C \max_{1 \leq l,k \leq p} \left| \frac{1}{n} \mathbf{M}_{i,l}^\top (\widehat{\mathbf{M}}_{i,k} - \mathbf{M}_{i,k}) \right|, \\
&\leq C \max_{1 \leq k \leq p} \frac{1}{\sqrt{n}} \left\| \widehat{\mathbf{M}}_{i,k} - \mathbf{M}_{i,k} \right\|_2 \leq \frac{4C\sqrt{s_2}\lambda_{\max}}{\kappa(s_2, \mathbf{Z})}.
\end{aligned} \tag{A.32}$$

For T_4, using the result in proposition 2, we have:

$$\begin{aligned}
\|T_4\|_\infty &= \left\| \frac{1}{n} (\widehat{\mathbf{M}}_i - \mathbf{M}_i)^\top (\widehat{\mathbf{M}}_i - \mathbf{M}_i) \boldsymbol{\theta}_i \right\|_\infty, \\
&\leq C \max_{1 \leq l,k \leq p} \frac{1}{n} \left\| (\widehat{\mathbf{M}}_{i,l} - \mathbf{M}_{i,l}) \right\|_2 \cdot \left\| (\widehat{\mathbf{M}}_{i,k} - \mathbf{M}_{i,k}) \right\|_2, \\
&\leq \frac{16 C s_2 \lambda^2_{\max}}{\kappa^2(s_2, \mathbf{Z})}.
\end{aligned} \tag{A.33}$$

For T_5, similar to T_3, we have

$$\begin{aligned}
\|T_5\|_\infty = \frac{1}{n} \left\| \mathbf{M}_i^\top (\widehat{\mathbf{D}}_i - \mathbf{D}_i) \right\|_\infty &\leq \max_{1 \leq l \leq p} \frac{1}{n} \left| \mathbf{M}_{i,l}^\top (\widehat{\mathbf{D}}_i - \mathbf{D}_i) \right|, \\
&\leq \frac{1}{\sqrt{n}} \|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2 \leq \frac{4C\sqrt{s_2}\lambda_{\max}}{\kappa(s_2, \mathbf{Z})}.
\end{aligned} \tag{A.34}$$

Finally for T_6,

$$\begin{aligned}
\|T_6\|_\infty = \frac{1}{n} \left\| (\widehat{\mathbf{M}}_i - \mathbf{M}_i)^\top (\widehat{\mathbf{D}}_i - \mathbf{D}_i) \right\|_\infty &\leq \max_{1 \leq l \leq p} \frac{1}{n} \left| (\widehat{\mathbf{M}}_{i,l} - \mathbf{M}_{i,l})^\top (\widehat{\mathbf{D}}_i - \mathbf{D}_i) \right|, \\
\leq \max_{1 \leq l \leq p} \frac{1}{n} \|\widehat{\mathbf{M}}_{i,l} - \mathbf{M}_{i,l}\|_2 \cdot \|\widehat{\mathbf{D}}_i - \mathbf{D}_i\|_2 &\leq \frac{16 C s_2 \lambda^2_{\max}}{\kappa^2(s_2, \mathbf{Z})}.
\end{aligned} \tag{A.35}$$

Combining the results from (A.30) to (A.35), there exists some positive constant C_4, C_5, C_5^*, such

that with the tuning parameter μ_i chosen as:

$$\mu_i = \frac{C_4^*}{\kappa(s_2, \mathbf{Z})} \sqrt{\frac{s_2(\log p + \log q)}{n}},$$

with $C_4^* = C_5^* \max(C, \sigma_{\zeta_i})$, then with probability at least $1 - C_4 (pq)^{-C_5}$,

$$\left\| \frac{1}{n} \widehat{\mathbf{M}}_i^\top \boldsymbol{\zeta}_i - \frac{1}{n} \widehat{\mathbf{M}}_i^\top \left(\widehat{\mathbf{M}}_i - \mathbf{M}_i \right) \boldsymbol{\theta}_i + \frac{1}{n} \widehat{\mathbf{M}}_i^\top \left(\widehat{\mathbf{D}}_i - \mathbf{D}_i \right) \right\|_\infty \leq \frac{\mu_i}{2} \tag{A.36}$$

Then under (A.36), we have:

$$\frac{1}{2n} \|\widehat{\mathbf{M}}_i (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\|_2^2 \leq \frac{\mu_i}{2} \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 + \mu_i \left(\|\boldsymbol{\theta}_i\|_1 - \|\widehat{\boldsymbol{\theta}}_i\|_1 \right). \tag{A.37}$$

Let R_i be the support of the true parameter $\boldsymbol{\theta}_i$ and without any abuse of using notations, we use $\boldsymbol{\theta}_{i,R_i}$ and $\widehat{\boldsymbol{\theta}}_{i,R_i}$ to represent the subvector of $\boldsymbol{\theta}_i$ and $\widehat{\boldsymbol{\theta}}_i$ restricted on the set R_i. Also let $|R_i| = r_i$. Adding $\frac{\mu_i}{2} \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1$ to both sides of (A.37) yields:

$$\begin{aligned} \frac{1}{2n} \|\widehat{\mathbf{M}}_i (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\|_2^2 + \frac{\mu_i}{2} \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 &\leq \mu_i \left(\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 + \|\boldsymbol{\theta}_i\|_1 - \|\widehat{\boldsymbol{\theta}}_i\|_1 \right), \\ &= \mu_i \left(\|\boldsymbol{\theta}_{i,R_i}\|_1 - \|\widehat{\boldsymbol{\theta}}_{i,R_i}\|_1 + \|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_1 \right), \\ &\leq 2\mu_i \|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_1 \leq 2\mu_i \sqrt{r_i} \|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_2. \end{aligned} \tag{A.38}$$

The last two inequalities in (A.38) imply:

$$\frac{1}{2n} \|\widehat{\mathbf{M}}_i (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\|_2^2 \leq 2\mu_i \sqrt{r_i} \|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_2, \tag{A.39}$$

$$\frac{\mu_i}{2} \|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 \leq 2\mu_i \|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_1, \tag{A.40}$$

and (A.40) is equivalent to

$$\|\widehat{\boldsymbol{\theta}}_{i,R_i^c} - \boldsymbol{\theta}_{i,R_i^c}\|_1 \leq 3 \|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_1. \tag{A.41}$$

