Statistical Methods for Sparse Binary, Count Data and Treatment Effect Heterogeneity

Yuxiang Xie

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Kwun Chuen Gary Chan, Chair

Peter Gilbert

Michael Wu

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Methods for Sparse Binary, Count Data
and Treatment Effect Heterogeneity

Yuxiang Xie

Chair of the Supervisory Committee:
Dr. Kwun Chuen Gary Chan
Department of Biostatistics

The concept of 'sparsity' is common to see in many topics of statistics. 'Sparsity' is a double-edged sword, depending on the statistical context. Sometimes, sparsity brings convenience; for example, a sparse statistical model is one having only a small number of nonzero parameters, which is easier to interpret than a dense model. On the other hand, sparsity may cause troubles; for example, a sparse sequencing read count table contains excessive zeros due to the issue that many rare bacterial taxa are not captured in the sequencing reads, and this sparsity may lead to inaccurate estimates of bacterial abundances.

This dissertation focuses on developing statistical methodologies for dealing with sparsity problems in three different statistical topics. We first present a false discovery rate (FDR) controlled variable selection method for a sparse model with binary covariates. We show that our proposal controls FDR under a pre-specified level in a finite sample and achieves asymptotic power equal to one under some mild assumptions. Next, we consider a sparse generalized linear model for studying treatment effect heterogeneity, and we propose two statistical frameworks that can detect factors contributing to heterogeneous treatment effect, and simultaneously control FDR. Finally, we develop a statistical method based on non-negative matrix factorization (NMF) for estimating bacterial compositions from sparse count data in microbiome studies. We establish upper bounds of estimation error for our

NMF estimators and show in simulation studies that our proposal outperforms some existing methods in various settings. We also demonstrate the interpretability of our model in a real data application.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Throughout my doctoral study, I have received a great deal of support and guidance. I would first like to thank my dissertation advisor, Dr. Kwun Chuen Gary Chan, whose expertise, insights, and encouragement are instrumental and crucial to my development as a statistician.

I would like to thank Dr. Peter Gilbert and Dr. Michael Wu for their invaluable advice and help throughout the writing of this dissertation. I would particularly like to thank Dr. Daniela Witten, for her training and support at the early stage of my doctoral study. I am also extremely grateful to my research assistant supervisors, Dr. Ying Qing Chen and Dr. Yoshio N. Hall, for all of the opportunities I have been given to conduct research with experienced collaborators.

I would like to acknowledge Dr. Raymond Wong, Dr. Xiaolin Shi and Nanyu Chen for their suggestions and comments on part of this dissertation work.

I would also like to single out my graduate program advisor, Gitana Garofalo, for her devotion to making the graduate program a warm family and for her kind-hearted help to me throughout the past five years.

I owe my thanks to other faculty and staff in the Department of Biostatistics and the Department of Statistics for making my doctoral study an unforgettable period in my lifetime.

Dedication

To my family.

Chapter 1

Introduction

This dissertation focuses on developing statistical methods for tackling three different problems that involve some sparsity assumptions. In this introductory chapter, we present the motivations for considering binary data, count data, and treatment effect heterogeneity data, and then finish it with an outlines of the dissertation.

1.1 Motivations

Variable selection has been widely used in data analysis for the past decades, and it becomes increasingly important in the Big Data era as there are usually hundreds of variables available in a dataset. With sparsity assumption, the true model should have only a small number of nonzero parameters. A sparse model is easier to estimate and interpret, and allows researchers to extract useful and reproducible patterns from data. Therefore, identifying potentially relevant features is often a step before fitting all the features into a regression model, especially when the number of features exceeds the number of observations. It is difficult for a variable selection method to avoid selecting false discoveries in a sparse model. A good variable selection method should effectively control the fraction of false discoveries while ensuring large enough power of its selection set.

In a lot of contemporary data applications, a great portion of features are coded as

binary variables. Binary features are widespread in many fields, from online controlled experiments to genome science to physical statistics. Although there has recently been a handful of literature for provable false discovery rate (FDR) control in variable selection, most of the theoretical analyses were based on some strong dependency assumption or Gaussian assumption among features. This motivates us to propose a FDR controlled variable selection method in regression framework for selecting binary features.

While studying FDR controlled variable selection method, we realize its natural extension to application in the detection of treatment effect heterogeneity. One way to study heterogeneous treatment effect is to formulate the problem as a generalized linear model. We usually assume that the treatment effect heterogeneity is caused by a small number of the factors in the model. Therefore, detecting factors contributing to treatment effect heterogeneity is equivalent to conducting variable selection in a sparse model. Treatment effect heterogeneity has become a hot topic in the study of online controlled experiments (a.k.a. A/B testing). Online controlled experiments have been used as the mantra for data-driven decision making on feature changing and product shipping in many Internet companies. The most commonly used A/B testing framework in many companies is based on Average Treatment Effect (ATE), which cannot detect the heterogeneity of treatment effect on users with different characteristics. Insights about user heterogeneity can help experimenters come up with strategies to improve the product, therefore, studying heterogeneity in treatment effect is meaningful and unavoidable for research in the Internet companies. In addition to online controlled experiments, the detection of heterogeneous treatment effect factors is also very important in developing personalized medicine and studying effect modifiers in vaccine efficacy trials.

The assumption of 'sparsity' plays a helpful role in the study of variable selection. However, the existence of 'sparsity' in count table of sequencing reads is troublesome. The human microbiome is the total DNA content of microbes inhabiting bodies. Studying the variability in microbiome has been a major topic in microbiome research as it may reveal the association between the microbiome and many complex diseases (Turnbaugh et al. [43], Lewis

et al. [25], Lloyd-Price et al. [28]). One important topic in microbiome studies is to estimate the bacterial composition matrix. Due to the issue that many rare bacterial taxa are not captured in the sequencing reads, the count tables of sequencing reads usually contains a lot of zeros. Simply using count normalization, thus, results in many zero proportions which are inaccurate estimates of bacterial abundances. In addition, literatures in co-occurrence pattern and symbiotic relationships in microbial communities (Faust et al. [18], Chaffron et al. [11]) indicate that compositional matrix is likely to possess low rank structure. The low rank structure is likely to enhance the interpretability of model for estimating compositional matrix. Given the aforementioned information, our goal is to propose an estimator of compositional matrix that overcomes the sparsity issue in count table and simultaneously has low rank structure.

1.2 Summary of the dissertation

In Chapter 2, we propose a false discovery rate (FDR) controlled variable selection method for a sparse model with binary covariates. Under mild conditions, we show that FDR is controlled exactly under a target level in a finite sample if the underlying distribution of the binary features is known. We show in simulations that FDR control is still attained when feature distribution is estimated from data. We also provide theoretical results on the power of our variables selection method in a linear regression model or a logistic regression model. In the restricted settings where competitors exist, we show in simulations and real data application on a HIV antiretroviral therapy dataset that our method has higher power than the competitor.

In Chapter 3, we present two statistical frameworks, HTE-TO and HTE-TC, that can detect factors contributing to heterogeneous treatment effect and at the same time control FDR of the detection set. Our methods are based on the Knockoff procedure and its variants (Barber and Candès [3], Candès et al. [9], Fan et al. [16], Xie and Chan [49]) combined with transformed outcome approach (Athey and Imbens [2]) and modified covariate method (Tian

et al. [40]). We provide a simple way to construct knockoffs for transformed covariates in HTE-TC methods, which allows us to apply existing Knockoff methods off-the-shelf without modifying their procedures. We show in simulations that both HTE-TO and HTE-TC have good power for detection and in the meantime control FDR under a target level. We apply our method on an HIV-1 vaccine efficacy trial data and its selection sets frequently include two variables that have previously been found to be a modifier of vaccine efficacy in literature.

In Chapter 4, we develop methods based on non-negative matrix factorization for estimating bacterial compositions from some sparse count data in microbiome studies. Due to the nature of non-negative matrix factorization, our estimator is constructed to possess low rank structure. We establish upper bounds of estimation error for our estimators and show in simulation studies that our proposal outperforms some existing methods in various settings. The application of our method on a breast milk dataset reveals clusters among the patients.

Chapter 2

Controlling the False Discovery Rate for Binary Feature Selection via Knockoff

2.1 Introduction

Generalized linear models, including linear regression model and logistic regression model, are widely used in statistical analysis of real data. In a regression framework, variable selection is one of the most popular tools for analyzing high dimensional data, in which a great number of features are available for modeling, while only a few of them are thought to be significantly associated with the response of interest. To enhance interpretability and predictability, it is crucial to identify the subset of relevant features before running a regression model. Many variable selection procedures with good theoretical properties have been proposed for the past two decades. For example, Tibshirani [41] proposed Lasso penalized linear regression model, which uses an l_1 penalty. Fan and Li [15] proposed SCAD, a non-convex penalty, for variable selection. Zou and Hastie [54] proposed regularization and variable selection via elastic net. An important question about variable selection is how many features should be

selected in the model. As a data-driven approach, the cross-validation method is commonly used for deciding the number of features selected (Shao [38], Zhang [53], Yu and Feng [52]). However, most cross-validation methods do not guarantee the control of false discovery rate (FDR) for the selected features.