As stated in proposition 2, $\widehat{\mathbf{M}}_i$ satisfies the RE condition with some constant $\kappa(r_i, \widehat{\mathbf{M}}_i) \geq \frac{1}{2} \kappa(r_i, \mathbf{M}_i)$,

together with (A.41) we have:

$$\frac{1}{2n}\|\widehat{\mathbf{M}}_i(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\|_2^2 \geq \frac{1}{2}\kappa^2(r_i, \widehat{\mathbf{M}}_i)\|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_2^2 \geq \frac{1}{8}\kappa^2(r_i, \mathbf{M}_i)\|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_2^2.$$

Combining with (A.39),

$$\|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_2 \leq \frac{16\mu_i\sqrt{r_i}}{\kappa^2(r_i, \mathbf{M}_i)}, \tag{A.42}$$

Plugging in the tuning parameter μ_i gives the final result in lemma 4:

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_1 &\leq 4\|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_1 \leq 4\sqrt{r_i}\|\widehat{\boldsymbol{\theta}}_{i,R_i} - \boldsymbol{\theta}_{i,R_i}\|_2, \\
&\leq \frac{64C_4^*}{\kappa^2(r_i, \mathbf{M}_i)\kappa(s_2, \mathbf{Z})} r_i\sqrt{\frac{s_2(\log p + \log q)}{n}}, \\
&\leq \frac{64C_4^*}{\kappa^2(Y, \mathbf{D})\kappa(s_2, \mathbf{Z})} r_i\sqrt{\frac{s_2(\log p + \log q)}{n}}.
\end{aligned}$$

□

Based on the previous results, we provide the proof for the main theorems. First we prove the asymptotic distribution of the test statistics for a single hypothesis.

Proof of Theorem 4. The form of the test statistic T_i is a de-biased version of the sample correlation. To show it follows a standard normal distribution, we list the following notation. Denote:

$$\widetilde{\xi}_k = \xi_k - \overline{\xi},\ \widetilde{\zeta}_{k,i} = \zeta_{k,i} - \overline{\zeta_i},$$

where $\overline{\xi} = \sum_{k=1}^n \xi_k$ and $\overline{\zeta_i} = \sum_{k=1}^n \zeta_{k,i}$. Recall that by the previous definition, we have:

$$\begin{aligned}
\xi_k &= y_k - \mu - \mathbf{D}_k^\top \beta, \\
\zeta_{k,i} &= \mathbf{D}_{k,i} - a_i - (y_k, D_{k,-i}^\top)^\top \boldsymbol{\theta}_i, \\
\widehat{\xi}_k &= y_k - \overline{Y} - (\widehat{\mathbf{D}}_k - \overline{\widehat{\mathbf{D}}})^\top \widehat{\boldsymbol{\beta}}, \\
\widehat{\zeta}_{k,i} &= \widehat{\mathbf{D}}_{k,i} - \overline{\widehat{\mathbf{D}}_i} - \left(y_k - \overline{Y}, \left(\widehat{\mathbf{D}}_{k,-i} - \overline{\widehat{\mathbf{D}}}_{-i}\right)^\top\right) \widehat{\boldsymbol{\theta}}_i.
\end{aligned}$$

Based on these notations, we have the following decomposition:

$$\frac{1}{n}\sum_{k=1}^{n}\widehat{\xi}_k\widehat{\zeta}_{k,i} = \frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k\widetilde{\zeta}_{k,i} - \underbrace{\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k(\widetilde{\zeta}_{k,i}-\widehat{\zeta}_{k,i})}_{A_1} - \underbrace{\frac{1}{n}\sum_{k=1}^{n}\widetilde{\zeta}_{k,i}(\widetilde{\xi}_k-\widehat{\xi}_k)}_{A_2} + \underbrace{\frac{1}{n}\sum_{k=1}^{n}(\widetilde{\xi}_k-\widehat{\xi}_k)(\widetilde{\zeta}_{k,i}-\widehat{\zeta}_{k,i})}_{A_3}.$$

For simplicity, denote the second, third and forth term as A_1, A_2 and A_3. Then for A_1, we have:

$$\begin{aligned}
A_1 = &\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k^2(\widehat{\boldsymbol{\theta}}_{1,i}-\boldsymbol{\theta}_{1,i}) + \frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k(\mathbf{D}_k-\overline{\mathbf{D}})^\top\beta(\widehat{\boldsymbol{\theta}}_{1,i}-\boldsymbol{\theta}_{1,i}),\\
&+\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k\left\{(\mathbf{D}_{k,i}-\overline{\mathbf{D}_i})-(\widehat{\mathbf{D}}_{k,i}-\overline{\widehat{\mathbf{D}}_i})\right\} + \frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k(\mathbf{D}_{k,-i}-\overline{\mathbf{D}}_{-i})^\top(\widehat{\boldsymbol{\theta}}_{-1,i}-\boldsymbol{\theta}_{-1,i}),\\
&+\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k\left\{(\mathbf{D}_{k,-i}-\overline{\mathbf{D}_{-i}})-(\widehat{\mathbf{D}}_{k,-i}-\overline{\widehat{\mathbf{D}}_{-i}})\right\}^\top\widehat{\boldsymbol{\theta}}_{-1,i}.
\end{aligned}$$

We denote these five terms as $A_{1.1}$ to $A_{1.5}$. For $A_{1.2}$, combining the result in lemma 4 and the fact that ξ and $\mathbf{D}$ are independent, we know that for some positive constant C, there exists some $C'>0$ such that:

$$\begin{aligned}
&|\widehat{\boldsymbol{\theta}}_{1,i}-\boldsymbol{\theta}_{1,i}| \lesssim_p r\sqrt{\frac{s_2(\log p+\log q)}{n}},\\
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k(\mathbf{D}_k-\overline{\mathbf{D}})^\top\boldsymbol{\beta}\right| \geq C\sqrt{\frac{\log p}{n}}\right) = \mathcal{O}(p^{-C'}).
\end{aligned}$$

Hence,

$$\begin{aligned}
A_{1,2} &\leq \left|\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k(\mathbf{D}_k-\overline{\mathbf{D}})^\top\boldsymbol{\beta}\right|\cdot|\widehat{\boldsymbol{\theta}}_{1,i}-\boldsymbol{\theta}_{1,i}|,\\
&\lesssim \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}}\cdot r\sqrt{\frac{s_2(\log p+\log q)}{n}}\right).
\end{aligned} \tag{A.43}$$