Barber and Candès [3] proposes 'Knockoff' to conduct variable selection and control the false discovery rate simultaneously. The original Knockoff procedure, though elegant and salient, has a couple of limitations: it assumes that the underlying model is Gaussian linear with homoscedasiticity and does not work for high dimensional setting (i.e. more features than the sample size). Candès et al. [9] then extends the Knockoff idea to a model free procedure (Model-X Knockoff) which allows the underlying model to be any type and also allows for high dimensional set-up. Instead of assuming the relationship between the response variable and the features, Model-X Knockoff requires the knowledge of the distribution of features. Shifting the burden of knowledge from the true regression model to the distribution of features is reasonable, particularly in the case where features are from case-control studies. Fan et al. [16] shows that when the features are generated from a Gaussian graphical model, under some mild assumptions the Model-X Knockoff not only controls the false discovery rate, but also has asymptotic power equal one. Weinstein et al. [46] further conducts a power and prediction analysis for Knockoff using lasso statistics, and their analyses mainly focus on the cases where the distribution of features is continuous. Some interesting applications of Knockoffs can be found in Gao et al. [19], Xiao et al. [48] and Xie et al. [50].

Although there has been a handful of 'Knockoff' methods, most of the theoretical analyses focus only on the case where the distribution of features is continuous. Nonetheless, binary datasets are also widespread in many fields, from online controlled experiments to genome science to physical statistics. When the features in a model all take binary values, it is not reasonable to assume normality of their distribution. Sesia and Candès [37] has developed algorithm to sample Knockoff variables with the assumption that the features can be described by a hidden Markov model, but there is still a lack of methodology for extending the exact construction and theoretical analysis of Knockoff to binary features setting.

Ising graphical model is a standard model of a phase transition for ferromagnetism in statistical mechanics, and it is very popular in modeling the pairwise interactions between binary variables via Ising model. In addition, multivariate Bernoulli model (Dai et al. [12]) is an extension of Ising graphical model, which further allows modeling clique effects among the binary variables. Therefore, rather than Gaussian graphical model or other continuous graphical models, it is more natural to assume that the binary features are generated from an Ising graphical model or a multivariate Bernoulli model.

Our contributions. Since there is no tailored method of applying the Knockoff idea to binary features in existing literature, not to mention a thorough power analysis, in this chapter we close this gap by developing a Knockoff procedure for features following Ising distribution or multivariate Bernoulli distribution. In particular, we

1. develop an exact construction of Knockoffs for binary features that are generated from Ising or multivariate Bernoulli models,

2. provide theoretical analyses on the FDR control and asymptotic power of our Knockoff selection set,

3. propose a second-order approximation construction to speed up the Knockoff procedure,

4. confirm the practical utility of the proposed method by comparing it to existing Knockoff procedures in simulations and real data application.

2.2 Preliminaries

2.2.1 Model-X Knockoff

The Model-X Knockoff procedure is a FDR-control variable selection method in a framework with a response variable Y and multiple features $X = (X_1, \ldots, X_p)$.

Definition 1. *(Candès et al. [9])* $\tilde{X} = \left(\tilde{X}_1, \ldots, \tilde{X}_p\right)$ *are Model-X knockoffs for the original features* $X = (X_1, \ldots, X_p)$ *if*

- $\tilde{X} \perp\!\!\!\perp X | Y$,
- *and for any subset* $S \subset \{1, \ldots, p\}$,

$$\left(X, \tilde{X}\right)_{swap(S)} \stackrel{d}{=} \left(X, \tilde{X}\right), \tag{2.1}$$

where $\left(X, \tilde{X}\right)_{swap(S)}$ *means swapping the* X_j *and* $\tilde{X}_j$ *for all* $j \in S$.

Note that the Model-X Knockoff does not assume knowledge of the conditional distribution of $Y|X$ or the relationship between Y and X. Instead, it does assume the joint distribution of the features is known. The exchangeability condition (2.1) is the key of the Knockoff procedure and most of its variants, and the technical difficulty in constructing $\tilde{X}$ is to ensure this exchangeability condition (2.1) to hold. Candès et al. [9] provide an exact construction of $\tilde{X}$ in the case where the features are Gaussian distributed. They also propose a second-order approximation for constructing knockoffs in the case where the features are not Gaussian, however their theoretical result of FDR control does not hold exactly for the approximation construction.

After constructing knockoff $\tilde{X}$, under generalized linear model of Y given X, Candès et al. [9] propose to first solve a lasso type regression problem on the augmented design matrix $\mathbf{X}^* = \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$, and denote the solution by $\hat{\beta}(\lambda)$, where the tuning parameter λ is selected by cross-validation. Then set $Z_j = \left|\hat{\beta}_j(\lambda)\right|$ and $\tilde{Z}_j = \left|\hat{\beta}_{j+p}(\lambda)\right|$. The Lasso Coefficient Difference (LCD) statistic is defined to be

$$W_j = Z_j - \tilde{Z}_j = \left|\hat{\beta}_j(\lambda)\right| - \left|\hat{\beta}_{j+p}(\lambda)\right|. \tag{2.2}$$

Let $\hat{\mathcal{S}}$ be the variable selection set and $\mathcal{S}$ be the set of non-zero coefficients in the true model. The false discovery rate (FDR) is defined to be $\mathbb{E}\left[\mathrm{FDP}\right]$ where $\mathrm{FDP} = \frac{\left|\hat{\mathcal{S}} \cap \mathcal{S}^C\right|}{\left|\hat{\mathcal{S}}\right|}$. For a given $q \in (0, 1)$, choose a positive threshold T as

$$T = \min\left\{t > 0 : \frac{\#\left\{j : W_j \leq -t\right\} + 1}{\#\left\{j : W_j \geq t\right\}} \leq q\right\}, \tag{2.3}$$

and the Knockoff selected set $\hat{\mathcal{S}} = \{j : W_j \geq T\}$ controls the FDR at the level of q.

2.2.2 Ising Model and Multivariate Bernoulli Distribution

Consider an Ising graphical model with p nodes denoted by X_j, $1 \leq j \leq p$. We assume in the rest of this chapter that each X_j takes either $+1$ or 0, though our analysis is also applicable to X_j's taking $+1$ or -1. The joint distribution of X_j's takes the form

$$\begin{aligned} &P_\Theta(X_1 = x_1, \ldots, X_p = x_p) \\ &= \frac{1}{Z(\Theta)} \exp\{\sum_{j=1,\ldots,p} \Theta_{jj} x_j + \sum_{(j,j')\in\mathcal{E}} \Theta_{jj'} x_j x_{j'}\}, \end{aligned} \tag{2.4}$$

where $Z(\Theta)$ is a normalization term. Given an Ising parameter matrix $\Theta \in \mathbb{R}^{p\times p}$, we can define an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, p\}$, and $(j, j') \in \mathcal{E}$ if and only if $\Theta_{jj'} \neq 0$ for $1 \leq j, j' \leq p$ and $j \neq j'$.

The Ising model was first adopted in physics [24]. Following the terminology in physics, the p nodes are p magnetic dipoles, and the Ising parameter $\Theta_{jj'}$ is the coupling coefficient that describes the physical interaction between dipoles j and j' under the external magnetic field.

It is worth noting that Ising model is a special case of Multivariate Bernoulli model, which has been extensively studied in Dai et al. [12]. The joint distribution of X_j's following a Multivariate Bernoulli distribution takes the form

$$\begin{aligned} &P_{\mathbf{f}}(X_1 = x_1, \ldots, X_p = x_p) \\ &= \frac{1}{b(\mathbf{f})} \exp\left\{\sum_{r=1}^{p} \left(\sum_{1\leq j_1<\ldots<j_r\leq p} f^{j_1\ldots j_r} B^{j_1\ldots j_r}(x) \right) \right\}, \end{aligned} \tag{2.5}$$

where $b(\mathbf{f})$ is a normalizing constant, $B^{j_1,\ldots,j_r}(x) = x_{j_1} \ldots x_{j_r}$, and $f^{j_1\ldots j_r}$ are the natural parameters that have a bijective mapping to the general parameters $P_{\mathbf{f}}(x_1, \ldots, x_p)$ ([12]). For convenience, we use $\pi_{x_1\ldots x_p}$ to denote $P_{\mathbf{f}}(x_1, \ldots, x_p)$ in the rest of the chapter.

2.2.3 Two Transformations

Suppose that $\left(X, \tilde{X}\right)$ follows a multivariate Bernoulli distribution with the joint probability $\pi_{x_1 \ldots x_p \tilde{x}_1 \ldots \tilde{x}_p}$. There are two important transformations of π used in the log-linear regression models and the multivariate logistic regression models (Chapter 6 in McCullagh and Nelder [30]). The log-linear approach is based on the transformation $\pi \to \gamma$ defined by

$$\begin{aligned}
\gamma^{X_1} &= \log \frac{\pi_{1*\ldots*}}{\pi_{0*\ldots*}}, \qquad \ldots, \qquad \gamma^{\tilde{X}_p} = \log \frac{\pi_{*\ldots*1}}{\pi_{*\ldots*0}} \\
\gamma^{X_1 X_2} &= \log \frac{\pi_{11*\ldots*}\pi_{00*\ldots*}}{\pi_{01*\ldots*}\pi_{10*\ldots*}}, \qquad \ldots, \qquad \gamma^{\tilde{X}_{p-1}\tilde{X}_p} = \log \frac{\pi_{*\ldots*11}\pi_{*\ldots*00}}{\pi_{*\ldots*01}\pi_{*\ldots*10}} \\
&\vdots \\
\gamma^{X_1 \ldots \tilde{X}_p} &= \log \frac{\prod \pi \text{ with even number of zeros in subscript}}{\prod \pi \text{ with odd number of zeros in subscript}},
\end{aligned}$$

where $*$ denotes the geometric mean taken over the subscript. γ's are related to conditional odds ratios.