Similarly for $A_{1.4}$, as

$$\mathbb{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{k=1}^{n}\widetilde{\xi}_k(\mathbf{D}_{k,j}-\overline{\mathbf{D}}_j)\right| \leq C\sqrt{\frac{\log p}{n}}\right) = \mathcal{O}(p^{-C'}),$$

so we have:

$$A_{1.4} \leq \left\| \frac{1}{n} \sum_{k=1}^{n} \widetilde{\xi}_k (\mathbf{D}_{k,-i} - \overline{\mathbf{D}}_{-i}) \right\|_{\infty} \cdot \|\widehat{\boldsymbol{\theta}}_{-1,i} - \boldsymbol{\theta}_{-1,i}\|_1,$$
$$\lesssim \mathcal{O}_p \left(\sqrt{\frac{\log p}{n}} \cdot r \sqrt{\frac{s_2(\log p + \log q)}{n}} \right). \tag{A.44}$$

Then for $A_{1.3}$, by the estimation error for $\mathbf{D}_i$ as we used in proposition 2, we have:

$$\left| \frac{1}{n} \sum_{k=1}^{n} \widetilde{\xi}_k (\mathbf{D}_{k,i} - \widehat{\mathbf{D}}_{k,i}) \right| = \left| \frac{1}{n} \sum_{k=1}^{n} \widetilde{\xi}_k \left(Z_k (\widehat{\boldsymbol{\Gamma}}_{0,i} - \boldsymbol{\Gamma}_{0,i}) \right) \right|,$$
$$\leq \left\| \frac{1}{n} \sum_{k=1}^{n} \widetilde{\xi}_k Z_k \right\|_{\infty} \cdot \|\widehat{\boldsymbol{\Gamma}}_{0,i} - \boldsymbol{\Gamma}_{0,i}\|_1,$$
$$\lesssim \mathcal{O}_p \left(\sqrt{\frac{\log p}{n}} \cdot s_2 \sqrt{\frac{\log p + \log q}{n}} \right). \tag{A.45}$$

For the last term $A_{1.5}$, similar to $A_{1.3}$,

$$A_{1.5} \lesssim \mathcal{O}_p \left(\sqrt{\frac{\log p}{n}} \cdot s_2 \sqrt{\frac{\log p + \log q}{n}} \right). \tag{A.46}$$

Combining the result from (A.43) to (A.46) we know that uniformly for $1 \leq i \leq p$:

$$A_1 = \frac{1}{n} \sum_{k=1}^{n} \widetilde{\xi}_k^2 (\widehat{\boldsymbol{\theta}}_{1,i} - \boldsymbol{\theta}_{1,i}) + \mathcal{O}_p \left(\sqrt{\frac{\log p}{n}} \cdot r \sqrt{\frac{s_2(\log p + \log q)}{n}} \right)$$
$$+ \mathcal{O}_p \left(\sqrt{\frac{\log p}{n}} \cdot s_2 \sqrt{\frac{\log p + \log q}{n}} \right). \tag{A.47}$$

And as for term A_2, we have a similar decomposition given by:

$$A_2 = \frac{1}{n} \sum_{k=1}^{n} \widetilde{\zeta}_{k,i}^2 (\widehat{\beta}_i - \beta_i) + \frac{1}{n} \sum_{k=1}^{n} \widetilde{\zeta}_{k,i} \left(y_k - \overline{Y}, \left(\mathbf{D}_{k,-i} - \overline{\mathbf{D}}_{-i} \right)^{\top} \right) \boldsymbol{\theta}_i (\widehat{\beta}_i - \beta_i),$$
$$+ \frac{1}{n} \sum_{k=1}^{n} \widetilde{\zeta}_{k,i} \left(\mathbf{D}_{k,-i} - \overline{\mathbf{D}}_{-i} \right)^{\top} (\widehat{\boldsymbol{\beta}}_{-i} - \boldsymbol{\beta}_{-i}) + \frac{1}{n} \sum_{k=1}^{n} \widetilde{\zeta}_{k,i} \left[(\widehat{\mathbf{D}}_k - \mathbf{D}_k) + (\overline{\widehat{\mathbf{D}}} - \overline{\mathbf{D}}) \right] \widehat{\boldsymbol{\beta}}.$$

By using the similar techniques as in A_1, we know that uniformly over $1 \leq i \leq p$,

$$A_2 = \frac{1}{n}\sum_{k=1}^{n} \widetilde{\zeta}_{k,i}^2(\widehat{\beta}_i - \beta_i) + \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}} \cdot s_1\sqrt{\frac{s_2(\log p + \log q)}{n}}\right) + \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}} \cdot s_2\sqrt{\frac{\log p + \log q}{n}}\right). \tag{A.48}$$

For the last term A_3, the decomposition is given as following:

$$A_3 = \frac{1}{n}\sum_{k=1}^{n}\left\{\underbrace{(\mathbf{D}_k - \overline{\mathbf{D}})^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}_{B_1} + \underbrace{\left((\widehat{\mathbf{D}}_k - \mathbf{D}_k) - (\overline{\widehat{\mathbf{D}}} - \overline{\mathbf{D}})\right)^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}_{B_2}\right\},$$
$$\cdot\left\{\underbrace{\left((\mathbf{D}_{k,i} - \overline{\mathbf{D}_i}) - (\widehat{\mathbf{D}}_{k,i} - \overline{\widehat{\mathbf{D}}_i})\right)}_{B_3} + \underbrace{\left(y_k - \overline{Y}, \left(\mathbf{D}_{k,-i} - \overline{\mathbf{D}}_{-i}\right)^\top\right)(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)}_{B_4}\right.$$
$$\left. + \underbrace{\left((\widehat{\mathbf{D}}_{k,-i} - \overline{\widehat{\mathbf{D}}}_{-i}) - (\mathbf{D}_{k,-i} - \overline{\mathbf{D}}_{-i})\right)\widehat{\boldsymbol{\theta}}_i}_{B_5}\right\}.$$