The multivariate logistic approach is based on the transformation $\pi \to \eta$ defined by

$$\begin{aligned}
\eta^{X_1} &= \log \frac{\pi_{1+\ldots+}}{\pi_{0+\ldots+}}, \qquad \ldots, \qquad \eta^{\tilde{X}_p} = \log \frac{\pi_{+\ldots+1}}{\pi_{+\ldots+0}} \\
\eta^{X_1 X_2} &= \log \frac{\pi_{11+\ldots+}\pi_{00+\ldots+}}{\pi_{01+\ldots+}\pi_{10+\ldots+}}, \qquad \ldots, \qquad \eta^{\tilde{X}_{p-1}\tilde{X}_p} = \log \frac{\pi_{+\ldots+11}\pi_{+\ldots+00}}{\pi_{+\ldots+01}\pi_{+\ldots+10}} \\
&\vdots \\
\eta^{X_1 \ldots \tilde{X}_p} &= \log \frac{\prod \pi \text{ with even number of zeros in subscript}}{\prod \pi \text{ with odd number of zeros in subscript}},
\end{aligned}$$

where $+$ denotes the summation over the subscript. η's are related to lower dimensional marginal probabilities.

Glonek [20] has studied the mapping $\pi \to (\eta, \gamma)$, where (η, γ) is a mixed parametrization. The combination of η and γ needs to follow the hierarchy principle in [20]. The mapping $\pi \to (\eta, \gamma)$ is invertible under mild conditions and Glonek [20] has proposed an inversion algorithm for getting π from (η, γ).

2.3 Binary Knockoff Procedure

In this section, we propose a method for binary feature selection, which can control FDR at a pre-specified level and maintain large enough power at the same time.

Given n random draws of $X = (X_1, \ldots, X_p)$ from a binary feature distribution F_X and n random draws of Y from a response distribution F_Y, we want to select features from $(X_1, \ldots, X_p)$ that are significantly associated with response Y, while keeping FDR below a target level. We assume the feature distribution of X is known, but we assume neither knowledge of the distribution of Y nor knowledge of the relationship between Y and X.

We assume in the rest of the chapter that the features are generated from an Ising model, though the construction of Knockoffs and the theoretical results in this section are also applicable to multivariate Bernoulli features. We focus on the case of Ising features because it has more practical applications due to the fact that there exists many estimation methods for parameters in Ising models.

The main contribution of our proposal is the construction of binary knockoffs. After constructing knockoff $\tilde{X}$ for the original X, we follow the same procedure of Model-X Knockoff in Section 2.1 to obtain the knockoff selection set $\hat{\mathcal{S}}$.

2.3.1 Exact Construction of Binary Knockoffs

Suppose that we have n independent draws from $X = (X_1, \ldots, X_p)$ following an Ising model with known coupling coefficient parameter Θ^*. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the design matrix such that each row is a draw. We first present an exact construction of Binary Knockoffs that satisfies the exchangeability condition (2.1), which is the key to FDR control of Knockoff procedure. Our proposed exact construction of Binary Knockoffs takes the following steps:

- **Step 1:** Choose a mixed parametrization (η, γ) for $\pi_{x_1 \ldots x_p \tilde{x}_1 \ldots \tilde{x}_p}$.

- **Step 2:** Calculate a part of η using the given Ising parameters of X.

- **Step 3:** Assign values to the rest of η and γ to ensure exchangeability condition (2.1) of $\left(X, \tilde{X}\right)$.

- **Step 4:** Invert the mapping $\pi \to (\eta, \gamma)$ to get π from constructed (η, γ).

- **Step 5:** Obtain the conditional distribution $\tilde{X}|X$ from π (joint distribution) and the known distribution of X (marginal distribution).

- **Step 6:** Sample knockoffs by using the conditional distribution $\tilde{X}|X$.

The inversion algorithm used in Step 4 can be found in Glonek [20]. Steps 5–6 are simple in theory. Our main effort is put on Steps 1–3 as we need the exchangeability condition to hold for FDR control.

Step 1. Note that $\left(X_1, \ldots, X_p, \tilde{X}_1, \ldots \tilde{X}_p\right)$ contains $2p$ variables including knockoffs. Let the index set $\{1, \ldots 2p\}$ correspond to the order of $\left(X, \tilde{X}\right)$. To better illustrate the choice of a mixed parametrization, we use ξ to denote a combination of (η, γ).

For any index subset $I \subset \{1, \ldots, 2p\}$ where $|I| \leq p$, we choose $\xi^{\left(X,\tilde{X}\right)_I} = \eta^{\left(X,\tilde{X}\right)_I}$, where $\left(X, \tilde{X}\right)_I$ corresponds to the superscript of η defined in Section 2.2.3.

For any index subset $J \subset \{1, \ldots, 2p\}$ where $|J| > p$, we choose $\xi^{\left(X,\tilde{X}\right)_J} = \gamma^{\left(X,\tilde{X}\right)_J}$, where $\left(X, \tilde{X}\right)_J$ corresponds to the superscript of γ defined in Section 2.2.3.

For example, when $p = 2$, we choose a mapping from π to a mixed (η, γ) as

$$\begin{aligned}\pi \to (&\eta^{X_1}, \eta^{X_2}, \eta^{\tilde{X}_1}, \eta^{\tilde{X}_2}, \eta^{X_1X_2}, \eta^{X_1\tilde{X}_1}, \eta^{X_1\tilde{X}_2}, \eta^{X_2\tilde{X}_1}, \eta^{X_2\tilde{X}_2}, \eta^{\tilde{X}_1\tilde{X}_2}, \\ &\gamma^{X_1X_2\tilde{X}_1}, \gamma^{X_1X_2\tilde{X}_2}, \gamma^{X_1\tilde{X}_1\tilde{X}_2}, \gamma^{X_2\tilde{X}_1\tilde{X}_2}, \gamma^{X_1X_2\tilde{X}_1\tilde{X}_2}).\end{aligned} \tag{2.6}$$

This type of combination of (η, γ) satisfies the hierarchy principle in [20], thus the inversion algorithm in [20] is applicable to it.

Step 2. Multivariate Bernoulli model is an extension of Ising model with $f^{jj'} = \Theta_{jj'}$ and $f^{\mathbf{J}} = 0$ for $|\mathbf{J}| > 2$, where $f^{\mathbf{J}}$ are the natural parameters in (2.5) and $\Theta_{jj'}$ are the Ising parameters in (2.4). In addition, there is a bijective mapping between the natural parameters

f and the joint probabilities (i.e. general parameters) π of a multivariate Bernoulli model. Therefore, given the Ising parameters of X, we are able to calculate $\pi_{\mathcal{I}}$ for any subset $\mathcal{I}$ of the power set of $\{x_1, \ldots, x_p\}$. The bijective transformation formula is explicitly stated in Dai et al. [12].

Continue using the example in (2.6). We are able to calculate η^{X_1}, η^{X_2}, and $\eta^{X_1X_2}$, because $\eta^{X_1}, \eta^{X_2}, \eta^{X_1X_2}$ relate to the lower dimensional marginal probabilities $\pi_{x_1}, \pi_{x_2}, \pi_{x_1x_2}$, which can be calculated from the given Ising parameters of X. We will use these η values in Step 3.

Step 3. The objective of our construction is to ensure exchangeability condition (2.1) of $\left(X, \tilde{X}\right)$. It requires appropriate assignment of η and γ values that are used for inverting back to π.

First, consider the η part. For any index subset $\mathcal{I} \subset \{1, \ldots, 2p\}$ where $|\mathcal{I}| \leq p$ and for any swapping index subset $S \subset \{1, \ldots, p\}$, we set

$$\eta^{\left(X,\tilde{X}\right)_{\mathcal{I}}} = \eta^{\left\{\left(X,\tilde{X}\right)_{\text{swap(S)}}\right\}_{\mathcal{I}}} \tag{2.7}$$

by using the calculated η values in Step 2. In the example (2.6), satisfying the condition (2.7) is equivalent to setting $\eta^{\tilde{X}_1} = \eta^{X_1}, \eta^{\tilde{X}_2} = \eta^{X_2}$, and $\eta^{X_1X_2} = \eta^{\tilde{X}_1\tilde{X}_2} = \eta^{X_1\tilde{X}_2} = \eta^{\tilde{X}_1X_2}$. Note that some η values are non-identifiable, for example, $\eta^{X_1\tilde{X}_1}, \eta^{X_2\tilde{X}_2}$ in (2.6). We can simply set them to be some arbitrary values like zeros as long as (2.7) holds.

Next, consider the γ part. We propose to set all γ to be some constant C. We recommend trying $C = 0$ when running the inversion algorithm in [20], since it slightly simplify one step of the algorithm.

Using the example in (2.6) one more time, we may consider an inverse mapping of

$$\begin{aligned}\pi \rightarrow (&\eta^{X_1}, \eta^{X_2}, \eta^{X_1}, \eta^{X_2},\\ &\eta^{X_1X_2}, 0, \eta^{X_1X_2}, \eta^{X_1X_2}, 0, \eta^{X_1X_2},\\ &0, 0, 0, 0, 0),\end{aligned} \tag{2.8}$$

and use the inversion algorithm in Glonek [20] to get $\pi_{x_1x_2\tilde{x}_1\tilde{x}_2}$.

It is easy to check that our assignment of η and γ leads to a set of joint probabilities π satisfying the exchangeability condition (2.1) of $\left(X, \tilde{X}\right)$. Furthermore, the construction of $\tilde{X}$ does not involve the response variable Y. Therefore, the Binary Knockoffs $\tilde{X}$ generated from our exact construction satisfy the two conditions for Model-X knockoffs, and consequently inherit the desirable properties of Model-X knockoffs. This leads to the following theoretical results.

2.3.2 Theoretical Results

After constructing knockoffs $\tilde{X}$, we calculate LCD statistic W_j and threshold T (depends on a pre-specified FDR control level q) following the same manner in Candès et al. [9]. $\hat{\mathcal{S}} = \{j : W_j \geq T\}$ is the Knockoff selection set. We first present the result for FDR control.

Theorem 1. *Given the Ising features X with known parameters Θ^*, and using the exact construction for sampling knockoffs $\tilde{X}$, the Knockoff selected set $\hat{\mathcal{S}} = \{j : W_j \geq T\}$ controls the FDR at a pre-specified level q. In addition, this FDR-control result is non-asymptotic and holds without knowledge of the underlying relationship between the response Y and the features X.*

The advantages of the Knockoff procedure are obvious based on Theorem 1: the FDR control result holds in finite samples and it works even if the model is mis-specified.

In addition to the FDR control, we also provide analyses on asymptotic power of the Binary Knockoff procedure. In contrast to the FDR control analysis, the power analysis requires knowledge of the true model. We first consider the case where the true relationship between Y and X is linear:

$$Y = X\beta + \epsilon,$$

where β is the unknown true coefficient vector and ϵ is an error term.

Denote X^* to be $\left(X, \tilde{X}\right)$, and $\mathbf{X}^* = [\mathbf{X}\ \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$ to be the augmented design matrix. Let $|\mathcal{S}| = s$, where $\mathcal{S}$ is the set of non-zero coefficients in the true model. Let q be the

pre-specified level of FDR that we want to control via Knockoff. Let $\hat{\beta} \in \mathbb{R}^{2p}$ be the augmented coefficient estimates from running a Lasso regression using Y and $\mathbf{X}^*$. Note that the augmented true coefficients β_T is equal to $[\beta^T, \mathbf{0}^T]^T$, because $\tilde{X}$ is constructed without looking at Y, thus irrelevant to Y.

To facilitate the power analysis, we impose the following regularity assumptions:

- **Condition 1:** The error components of ϵ are $i.i.d$ with a sub-Gaussian distribution.
- **Condition 2:** As n increases, it holds that $\left(\frac{n}{\log p}\right)^{\frac{1}{2}} \min_{j \in \mathcal{S}} |\beta_j| \to \infty$.
- **Condition 3:** With asymptotic probability one, $|\hat{\mathcal{S}}| \geq cs$ for some constant $c \in (2(qs)^{-1}, 1)$.
- **Condition 4:** Let $\Sigma_0 = \mathbb{E}\left[X^{*T} X^*\right]$ and Σ_0 satisfies compatibility condition with some constant $\phi_{\Sigma_0} > 0$, i.e.

$$||\alpha_S||_1^2 \leq \frac{s\alpha^T \Sigma_0 \alpha}{\phi_{\Sigma_0}^2} \tag{2.9}$$

  for all vectors α satisfying $||\alpha_{\mathcal{S}^C}||_1 \leq 3||\alpha_{\mathcal{S}}||_1$, where $\mathcal{S}^C$ is the complement set of $\mathcal{S}$.

The error term ϵ does not need to follow exactly a sub-Gaussian distribution. We need a concentration inequality of sub-Gaussian distribution in the proof. Any other distributions with similar concentration inequalities can replace the sub-Gaussian in Condition 1. Condition 2 ensures the asymptotic power of Lasso to be one. This condition is needed since the Knockoff procedure uses Lasso in variable selection, so its asymptotic power is upper bounded by Lasso. Condition 3 puts a lower bound on the number of selected features. The Conditions 1–3 are exactly same as the ones in the asymptotic power analysis of Model-X Knockoff in Fan et al. [16]. The analysis in Fan et al. [16] is based on the assumption of Gaussian features. In contrast, X^* is binary in our case and follows a multivariate Bernoulli distribution by construction. In order to obtain error bounds of Lasso results without Gaussian assumption, we further assume Condition 4 that imposes constraints on the smallest

eigenvalue of the covariance Σ_0. It is reasonable to assume such condition in power analysis, as many theories on Lasso require similar restriction on the smallest eigenvalue of covariance.

Theorem 2. *Assume that Condition 1–4 hold. Use the exact construction to obtain Binary Knockoffs $\tilde{X}$ and follow the Model-X Knockoff procedure to get Knockoff selection set $\hat{\mathcal{S}}$. With asymptotic probability one, $\frac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{|\mathcal{S}|} \geq 1 - O(a_n^{-1})$ for some $a_n \to \infty$, i.e. Power($\hat{\mathcal{S}}$) $\to 1$ as $n \to \infty$.*

Theorem 2 shows that under some mild conditions the asymptotic power of our knockoff procedure approaches to one when the underlying model is a linear regression model. We also provide power analysis in a logistic regression model setup. Suppose that $Y \in \{-1, 1\}$ and X are generated from an Ising model with parameters Θ^*. Consider the true model to be

$$P_{\beta_T}(Y|X^*) = \frac{\exp\{YX^*\beta_T\}}{\exp\{YX^*\beta_T\} + 1}.$$

Let $Q^* = \mathbb{E}_{\beta_T}\{\nabla^2 \log P_{\beta_T}[Y|X^*]\}$, the Fisher information matrix associated with the conditional probability distribution of $Y|X^*$. Let $Q^*_{\mathcal{S}\mathcal{S}}$ be the sub-matrix of Q^* indexed by the true non-zero coefficient set $\mathcal{S}$.

To facilitate the analysis, we further impose the following basic regularity assumptions:

- **Condition 5:** There exists some constant $C_{\min} > 0$ s.t. the minimum eigenvalue of $Q^*_{\mathcal{S}\mathcal{S}} \geq C_{\min}$ and the maximum eigenvalue of $\mathbb{E}\left[X^{*T}X^*\right] \leq D_{\max}$ for some positive constant $D_{\max}$.

- **Condition 6:** $\left|\left|\left|Q^*_{\mathcal{S}^C\mathcal{S}}(Q^*_{\mathcal{S}\mathcal{S}})^{-1}\right|\right|\right|_\infty \leq 1 - \alpha$ for some $\alpha \in (0, 1]$.

The first part of Condition 5 puts a lower bound on the eigenvalues of the Fisher information matrix corresponding to the relevant features. Moreover, the second part of Condition 5 ensures that the relevant features do not become overly dependent. Condition 6 indicates that the irrelevant features cannot have a strong effect on the relevant features.

Theorem 3. *Assume that Conditions 2–3 and 5–6 hold. With asymptotic probability one, $\frac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{|\mathcal{S}|} \geq 1 - O(b_n^{-1})$ for some $b_n \to \infty$, i.e. Power($\hat{\mathcal{S}}$) $\to 1$ as $n \to \infty$.*

2.3.3 Approximation Construction

Although the exact construction has desirable properties in terms of FDR control and asymptotic power, it has a major limitation with respect to computational cost: in Step 1 of exact construction, we need to calculate 2^p joint probabilities π values, which is computational infeasible when p is large. A computationally feasible version of construction method is in need for practical use.

Inspired by the approximation construction in Candès et al. [9], we modify the exact construction and propose a second-order approximation construction for Binary Knockoff procedure. Instead of ensuring the exchangeability condition (2.1) to hold exactly, we only ask for the first two moments in $\left(X, \tilde{X}\right)$ and $\left(X, \tilde{X}\right)_{\text{swap(S)}}$ to match. The payoff of violating the exact exchangeability condition to a small extent is a tremendous reduction in computational cost.

The second-order approximation construction of Binary Knockoffs is different from the exact construction in the first three steps. Here are the differences:

- In **Step 1**, we choose a different combination of $\xi := (\eta, \gamma)$. For any index subset $I \subset \{1, \ldots, 2p\}$ where $|I| \leq 2$, we choose $\xi^{(X,\tilde{X})_I} = \eta^{(X,\tilde{X})_I}$, where $\left(X, \tilde{X}\right)_I$ corresponds to the superscript of η defined in Section 2.2.3. For any index subset $J \subset \{1, \ldots, 2p\}$ where $|J| > 2$, we choose $\xi^{(X,\tilde{X})_J} = \gamma^{(X,\tilde{X})_J}$, where $\left(X, \tilde{X}\right)_J$ corresponds to the superscript of γ defined in Section 2.2.3. In contrast to the Step 1 in the exact construction, the number of η values in the mixed parameterization is only $\binom{2p}{2} + 2p$.

- In **Step 2**, we calculate $\eta^{X_1}, \ldots, \eta^{X_p}, \eta^{X_1 X_2}, \ldots, \eta^{X_{p-1} X_p}$, which consists of $\frac{p+p^2}{2}$ values compared to 2^p values in the exact construction.