For simplicity we denote the five terms above as B_1 to B_5. Then,

$$A_3 = B_1(B_3 + B_4 + B_5) + B_2(B_3 + B_4 + B_5).$$

For B_1B_3, based on the bounds in proposition 2, we have:

$$B_1B_3 \leq \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \cdot \left\|\frac{1}{n}\sum_{k=1}^{n}(\mathbf{D}_k - \overline{\mathbf{D}})^\top\left((\mathbf{D}_{k,i} - \overline{\mathbf{D}_i}) - (\widehat{\mathbf{D}}_{k,i} - \overline{\widehat{\mathbf{D}}_i})\right)\right\|_\infty,$$
$$\lesssim \mathcal{O}_p\left(s_1\sqrt{\frac{s_2(\log p + \log q)}{n}} \cdot \sqrt{\frac{s_2(\log p + \log q)}{n}}\right). \tag{A.49}$$

For B_1B_4, it follows from the proof in Liu and Luo (2014) that:

$$B_1B_4 \lesssim \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}} \cdot s_1\sqrt{\frac{s_2(\log p + \log q)}{n}} \cdot r\sqrt{\frac{s_2(\log p + \log q)}{n}}\right) + \mathcal{O}_p\left(r\sqrt{\frac{s_2(\log p + \log q)}{n}} \cdot a_n\right), + \mathcal{O}_p\left(\lambda_{\max}(\Sigma_{\mathbf{D}}) \cdot a_n^2\right), \tag{A.50}$$

where $\max(\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2, \|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}\|_2) = \mathcal{O}_p(a_n)$. For term B_2B_3, we have:

$$\begin{aligned}
B_2B_3 &\leq \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \cdot \left\| \frac{1}{n}\sum_{k=1}^{n} \left((\widehat{\mathbf{D}}_k - \mathbf{D}_k) - (\overline{\widehat{\mathbf{D}}} - \overline{\mathbf{D}}) \right) \left((\mathbf{D}_{k,i} - \overline{\mathbf{D}_i}) - (\widehat{\mathbf{D}}_{k,i} - \overline{\widehat{\mathbf{D}}_i}) \right) \right\|_\infty , \\
&\leq \|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \cdot \max_{1\leq j\leq p} \left| \frac{1}{n}\sum_{k=1}^{n} \left((\mathbf{D}_{k,j} - \overline{\mathbf{D}_j}) - (\widehat{\mathbf{D}}_{k,j} - \overline{\widehat{\mathbf{D}}_j}) \right) \cdot \left((\mathbf{D}_{k,i} - \overline{\mathbf{D}_i}) - (\widehat{\mathbf{D}}_{k,i} - \overline{\widehat{\mathbf{D}}_i}) \right) \right| , \\
&\lesssim \mathcal{O}_p \left(s_1 \sqrt{\frac{s_2(\log p + \log q)}{n}} \cdot \frac{s_2(\log p + \log q)}{n} \right) . \qquad \text{(A.51)}
\end{aligned}$$

Then for B_2B_4, it follows from the estimator bounds of $\widehat{\boldsymbol{\theta}}$,

$$B_2B_4 \leq \|\widehat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i\|_1 \cdot \left\| \frac{1}{n}\sum_{k=1}^{n} (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top \left((\widehat{\mathbf{D}}_k - \mathbf{D}_k) - (\overline{\widehat{\mathbf{D}}} - \overline{\mathbf{D}}) \right) \left(y_k - \overline{Y}, \left(\mathbf{D}_{k,-i} - \overline{\mathbf{D}}_{-i} \right)^\top \right) \right\| . \qquad \text{(A.52)}$$

Notice that the second term is in the same order as the term B_1B_3, so the order of the whole term B_2B_4 is actually dominated by the term B_1B_3. And terms B_1B_5 and B_2B_5 are in the same order as B_1B_3 and B_2B_3. So together with the result in (A.49) to (A.52) and summing up the previous results in (A.47), (A.48) and the form of test statistic T_i, we have:

$$\begin{aligned}
T_i &= \sqrt{n} \left(\frac{1}{n}\sum_{k=1}^{n} \widehat{\xi}_k \widehat{\zeta}_{k,i} + \frac{1}{n}\sum_{k=1}^{n} \widehat{\xi}_k^2 \widehat{\boldsymbol{\theta}}_{1,i} + \frac{1}{n}\sum_{k=1}^{n} \widehat{\zeta}_{k,i}^2 \widehat{\boldsymbol{\beta}}_i \right) / \widehat{\sigma}_\xi \widehat{\sigma}_{\zeta_i} , \\
&= \frac{\sqrt{n}}{\widehat{\sigma}_\xi \widehat{\sigma}_{\zeta_i}} \left\{ \frac{1}{n}\sum_{k=1}^{n} \widetilde{\xi}_k \widetilde{\zeta}_{k,i} - A_1 - A_2 + A_3 + \frac{1}{n}\sum_{k=1}^{n} \widehat{\xi}_k^2 \widehat{\boldsymbol{\theta}}_{1,i} + \frac{1}{n}\sum_{k=1}^{n} \widehat{\zeta}_{k,i}^2 \widehat{\boldsymbol{\beta}}_i \right\} , \\
&= \frac{\sqrt{n}}{\widehat{\sigma}_\xi \widehat{\sigma}_{\zeta_i}} \left\{ \frac{1}{n}\sum_{k=1}^{n} \widetilde{\xi}_k \widetilde{\zeta}_{k,i} + \frac{1}{n}\sum_{k=1}^{n} \widetilde{\xi}_k^2 \boldsymbol{\theta}_{1,i} + \frac{1}{n}\sum_{k=1}^{n} \widetilde{\zeta}_{k,i}^2 \boldsymbol{\beta}_i + \frac{1}{n}\sum_{k=1}^{n} \widehat{\boldsymbol{\theta}}_{1,i} (\widehat{\xi}_k^2 - \widetilde{\xi}_k^2) \right. \\
&\left. + \frac{1}{n}\sum_{k=1}^{n} \widehat{\boldsymbol{\beta}}_i (\widehat{\zeta}_{k,i}^2 - \widetilde{\zeta}_{k,i}^2) + order \right\} , \qquad \text{(A.53)}
\end{aligned}$$

where $order$ is the sum of all the reminder terms, which is given by:

$$
\begin{aligned}
order = & \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}} \cdot r\sqrt{\frac{s_2(\log p+\log q)}{n}}\right) + \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}} \cdot s_2\sqrt{\frac{\log p+\log q}{n}}\right) \\
& + \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}} \cdot s_1\sqrt{\frac{s_2(\log p+\log q)}{n}}\right) + \mathcal{O}_p\left(s_1\sqrt{\frac{s_2(\log p+\log q)}{n}} \cdot \sqrt{\frac{s_2(\log p+\log q)}{n}}\right) \\
& + \mathcal{O}_p\left(\sqrt{\frac{\log p}{n}} \cdot s_1\sqrt{\frac{s_2(\log p+\log q)}{n}} \cdot r\sqrt{\frac{s_2(\log p+\log q)}{n}}\right) + \mathcal{O}_p\left(r\sqrt{\frac{s_2(\log p+\log q)}{n}} \cdot a_n\right) \\
& + \mathcal{O}_p\left(\lambda_{\max}(\Sigma_D) \cdot a_n^2\right) + \mathcal{O}_p\left(s_1\sqrt{\frac{s_2(\log p+\log q)}{n}} \cdot \frac{s_2(\log p+\log q)}{n}\right). \qquad \text{(A.54)}
\end{aligned}
$$

Define $\widetilde{\omega}_{ii} = \mathbf{\Omega}^D_{i,i} + \dfrac{\boldsymbol{\beta}_i^2}{\sigma_\xi^2}$. Notice that:

$$
\begin{aligned}
& \frac{1}{n}\sum_{k=1}^n \widetilde{\xi}_k\widetilde{\zeta}_{k,i} + \frac{1}{n}\sum_{k=1}^n \widetilde{\xi}_k^2\boldsymbol{\theta}_{1,i} + \frac{1}{n}\sum_{k=1}^n \widetilde{\zeta}_{k,i}^2\boldsymbol{\beta}_i, \\
& = \left(\frac{1}{n}\sum_{k=1}^n \xi_k\zeta_{k,i} - \mathbb{E}\xi\zeta_i\right) + \left(\mathbb{E}\xi\mathbb{E}\zeta_i - \bar{\xi}\bar{\zeta}_i\right) - \frac{\boldsymbol{\beta}_i}{\widetilde{\omega}_{ii}}\left(1 - \frac{\tilde{\sigma}_\xi^2}{\sigma_\xi^2} - \frac{\tilde{\sigma}_{\zeta_i}^2}{\sigma_{\zeta_i}^2}\right), \\
& = \left(\frac{1}{n}\sum_{k=1}^n \xi_k\zeta_{k,i} - \mathbb{E}\xi\zeta_i\right) + \mathcal{O}_p\left(\frac{\log p}{n}\right) - \frac{\boldsymbol{\beta}_i}{\widetilde{\omega}_{ii}}\left(1 - \frac{\tilde{\sigma}_\xi^2}{\sigma_\xi^2} - \frac{\tilde{\sigma}_{\zeta_i}^2}{\sigma_{\zeta_i}^2}\right). \qquad \text{(A.55)}
\end{aligned}
$$

Denote order $A_n = order + \mathcal{O}_p\left(\dfrac{\log p}{n}\right)$, followed by the argument in Liu and Luo (2014) we know that:

$$
\widehat{\sigma}_\xi^2 = \sigma_\xi^2 + \mathcal{O}_p(A_n + \sqrt{\frac{\log p}{n}}),\ \widehat{\sigma}_{\zeta_i}^2 = \sigma_{\zeta_i}^2 + \mathcal{O}_p(A_n + \sqrt{\frac{\log p}{n}}). \qquad \text{(A.56)}
$$

In addition, notice that the required assumption $\lambda_{\max}(\Sigma_{\mathbf{D}})a_n^2 = o(n^{-\frac{1}{2}})$ and $r\sqrt{s_2(\log p+\log q)}\cdot a_n = o(1)$ are naturally hold for the estimators we are using and under assumptions C1-2. Hence, based on assumptions C1-2, together with (A.53), (A.54), (A.55) and (A.56) we know that

$$
T_i \rightsquigarrow N(0,1).
$$

And further recalls the relation between T_i and $\widehat{T}_i$ given by:

$$
\widehat{T}_i = \frac{T_i}{1 - \dfrac{T_i^2}{n}\mathbf{1}\left(\dfrac{T_i^2}{n} < 1\right)}.
$$

So finally Slutsky's theorem, we know that

$$T_i \rightsquigarrow N(0,1).$$

So we have finished the proof of theorem 4. □

Once we were able to prove that our test statistic follows a standard normal distribution,the proofs for theorem 5 and 6 become rather straight forward. We refer the details of the proofs to section 5.2 of Liu and Luo (2014). The proofs for ours differ from theirs in the error terms which have already been shown to be controlled in preferred orders in the proofs of theorem 4.

A.5. Additional Simulation Studies Chapter 4

A.5.1. Evaluation of testing single hypothesis

Figures A.1 and A.2 show additional simulation results for testing single hypothesis.

A.5.2. Sensitivity analysis

We further examine the performance of our method with the existence of direct effects between potential instruments and outcome of interests. From a theoretical point of view, such effects would distort all the statistical inference procedures but it is still worthwhile to check the performance via simulations.

The data is generated in a way similar to the previous simulations but now the potential instruments $\mathbf{Z}$ has direct effects on the response Y such that $Y_i = \mathbf{X}_i\boldsymbol{\beta}_0 + \mathbf{Z}_i\boldsymbol{\tau} + \varepsilon_i$. In the setting of weak direct effects, only 2 true instruments are related with Y directly with a coefficient of $(0.5, -0.5)$. In the setting of relatively strong direct effects, 5 true instruments are related with Y directly with a coefficient of $(1, 1, 0.5, 0.5, -0.5)$. We perform our analysis ignoring the fact that these are actually invalid instruments and evaluate the results using eFDR and eFDV.