- In **Step 3**, for any index subset $\mathcal{I} \subset \{1, \ldots, 2p\}$ where $|\mathcal{I}| \leq 2$ and for any swapping index subset $S \subset \{1, \ldots, p\}$, we set

$$\eta^{(X,\tilde{X})_{\mathcal{I}}} = \eta^{\left\{(X,\tilde{X})_{\text{swap(S)}}\right\}_{\mathcal{I}}} \tag{2.10}$$

by using the calculated η values in the modified Step 2.

For the non-identifiable η's, we set them to be some arbitrary values as long as (2.10) holds. And we still recommend setting all the γ values to be zeros.

The condition (2.10) ensures that the first two moments of $\left(X, \tilde{X}\right)$ and $\left(X, \tilde{X}\right)_{\text{swap(S)}}$ are matched. This construction is not exact because marginalizing the constructed joint distribution of $\left(X, \tilde{X}\right)$ over $\tilde{X}$ does not give back the given distribution of X. Therefore, the exchangeability condition (2.1) does not hold exactly. However, we show in the simulations that the approximation approach robustly controls FDR in practice.

2.3.4 Parameters Unknown

The ideal scenario considered in previous part may not be realistic all the time since the knowledge of the covariates distribution may not be available. Even though we model the covariate distribution using an Ising model, the true parameters are often unknown. Then it is natural to ask the question whether our Ising Knockoff procedure still controls FDR and holds the properties of the power if we use an estimated Ising parameter matrix for knockoff construction. Similar to Fan et al. [16], we may consider a data-split approach where half of the data are used for estimating Θ^* as $\hat{\Theta}$, and another half of the data for conducting Knockoff procedure. In practice, however, the data-split procedure may not be necessary as noted in the simulations of Fan et al. [16] that FDR is still controlled without data-split.

Note that in the previous power analyses, only Condition 4 for linear regression and Condition 5–6 for logistic regression involve the augmented variable X^* that contains the knockoffs. These three conditions are imposed directly on the expectations in terms of X^*. Therefore, if they hold for X^* obtained from using the estimated Ising parameters, the power analysis conclusions will be the same. However, we need a further analysis of the FDR control when using $\hat{\Theta}$ for generating the knockoffs.

FDR Analysis

Denote the FDR function of using the estimated parameter $\hat{\Theta}$ to be $\text{FDR}(\hat{\Theta})$ and the FDR function of using the true parameter Θ^* to be $\text{FDR}(\Theta^*)$. Fan et al. [16] uses some Lipschitz function for analyzing the FDR control when using estimated precision matrix of a Gaussian graphical model, however, it is not easy to extend their idea in our case. We may mimic the way in Fan et al. [16] by proposing a strong condition as the following:

- **Condition 7:** There exists some constant $L > 0$ such that for all $\left|\left|\hat{\Theta} - \Theta^*\right|\right|_F = \mathcal{O}(c_n)$ with $c_n \to \infty$,

$$\left|\text{FDR}(\hat{\Theta}) - \text{FDR}(\Theta^*)\right| \leq L \left|\left|\Theta^* - \hat{\Theta}\right|\right|_F. \tag{2.11}$$

By doing so, we bound the error term $\left|\text{FDR}(\hat{\Theta}) - \text{FDR}(\Theta^*)\right|$ by the Frobenius norm of $\left|\left|\Theta^* - \hat{\Theta}\right|\right|_F$. Note that a couple of existing methods are able to get an estimator of Ising parameters that satisfies the condition $\left|\left|\hat{\Theta} - \Theta^*\right|\right|_F = \mathcal{O}(c_n)$. For example, Xue et al. [51] proposes an estimator of Θ^* that with probability tending to 1, $\left|\left|\hat{\Theta} - \Theta^*\right|\right|_F = \mathcal{O}(\sqrt{\frac{s_1}{n}})$ for some constant s_1. Therefore, if Condition 7 holds, with high probability the estimated Ising Knockoff procedure can asymptotically control FDR at a target level. The difficulty remained is to check whether Condition 7 holds or not. We leave this part for future study.

2.4 Empirical Results

2.4.1 Simulations

Simulation setup. We compare the second-order approximation method of Binary Knockoff procedure with the approximation method of Model-X Knockoff procedure proposed by Candès et al. [9]. Both linear regression model and logistic regression model are considered in simulations. Note that in real data applications we usually do not know the true parameters of the features distribution. To show that our second-order approximation method has a robust performance on FDR control, in simulations we first estimate the first two moments

of X via sample mean and sample variance, and then use these estimated first two moments in the second-order approximation method. A similar estimation procedure is used in the real data analysis in next section.

In a low dimensional linear regression model setup, we generate $n = 400$ samples for 40 subgroups of features and each subgroup contains five features generated from an Ising model (i.e. $p = 200$). We randomly set 30 out of 200 coefficients β_j's to be $\pm L$ with L ranging from 0.2 to 0.5, and all the rest coefficients are set to be zeros. In a high dimensional linear regression model setup, we generate $n = 400$ samples for 120 subgroups of features and each subgroup contains five features generated from an Ising model (i.e. $p = 600$). We randomly set 30 out of 600 coefficients β_j's to be $\pm L$ with L ranging from 0.2 to 0.5, and all the rest coefficients are set to be zeros. In a logistic regression model setup, we generate $n = 400$ samples for 40 subgroups of features and each subgroup contains five features generated from an Ising model (i.e. $p = 200$). We randomly set 30 coefficients β_j's to be $\pm L$ with L ranging from 0.5 to 2.5, and all the rest coefficients are zeros. In all three setting, the target FDR control level is 0.2.

Although the second-order approximation construction violates the exchangeability condition (2.1), we see from Figure 2.1 that it still controls FDR under a pre-specified level in practice. Moreover, our approximation method of Binary Knockoff procedure has much higher power than the approximation method of Model-X Knockoff in all three simulation scenarios. One major difference between these two approximation methods is that the knockoffs generated by Binary Knockoff procedure are binary while the knockoffs constructed by Model-X Knockoff are continuous. It is more natural and reasonable to construct binary knockoffs for binary features; this may partially explain the gain of power in Binary Knockoff procedure in our simulations.

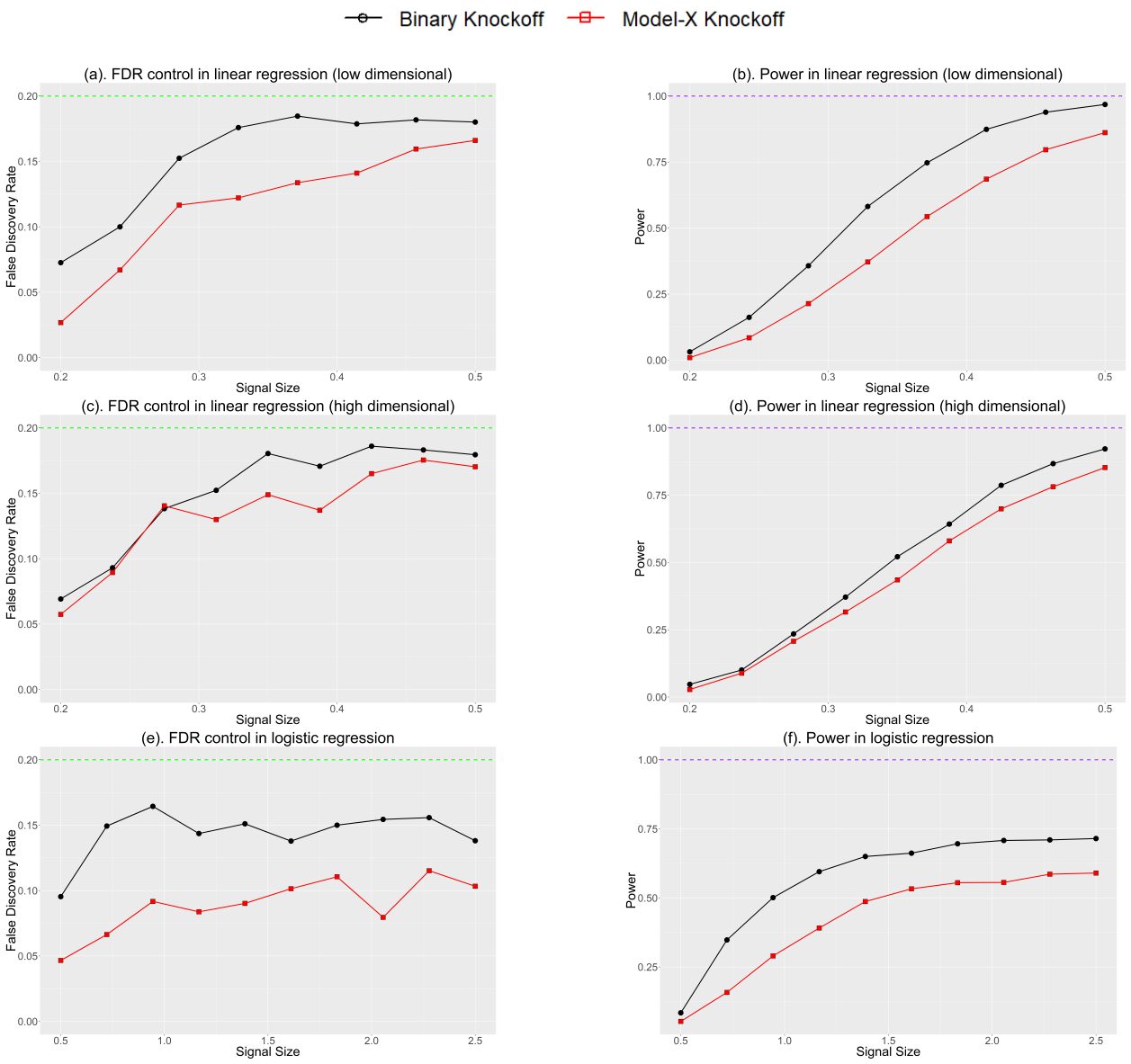

Figure 2.1: Simulation results from second-order approximation Binary Knockoff and approximation Model-X Knockoff. For (a) and (b), the model is linear regression with $n = 400$ and $p = 200$. For (c) and (d), the model is linear regression with $n = 400$ and $p = 600$. For (e) and (f), the model is logistic regression with $n = 400$ and $p = 200$.

2.4.2 Real Data

We also illustrate the practical utility of Binary Knockoff procedure using a HIV antiretroviral therapy (ART) susceptibility dataset from the Stanford HIV drug resistance database. This dataset contains virus mutation information at protease residues for 702 isolates from the plasma of HIV-1-infected patients. Rhee et al. [36] and Wu et al. [47] have used this dataset for studying the association between protease mutations and susceptibility to ART drugs. It has also been used in Xue et al. [51] to study the graphical model of the protease residues, and Xue et al. [51] model these protease residues using Ising graphical models.

We treat the protease residues as features and the amprenavir (APV) level as the response, and assume that their relationship follows a linear model. The mutations on each protease residue are recorded as binary values, so all the features of this dataset are binary, in which case the Binary Knockoff procedure is a more natural choice than other existing Knockoff procedures. Similar to a previous study in Xue et al. [51], we assume that all the features are generated from an Ising model, and can be partitioned into subgroups based on the stable edge graphs in the Figure 2 of Xue et al. [51]. Our analysis uses $p = 19$ of the residues that have at least 20% of the values to be 1.

We apply the proposed second-order approximation Binary Knockoff method on this real data for controlling FDR at the level of 0.2. The first two moments of the features X are estimated via sample mean and sample variance. We also provide the variable selection result by the approximation Model-X Knockoff procedure for comparison.

Table 2.1 summarizes the results of two approximation methods. Since we do not know the ground truth, we search over genome science literature to find claims that support the association between the APV susceptibility and the residues frequently selected in this table. For example, Mittal et al. [31] studies the association between APV and mutations at residue 50. The Table 1 in Martinez-Cajas et al. [29] presents the APV resistance mutations at residue 33 and 36. Moreover, the Figure 1 in Rhee et al. [36] shows their study about the association between APV and some residues listed in our Table 2.1. Most of the frequently

Residue	Selection by BKF	Selection by MKF
No.33	82%	70%
No.84	82%	70%
No.46	82%	68%
No.13	80%	63%
No.36	77%	64%
No.54	76%	63%
No.77	71%	52%
No.50	70%	52%

Table 2.1: The table of selection frequencies by Binary Knockoff procedure (BKF) and by Model-X Knockoff procedure (MKF) for residues being selected more than half times by both methods.

selected residues in Table 2.1 have literature supporting their association with APV, so we argue that most of the residues listed in Table 2.1 are not false discoveries. In addition, both Knockoff procedures tend to select same residues, while the Binary Knockoff procedure has much higher selection frequencies than the existing Knockoff procedure. This result indicates that the Binary Knockoff procedure has a higher power, which matches the comparison results in the previous simulation studies.

2.5 Conclusions and Discussions

In this chapter, we proposed Binary Knockoff procedure, an FDR controlled variable selection method tailored to binary features in regression framework. Since Ising model is commonly adopted for modeling the relationship among binary variables and has gained popularity in machine learning literature, this is a natural alternative to the Model-X knockoff in Candès

et al. [9] and RANK in Fan et al. [16] in the binary features setting. We provide both exact construction and second-order approximation construction of Binary Knockoff procedure. The exact construction leads to attractive theoretical results of FDR control and asymptotic power, and we show in empirical results that the second-order approximation method also controls FDR well in practice.

We note that the way of constructing Binary Knockoffs in this chapter can be easily extended to features generated from multivariate Bernoulli model, as the Ising model is a special case of the multivariate Bernoulli model. We expect that Ising model is probably useful enough for most practical applications.

In spite of the good theoretical properties and empirical performance, our current proposal still have some limitations and thus can be improved in future research work. The inversion algorithm [20] we used in Step 4 of the construction procedure requires a good initial value for convergence. And it does not guarantee a valid output π (i.e. all components of π are non-negative) if there are some extreme η values in input. It may happen when some binary features have very few 1's or 0's in a large sample, which indicates an extreme η value is possible during the calculation in Step 2. To our best knowledge, this problem has not been solved in literature related to multivariate logistic regression model where the transformation $\pi \to \eta$ is frequently used. We leave this problem to future research work.

Chapter 3

Heterogeneous Treatment Effect Detection Models with False Discovery Rate Control

Part of this work has been published in the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Xie et al. [50]).

3.1 Introduction

As grown personalized medicine popularity comes with greater needs, researchers are no longer satisfied with knowing whether a treatment is different from the placebo in the overall study population or not, but have also become interested in knowing 'which factors' drive the changes in treatment effects or 'which subpopulations' are chief beneficiaries of a treatment. A similar story is taking place in industries where online controlled experiments have been used as the mantra for data-driven decision making on feature changing and product shipping. The experimenters, who used to make strategies based on average treatment effect (ATE) only, have become increasingly interested in studying user heterogeneity to improve

the product. The increasing needs have brought popularity to the study of heterogeneous treatment effect (HTE) for the past decade, which leads to many insightful ideas about learning the HTE.

Imai and Ratkovic [23] proposes to estimate the treatment effect heterogeneity in a randomized evaluation program by using Squared Loss Support Vector Machine with Lasso. The key part of their approach is to put two separate l_1 penalties on the coefficient estimates for the pre-treatment covariates and the coefficient estimates for the interaction between treatment and pre-treatment covariates. Deng et al. [13] proposes a total variation regularized regression model to understand the structure of the HTE. In addition to the use of Lasso for covariate selection, they further include a total variation penalty in the regression model to encourage block-wise structure for the non-zero coefficients of the potential covariates. Athey and Imbens [2] uses machine learning methods to estimate HTE, and Wager and Athey [45] develops a causal forest method based on the idea of the causal tree model in Athey and Imbens [2] and the extension of the random forest algorithms proposed in Breiman [5]. However, this work also focuses on estimation inference on the HTE instead of understanding systematically which subgroups differ from average users or which variables contributes to the heterogeneity in treatment effects.

On the other hand, with the overly affluent amount of data, there is a strong threat from false discoveries, due to a statistical artifact known as 'multiple testing'. With the hundreds of thousands of user characteristics available to researchers, one can construct user groups in millions of ways. If we take a 'naive' approach by simply computing and comparing the estimated effect based on users within groups, we can always easily find groups with treatment effects that differ widely from the average population, regardless of whether there is a real heterogeneity or not. Although the above-mentioned works have contributed a lot into the study of estimating the HTE or drawing inference from the HTE, they do not deal with the potential multiple testing problem when conducting analysis for the HTE. For example, using l_1 penalty in a regression model may help us to select a set of variables that potentially causes the heterogeneity in treatment effect, but it is possible that a large

proportion of the selected variables are false positives.

The goal of our work is to fill such gap by providing two statistical methods that can detect variables contributing to HTE while dealing with the potential multiple testing problem by controlling the false positive rate (FDR). We focus on the case of randomized trial in this project. The rest of this chapter is organized as follows. In Section 3.2, we introduce notations and preliminaries that play important roles in our statistical framework. In Section 3.3, we present two proposed methods and establish theoretical properties, which include the control of FDR and asymptotic power analysis. Simulation studies are presented in Section 3.4, and a real data application is shown in Section 3.5. We close with a discussion in Section 3.6.

3.2 Preliminaries

3.2.1 Average Treatment Effect vs. Heterogeneous Treatment Effect

In a randomized clinical trial, we randomly split users into a treatment group and a control group, and observe the response of interest for all the users. The Rubin Causal Model [21] is commonly used as a statistical framework for causal inference. Define $Y_i(T_i)$ to be the potential outcome for i-th user, where $T_i = 1$ if the i-th user is in the treatment group and $T_i = 0$ if the i-th user is in the control group. Therefore, $\tau_i = Y_i(1) - Y_i(0)$ is the causal effect of taking the treatment for i-th unit, and the average causal effect of all users $\bar{\tau}$ is defined as the Average Treatment Effect (ATE). Note that ATE is not observable since we do not know $Y_i(0)$ and $Y_i(1)$ at the same time. This is known as the 'fundamental problem of causal inference' [21]. However, The estimator

$$\overline{Y_{i|T_i=1}} - \overline{Y_{i|T_i=0}} \tag{3.1}$$

is unbiased for ATE when the following two assumptions hold and is usually used for estimating ATE in a randomized trial.