Table A.2 shows the empirical FDR and FDV for the proposed procedure with the presence of direct effects (invalid instruments). We could see that when the direct effects between the instruments and outcome is weak, our method could still provide valid inference procedure although the empirical FDR and FDV are slightly inflated. However, when having strong direct effects, the performance of

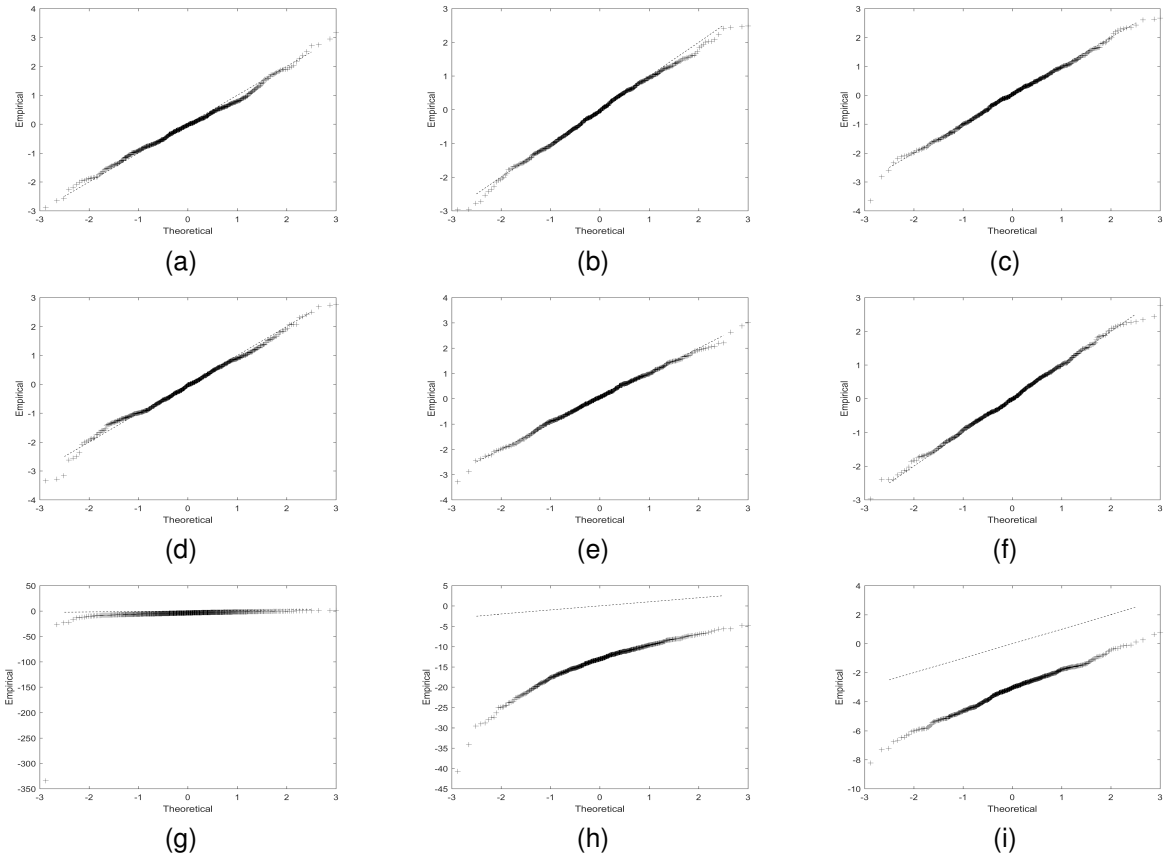


Figure A.1: QQ-plots of the test statistic $\widehat{T}_i$ based on the two-stage IV model for several randomly selected variables to demonstrate the validity of its asymptotic distribution. The panels in the first and second row correspond to selected variables whose true value are zero and the third row are variables that are not zero. For different columns, (a)(d)(g), (b)(e)(h) and (c)(f)(i) correspond to different (n, p, q) values as $(200, 100, 100)$, $(400, 200, 200)$ and $(200, 500, 500)$.

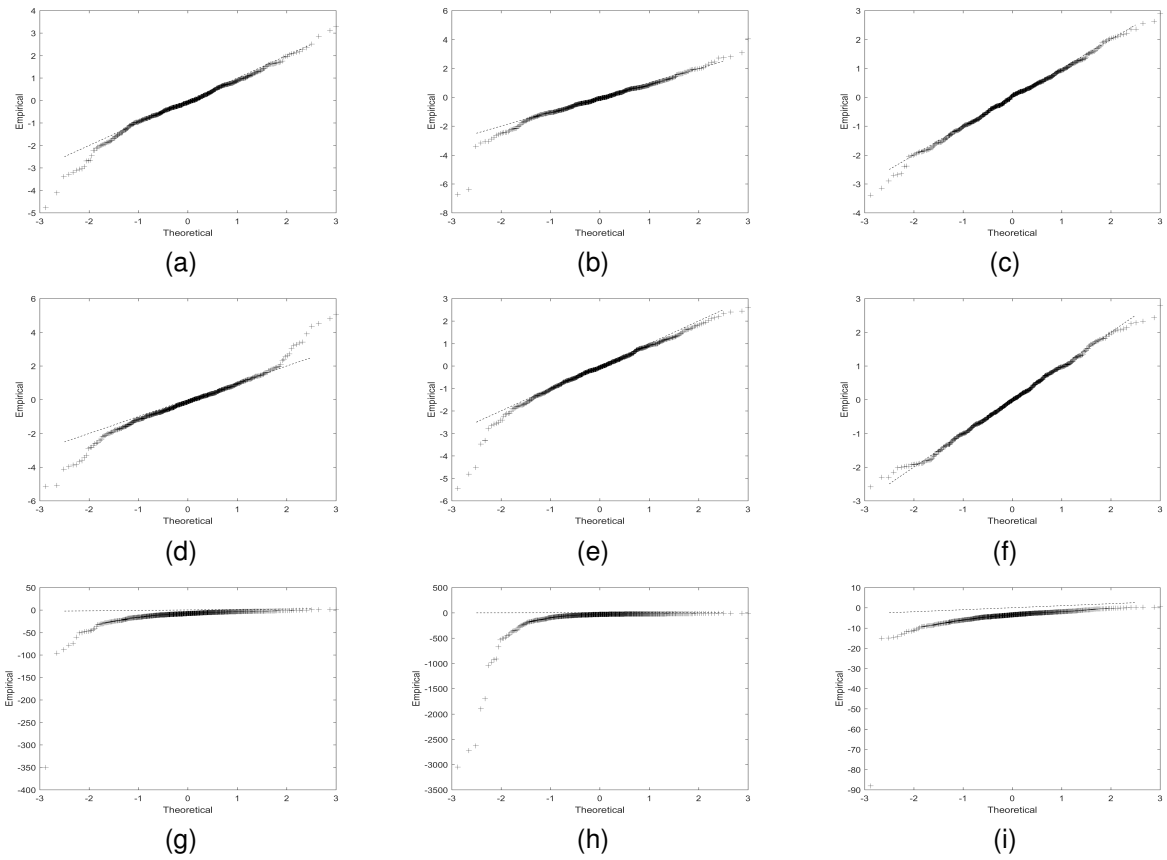


Figure A.2: Selected QQ-plots of the test statistics $\widehat{T}_i$ developed for fitting naive high dimensional regression models. The panels in the first and second row corresponds to selected variables whose true value are zero and the third row are variables that are not zero. For different columns, (a)(d)(g), (b)(e)(h) and (c)(f)(i) correspond to different (n, p, q) values as $(200, 100, 100)$, $(400, 200, 200)$ and $200, 500, 500)$.

our method is bad, reflected by the over-inflated values of empirical FDR and FDV. These results emphasize that necessity of using valid instruments when applying the IV regression methods since the strength of the direct effects is unknown in real applications.

Table A.2: Sensitivity analysis results based on 500 replications. The eFDR and eFDV for multiple testing procedures based on IV regression for different combinations of (n, p, q) and different α, k levels and weak and strong direct effects.

(n, p, q)	α-level	eFDR	k-level	eFDV
Weak direct effects				
$(n, p, q) = (200, 100, 100)$	0.05	0.13	2	2.44
	0.1	0.16	3	3.08
	0.2	0.23	4	3.82
$(n, p, q) = (400, 200, 200)$	0.05	0.09	2	2.23
	0.1	0.13	3	3.04
	0.2	0.21	4	3.76
Strong direct effects				
$(n, p, q) = (200, 100, 100)$	0.05	0.51	2	8.50
	0.1	0.54	3	9.51
	0.2	0.58	4	10.30
$(n, p, q) = (400, 200, 200)$	0.05	0.56	2	13.10
	0.1	0.60	3	14.49
	0.2	0.65	4	15.68

BIBLIOGRAPHY

Aitchison, J (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 139–177.

Aitchison, J and Bacon-shone, J (1984). Log contrast models for experiments with mixtures. *Biometrika* 71.2, 323–330.

Anderson, T (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471360919. URL: https://books.google.com/books?id=Cmm9QgAACAAJ.

Bickel, PJ, Ritov, Y, and Tsybakov, AB (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 1705–1732.

Brem, RB and Kruglyak, L (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences* 102.5, 1572–1577.

Bühlmann, P and Van De Geer, S (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, TT and Zhang, A (2013). Compressed sensing and affine rank minimization under restricted isometry. *Signal Processing, IEEE Transactions on* 61.13, 3279–3290.

Cao, Y, Lin, W, and Li, H (2018). Two-sample tests of high-dimensional means for compositional data. *Biometrika* 105, 115–132.

Cao, Y, Lin, W, and Li, H (2019). Large covariance estimation for compositional data via composition-adjusted thresholding. *Journal of the American Statistical Association* 114.526, 759–772.

Carabotti, M, Scirocco, A, Maselli, MA, and Severi, C (2015). The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology* 28.2, 203.

Casella, G and Berger, RL (2002). *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA.

Chatterjee, A and Lahiri, S (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proceedings of the American Mathematical Society* 138.12, 4497–4509.

Chen, B-J, Causton, HC, Mancenido, D, Goddard, NL, Perlstein, EO, and Pe'er, D (2009). Harnessing gene expression to identify the genetic basis of drug resistance. *Molecular systems biology* 5.1, 310.

Chen, J and Chen, Z (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* 95.3, 759–771.

Chen, J and Chen, Z (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*, 555–574.

Cho, I and Blaser, MJ (2012). The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* 13.4, 260–270.

Cryan, JF and Dinan, TG (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nature reviews neuroscience* 13.10, 701–712.

Dezeure, R, Bühlmann, P, and Zhang, C-H (2017). High-dimensional simultaneous inference with the bootstrap. *Test* 26.4, 685–719.

Fan, J and Liao, Y (2014). Endogeneity in high dimensions. *Annals of statistics* 42.3, 872.

Foster, JA and Neufeld, K-AM (2013). Gut–brain axis: how the microbiome influences anxiety and depression. *Trends in neurosciences* 36.5, 305–312.

Friedman, J, Hastie, T, and Tibshirani, R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33.1, 1.

Gamazon, E, Wheeler, H, Shah, K, Mozaffari, S, Aquino-Michaels, K, Carroll, R, Eyler, A, Denny, J, Consortium, G, Nicolae, D, Cox, N, and Im, H (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet.* 47, 1091–1098.

Geer, S Van de, Bühlmann, P, Ritov, Y, and Dezeure, R (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42.3, 1166–1202.

Gilbert, JA, Blaser, MJ, Caporaso, JG, Jansson, JK, Lynch, SV, and Knight, R (2018). Current understanding of the human microbiome. *Nature medicine* 24.4, 392.

Gold, D, Lederer, J, and Tao, J (2017). Inference for high-dimensional nested regression. *arXiv preprint arXiv:1708.05499*.

Goodrich, JK, Davenport, ER, Beaumont, M, Jackson, MA, Knight, R, Ober, C, Spector, TD, Bell, JT, Clark, AG, and Ley, RE (2016). Genetic determinants of the gut microbiome in UK twins. *Cell host & microbe* 19.5, 731–743.

Hoggart, CJ, Parra, EJ, Shriver, MD, Bonilla, C, Kittles, RA, Clayton, DG, and McKeigue, PM (2003). Control of confounding of genetic associations in stratified populations. *The American Journal of Human Genetics* 72.6, 1492–1504.

Imbens, G (2014). *Instrumental variables: An econometrician's perspective*. Tech. rep. National Bureau of Economic Research.

Ishiguro, A, Kassavetis, GA, and Geiduschek, EP (2002). Essential roles of Bdp1, a subunit of RNA polymerase III initiation factor TFIIIB, in transcription and tRNA processing. *Molecular and cellular biology* 22.10, 3264–3275.

Javanmard, A and Montanari, A (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* 15.1, 2869–2909.

Jie, Z, Xia, H, Zhong, S-L, Feng, Q, Li, S, Liang, S, Zhong, H, Liu, Z, Gao, Y, Zhao, H, et al. (2017). The gut microbiome in atherosclerotic cardiovascular disease. *Nature communications* 8.1, 1–12.