Assumption 1. *Stable unit treatment value assumption (SUTVA):*

- *Only one version of treatment and control, i.e. only one version of* $T = 1$ *and* $T = 0$.
- *Treatment applied to one user does not affect the outcome of other user (no interference).*

Assumption 2. *Unconfoundedness:*

$$T_i \perp (Y_i(0), Y_i(1)) \,|X_i, \tag{3.2}$$

where X_i *is a set of the pre-treatment variables for* i*-th user, for example age, gender, country, etc.*

However, the analysis based on ATE only is not enough for obtaining accurate and meaningful insights when the population is heterogeneous. It is possible for the ATE to exaggerate the positive treatment effect of one sub-population while neglecting the negative treatment effect of another sub-population. In order to study heterogeneous treatment effect, we need to consider the conditional average treatment effect, which is defined as

$$\tau(x) = \mathbf{E}\left[Y_i(1) - Y_i(0) | X_i = x\right], \tag{3.3}$$

where X_i is a set of pre-treatment variables for i-th user. Obtaining accurate estimates of the conditional average treatment effect $\tau(x)$ for all values of x is very useful for heterogeneous treatment effect detection, because $\tau(x)$ gives the conditional average treatment effect for the subpopulation defined by the covariates $X = x$. For example, any statistically significant change in $\tau(x)$ due to a change in covariate X_j may indicate that X_j is a factor that contributes to HTE.

3.2.2 Transformed Outcomes

To detect variables contributing to heterogeneity in treatment, we need to estimate the conditional average treatment effects defined in (3.3) for the subgroups $X = x$. Due to

the fundamental problem of causal inference, we are not able to directly observe individual treatment effect. However, we are able to construct a transformed outcome (TO) for each user as an alternative measure of individual treatment effect. Let Y_i^{obs} be the observed outcome for i-th unit. In addition, let ν be the assignment probability, which, following the terminology of online controlled experiments, is the traffic percentage assigned to treatment group in a randomized trial. Let $T = 1$ for treatment group and $T = 0$ for control group. Then the transformed outcome for the i-th unit, Y_i^*, is then defined as:

Definition 2. *(Transformed Outcome):*

$$Y_i^* = Y_i^{obs} \times \frac{(T_i - \nu)}{\nu(1-\nu)}. \tag{3.4}$$

A desirable property of the TO is that under the unconfoundedness assumption the conditional expectation $\mathbf{E}\left[Y_i^* | X_i = x\right]$ equals the conditional average treatment effect $\tau(x)$ [2], which is the key for developing one of our proposed methods.

3.2.3 Transformed Covariates Model

Fitting a simple multivariate linear regression model to characterize the interaction between the treatment T and covariates X is one way to study heterogeneous treatment effects [23, 40]. Let $T = 1$ for treatment group and $T = -1$ for control group. For rest of the paper, we assume the assignment probability $P(T = 1) = \frac{1}{2}$ without loss of generality. Then given a response variable Y, consider a working model

$$Y = X\beta + XT\gamma/2 + \epsilon, \tag{3.5}$$

where ϵ is a mean zero random error. The interaction part $XT\gamma/2$ models the HTE, and by figuring out the statistically significant interactions we can find out the covariates that contribute the heterogeneity in treatment effects. Note that under model (3.5), a version of transformed outcome is $Y_i^* = 2Y_iT_i$, which has the same desirable property of unbiasedness as the one in (3.4).

Tian et al. [40] has proposed a transformed covariates model for estimating interactions between a treatment and a large number of covariates in randomized trials. The novel part of their proposal is that in their model there is no need for modeling the main effects. Define the transformed covariates to be $Z = XT/2$. When Y is a continuous response, they consider a simple working model which is equivalent to (3.5) without main effects:

$$Y = Z\gamma + \epsilon. \tag{3.6}$$

When Y is a binary response, they propose to fit a logistic regression model:

$$P(Y = 1|X, T) = \frac{\exp(Z\gamma)}{1 + \exp(Z\gamma)}. \tag{3.7}$$

And when Y is some survival outcome, they propose to fit a Cox regression model with transformed covariates Z.

Even when their working model is mis-specified, which is likely to happen due to its lack of main effects, Tian et al. [40] has shown that the maximum likelihood estimator or the maximum partial likelihood estimator (for Cox model) of the working model converges to a deterministic γ^*, which can be used to characterize the covariate-specific treatment effect and detect heterogeneity in treatment effect. Therefore, the transformed covariates approach has a causal interpretation in a randomized trial for the resulting estimator, and it is robust to model misspecification. However, as demonstrated in their numerical analysis, this method is not immune to problems such as multiple testing and false discoveries.

3.2.4 The Family of Knockoff Methods

Barber and Candès [3] proposes 'Knockoff', an FDR controlled variable selection method in regression framework. The original knockoff procedure in [3] assumes that the underlying model is a linear regression with Gaussian and homoscedastic error term, and the dataset needs to be low dimensional (i.e. the number of observations is greater than the number of covariates fitted in a model).

Suppose that we are interested in a regression model of a response variable Y on covariates $X = (X_1, \ldots, X_p)$. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the design matrix of X. The original knockoff procedure constructs the knockoffs $\tilde{X} = \left(\tilde{X}_1, \ldots, \tilde{X}_p\right)$ such that the knockoff matrix satisfies $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X} = \mathbf{\Sigma}$ and $\mathbf{X}^T\tilde{\mathbf{X}} = \mathbf{\Sigma} - \mathbf{diag\{s\}}$, where $\mathbf{s}$ is some non-negative vector that needs to be derived during the construction. After constructing a knockoff $\tilde{X}$ for the original X, Barber and Candès [3] discusses several ways of computing statistics W_j for each variable $X_j, j \in 1, \ldots, p$. We focus on the Lasso coefficient-difference (LCD) statistic since Candès et al. [9] has studied LCD extensively in their theoretical and numerical analysis and has shown empirically that LCD has higher power than some other options mentioned in Barber and Candès [3]. To compute the LCD statistic, we first solve a lasso type regression problem on the augmented design matrix $\mathbf{X}^* = \left[\mathbf{X}, \tilde{\mathbf{X}}\right]$, and denote the solution by $\hat{\beta}(\lambda)$, where the tuning parameter λ is selected by cross-validation in the proposal of [9]. Then set $Z_j = \left|\hat{\beta}_j(\lambda)\right|$ and $\tilde{Z}_j = \left|\hat{\beta}_{j+p}(\lambda)\right|$. The LCD statistic is defined to be

$$W_j = Z_j - \tilde{Z}_j = \left|\hat{\beta}_j(\lambda)\right| - \left|\hat{\beta}_{j+p}(\lambda)\right|. \tag{3.8}$$

Following many Knockoff papers [3, 9, 16, 49], we also assume that there are no ties in the magnitude of nonzero W_j's and no ties in the nonzero components of the Lasso solution with asymptotic probability one. For a given $q \in (0, 1)$, we choose a positive threshold T as

$$T = \min\left\{t > 0 : \frac{\#\{j : W_j \leq -t\} + 1}{\#\{j : W_j \geq t\}} \leq q\right\}, \tag{3.9}$$

and the Knockoff selected set $\hat{S} = \{j : W_j \geq T\}$ controls the FDR at the level of q.

Candès et al. [9] extends the idea of 'knockoff' to a high dimensional setting. The Model-X Knockoff in [9] allows for misspecification of model, but instead requires the knowledge of the covariates distribution. $\tilde{X} = \left(\tilde{X}_1, \ldots, \tilde{X}_p\right)$ are Model-X knockoffs for the original random variables $X = (X_1, \ldots, X_p)$ if $\tilde{X}$ are constructed without looking at the response variable Y and for any subset $S \subset \{1, \ldots, p\}$,

$$\left(X, \tilde{X}\right)_{\text{swap(S)}} \stackrel{d}{=} \left(X, \tilde{X}\right), \tag{3.10}$$

where $\left(X, \tilde{X}\right)_{\text{swap(S)}}$ means swapping the X_j and $\tilde{X}_j$ for all $j \in S$. Candès et al. [9] provides a way of constructing knockoffs for Gaussian distributed covariates so that (2.1) holds exactly, and a second-order approximation for constructing knockoffs in the case where the covariates are not Gaussian such that the first two moments of the two sides of (3.10) are same.

Fan et al. [16] later studies the power of the Model-X Knockoff and show that under some mild conditions, the Model-X Knockoff with Gaussian covariates achieves asymptotic power one in a linear regression model. In Chapter 2 of this dissertation, we propose a variant of Model-X Knockoff that is tailored to binary covariates generated from an Ising graphical model or a multivariate Bernoulli model, and it also provides power analysis on the binary knockoff procedure assuming the underlying model is a linear regression or a logistic regression.

3.3 Methods

Suppose that we have n random draws of $(Y_i, \ldots, X_{i1}, \ldots, X_{ip}, T_i)$, for $i = 1, \ldots, n$, where Y is the response of interest, $X_1, \ldots, X_p$ are the covariates, and T is the treatment indicator. Each user is either in the treatment group or the control group. Our objective is to identify X_j's that contributes to heterogeneous treatment effects in Y. Inspired by the methods we have discussed in the previous section, we propose two methods for controlling FDR while selecting heterogeneous variables in a regression framework. In this paper we assume that the treatment is randomly assigned to the subjects, i.e. $T \perp\!\!\!\perp X$. Without loss of generality, we assume that the assignment probability $\nu = \frac{1}{2}$, though our methods are applicable to arbitrary $0 < \nu < 1$.

3.3.1 Method 1: HTE-TO

We first propose the method 'HTE-TO' to detect the variables that contribute to the heterogeneity in treatment effects while controlling FDR. This approach is based on the transformed

outcome framework and Knockoff procedure.

HTE-TO method:

- **Step 1:** Construct a design matrix $\mathbf{X}$ based on the set of the pre-treatment variables $X = (X_1, \ldots, X_p)$.