Kaakoush, NO, Day, AS, Huinao, KD, Leach, ST, Lemberg, DA, Dowd, SE, et al. (2012). Microbial dysbiosis in pediatric patients with Crohn's disease. *Journal of clinical microbiology* 50.10, 3258–3266.

Kabeerdoss, J, Sankaran, V, Pugazhendhi, S, and Ramakrishna, BS (2013). Clostridium leptum group bacteria abundance and diversity in the fecal microbiota of patients with inflammatory bowel disease: a case–control study in India. *BMC gastroenterology* 13.1, 20.

Kang, H, Zhang, A, Cai, TT, and Small, DS (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American statistical Association* 111.513, 132–144.

Kato, K, Odamaki, T, Mitsuyama, E, Sugahara, H, Xiao, J-Z, and Osawa, R (2017). Age-Related Changes in the Composition of Gut Bifidobacterium Species. *Current microbiology* 74.8.

Kurtz, ZD, Müller, CL, Miraldi, ER, Littman, DR, Blaser, MJ, and Bonneau, RA (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 11.5, e1004226.

Lee, JD, Sun, DL, Sun, Y, Taylor, JE, et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44.3, 907–927.

Leek, JT and Storey, JD (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* 3.9, e161.

Lewis, JD, Chen, EZ, Baldassano, RN, Otley, AR, Griffiths, AM, Lee, D, et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell host & microbe* 18.4, 489–500.

Lin, W, Feng, R, and Li, H (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association* 110.509, 270–288.

Lin, W, Shi, P, Feng, R, and Li, H (2014). Variable selection in regression with compositional covariates. *Biometrika* 101.4, 785–797.

Liu, W (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics* 41.6, 2948–2978.

Liu, W and Luo, S (2014). *Hypothesis testing for high-dimensional regression models*. Tech. rep. Technical report.

Liu, Y, Morley, M, Brandimarto, J, Hannenhalli, S, Hu, Y, Ashley, EA, Tang, WW, Moravec, CS, Margulies, KB, Cappola, TP, et al. (2015). RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics* 105.2, 83–89.

Lloyd-Price, J, Arze, C, Ananthakrishnan, AN, Schirmer, M, Avila-Pacheco, J, Poon, TW, Andrews, E, Ajami, NJ, Bonham, KS, Brislawn, CJ, et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569.7758, 655–662.

Lozupone, C, Faust, K, Raes, J, Faith, JJ, Frank, DN, Zaneveld, J, et al. (2012). Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts. *Genome research* 22.10, 1974–1984.

Lu, J and Li, H (2020a). *Hypothesis Testing in High-Dimensional Instrumental Variables Regression with an Application to Genomics Data*. Tech. rep. Technical report.

Lu, J and Li, H (2020b). *Post-selection Inference for Regression Models with Linear Constraints, with an Application to Microbiome Data*. Tech. rep. Technical report.

Lu, J, Shi, P, and Li, H (2019). Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* 75.1, 235–244.

Marykwas, D and Fox, T (1989). Control of the Saccharomyces cerevisiae regulatory gene PET494: transcriptional repression by glucose and translational induction by oxygen. *Molecular and cellular biology* 9.2, 484–491.

Matsuoka, K and Kanai, T (2015). "The gut microbiota and inflammatory bowel disease". In: *Seminars in immunopathology*. Vol. 37. 1. Springer, 47–55.

Meinshausen, N and Bühlmann, P (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, 417–473.

Mondot, S, Kang, S, Furet, J-P, Cárcer, D Aguirre de, McSweeney, C, Morrison, M, et al. (2011). Highlighting new phylogenetic specificities of Crohn's disease microbiota. *Inflammatory bowel diseases* 17.1, 185–192.

Nesterov, Y (2013). *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media.

Neykov, M, Ning, Y, Liu, JS, Liu, H, et al. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science* 33.3, 427–443.

Ning, Y and Liu, H (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* 45.1, 158–195.

Perlstein, EO, Ruderfer, DM, Roberts, DC, Schreiber, SL, and Kruglyak, L (2007). Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nature genetics* 39.4, 496.

Research Network Consortium, IH iHMP et al. (2019). The integrative human microbiome project. *Nature* 569, 641–648.

Rhodes, JM (2007). The role of Escherichia coli in inflammatory bowel disease. *Gut* 56.5, 610–612.

Romeo, S, Kozlitina, J, Xing, C, Pertsemlidis, A, Cox, D, Pennacchio, LA, Boerwinkle, E, Cohen, JC, and Hobbs, HH (2008). Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature genetics* 40.12, 1461–1465.

Schwabe, RF and Jobin, C (2013). The microbiome and cancer. *Nature Reviews Cancer* 13.11, 800–812.

Shi, P, Zhang, A, Li, H, et al. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics* 10.2, 1019–1040.

Sokol, H, Seksik, P, Furet, J, Firmesse, O, Nion-Larmurier, I, Beaugerie, L, et al. (2009). Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflammatory bowel diseases* 15.8, 1183–1189.

Speliotes, EK, Yerges-Armstrong, LM, Wu, J, Hernaez, R, Kim, LJ, Palmer, CD, Gudnason, V, Eiriksdottir, G, Garcia, ME, Launer, LJ, et al. (2011). Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS genetics* 7.3, e1001324.

Su, W, Boyd, S, and Candes, E (2014). "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights". In: *Advances in Neural Information Processing Systems*, 2510–2518.

Sun, T and Zhang, C-H (2012). Scaled sparse linear regression. *Biometrika* 99.4, 879–898.

Surette, MG (2014). The cystic fibrosis lung microbiome. *Annals of the American Thoracic Society* 11.Supplement 1, S61–S65.

Taylor, J and Tibshirani, R (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics* 46.1, 41–61.

Thaiss, CA, Zmora, N, Levy, M, and Elinav, E (2016). The microbiome and innate immunity. *Nature* 535.7610, 65–74.

Tibshirani, R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, 267–288.

Tibshirani, RJ et al. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* 7, 1456–1490.

Woodmansey, E (2007). Intestinal bacteria and ageing. *J Appl Microbiol.* 102, 1178–1186.

Xia, LC, Cram, JA, Chen, T, Fuhrman, JA, and Sun, F (2011). Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PloS one* 6.12.

Zhang, C-H and Zhang, SS (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, 217–242.