- **Step 2:** Calculate the transformed outcomes Y^* for all users based on the formula in Equation (3.4). Let $\mathbf{Y}^*$ be the vector of the resulted outcomes.

- **Step 3:** Choose an appropriate Knockoff procedure based on prior knowledge on X, and create a knockoff matrix $\tilde{\mathbf{X}}$ of $\mathbf{X}$.

- **Step 4:** Run a Lasso type regression using $\mathbf{Y}^*$ as the response and $\mathbf{X}^* = [\mathbf{X}\ \tilde{\mathbf{X}}]$ as the design matrix.

- **Step 5:** Follow the procedure of the Knockoff method adopted in Step 3 to get the knockoff selection set of the heterogeneous variables.

Due to the property that $\mathbf{E}[Y_i^*|X_i = x] = \tau(x)$, identifying the variables X_j's that are associated with Y^* is one way to figure out variables that contribute to heterogeneous treatment effect. Therefore, the transformed outcome framework enables researchers to apply existing variable selection methods (e.g. Lasso) off-the-shelf without any modification.

The use of Knockoff procedures helps control the FDR at a pre-specified level. There exists a handful of knockoff methods in literature, and we need to choose the most appropriate one in order to achieve large enough power. For example, if we know the covariates X are generated from a Gaussian distribution, then the exact construction of Model-X knockoff [9] may be the best available method used in Step 3. If X_j are all taking binary values, for example, the variables in datasets from many online controlled experiments, then the binary knockoff procedure [49] may be a better choice.

Furthermore, a good choice of Lasso-type regression method is important in Step 4. Basically, if we know the conditional distribution of $Y^*|X$ to be linear or logistic, then we

may choose linear Lasso or logistic Lasso in Step 4, respectively. But it is usually difficult to know the true relationship between the transformed outcome Y^* and the covariates X, even if we know the conditional distribution of $Y|X$. Both Model-X knockoff [9] and binary knockoff [49] guarantee the FDR control even when the underlying model is mis-specified, however, most power analyses on knockoff procedures assume the knowledge of the true model. The assumptions required by these power analyses, therefore, may not be reasonable for HTE-TO.

3.3.2 Method 2: HTE-TC

Our second proposal 'HTE-TC' is based on the transformed covariates model [40] and Knockoff procedures. In contrast to HTE-TO, we keep the outcome Y unchanged but modify the covariates X.

We first consider a linear working model:

$$Y = Z\beta + \epsilon, \tag{3.11}$$

where $Z = XT/2$, β is the unknown coefficient vector, and ϵ is some random error. Following Tian et al. [40], figuring out nonzero coefficients in β is equivalent to identifying the subgroups of covariates contributing to HTE, so the HTE detection problem boils down to a variable selection problem. In Tian et al. [40], they use Lasso for selecting non-zero β, however, much more false discoveries are selected than true positives as shown in their numerical analysis. We deal with this issue by using the idea of Knockoff; this requires a construction of knockoffs $\tilde{Z}$ for the transformed covariates Z. We discuss the construction of $\tilde{Z}$ in different cases of Z.

Case 1.

Suppose that the design matrix $\mathbf{Z}$ is fixed and low dimensional. In this case, it is easy to construct knockoff $\tilde{\mathbf{Z}}$ by using the construction in Barber and Candès [3]. And then follow the knockoff procedure to obtain the selection set $\hat{S}$. □

Case 2.

If we treat $\mathbf{Z}$ as random, and suppose that we know the distribution of X, then the following lemma is useful for deriving a simple method of constructing knockoffs for Z when we are able to construct valid knockoffs for X.

Lemma 1. *Suppose that $\tilde{X}$ are knockoffs for X that satisfies the Model-X Knockoff conditions (3.10). In a randomized trial, $\tilde{X}T/2$ are valid knockoffs for the transformed covariates Z.*

Proof. Recall (3.10) that, for any subset $S \subset \{1, \ldots, p\}$,

$$\left(X, \tilde{X}\right)_{\text{swap(S)}} \stackrel{d}{=} \left(X, \tilde{X}\right).$$

In a randomized trial, X and T are independent, and T takes ± 1 in this case. Therefore, $\left(XT, \tilde{X}T\right)_{\text{swap(S)}} \stackrel{d}{=} \left(XT, \tilde{X}T\right)$. Since this construction does not involve the response Y, it satisfies the two conditions required by valid knockoff construction. □

This lemma provides a simple solution to the construction of knockoffs for transformed covariates Z when we are able to construct valid knockoffs for original covariates X. For example, if X are mean zero Gaussian covariates, then we can use the exact construction of Model-X Knockoffs for Gaussian covariates in Candès et al. [9], and multiply the obtained knockoffs $\tilde{X}$ by the given $\frac{T}{2}$ to get knockoffs for the transformed covariates Z. If X are generated from an Ising graphical model, then we may follow the construction method in Xie and Chan [49] to get knockoffs $\tilde{X}$ for X, and obtain knockoffs for Z easily. Based on the FDR control property of Model-X Knockoff [9] and Binary Knockoff [49], our proposal also controls FDR under a pre-specified level. In addition, if we assume the knowledge of underlying model, then the power analysis results of knockoff procedures in [16] and [49] are also valid in our case. □

Case 3.

Assume that $\mathbf{Z}$ is random, and we do not know the distribution of X exactly. In this case, we can split the data of X into a training set and a test set. In the training set, we estimate

the distribution of X, and then use the estimated distribution of X to construct knockoffs in the test set as in Case 2. Literatures [9, 16, 49] have shown in simulations that Model-X and Binary Knockoff procedures with such estimation method robustly control FDR in many scenarios. □

With correct construction of knockoff $\tilde{Z}$ for Z, HTE-TC controls FDR at a target level. The application of HTE-TC is not limited to linear regression model, but also other generalized linear models including logistic regression model and Cox model. In contrast to HTE-TO, due to the construction of Z it is usually more reasonable to assume the relationship between Y and Z to be linear or logistic, particularly if we have some prior knowledge about $Y|X$. Therefore, some existing power analyses results on knockoff procedures would hold in HTE-TC as well.

3.3.3 HTE-TO vs. HTE-TC

Whether to transform the outcome or to modify the covariates surely depends on the context. When the outcome is continuous, both HTE-TO and HTE-TC are applicable by fitting some linear regression models. When the outcome is binary, we can still fit a linear regression model in both methods, but we have another model option in HTE-TC, that is, to fit a logistic regression model with transformed covariates.

The wider application of HTE-TC makes it a more favorable choice when the outcome is survival time, which we do not observe the exact number for all patients due to censoring. In this case, we usually have more than one random variable for the outcome: the survival time S, the censoring time C, and the censoring indicator $I(S < C)$. It is difficult to transform outcomes in HTE-TO given this kind of survival response, therefore, HTE-TC is a more natural choice. In this survival outcome example, the proposed Cox regression model with transformed covariates is

$$\lambda\left(t|Z,T\right) = \lambda_0(t)\exp\left(Z\beta\right), \tag{3.12}$$

where $\lambda(t|\cdot)$ is the hazard function for survival time S and $\lambda_0(\cdot)$ is a baseline hazard function

independent of Z and T. With the working model (3.12) and the knockoff $\tilde{Z}$ for Z, we can conduct a Knockoff method to control FDR while selecting heterogeneous features by applying Lasso in Cox model to get the knockoff statistics (e.g. LCD) for all covariates. This is how HTE-TC works for survival outcome cases.

3.3.4 Covariate Adjustment

Covariate adjustment is unavoidable in many real data analyses. It affords opportunities for improving efficiency of inferences and correcting for chance imbalances in some exposure factors across treatment arms. Moreover, the covariate-adjusted versions of the methods have several applications to non-randomized comparisons in trials. In order to make our methods applicable to the cases where covariate adjustment is needed, in this section we propose simple modifications of HTE-TO to address the covariate adjustment issue, and discuss about the compatibility of HTE-TC on covariate adjustment. We will show in the simulation section that both methods empirically control FDR and attain large enough power in many settings.

Let $(X_1, \ldots, X_p)$ be the potential modifiers of treatment effects and $(V_1, \ldots, V_k)$ be the covariates that need to be adjusted.

Covariate-adjusted version of HTE-TO:

- **Step 1:** Construct a design matrix $\mathbf{W}$ based on the set of the pre-treatment variables $W = (X_1, \ldots, X_p, V_1, \ldots, V_k)$.

- **Step 2:** Calculate the transformed outcomes Y^* for all users based on the formula in Equation (3.4). Let $\mathbf{Y}^*$ be the vector of the resulted outcomes.

- **Step 3:** Run a linear regression of $\mathbf{Y}^*$ against $\mathbf{W}$ and obtain the coefficient estimates $\hat{\beta}_1, \ldots, \hat{\beta}_{p+k}$.

- **Step 4:** Calculate the residual $\tilde{\mathbf{Y}} = \mathbf{Y}^* - (\hat{\beta}_{p+1}\mathbf{W}_{\cdot(p+1)} + \ldots + \hat{\beta}_{p+k}\mathbf{W}_{\cdot(p+k)})$.

- **Step 5:** Choose an appropriate Knockoff procedure based on prior knowledge on $(X_1, \ldots, X_p)$, and create a knockoff matrix $\tilde{\mathbf{X}}$ of $\mathbf{X}$, where $\mathbf{X}$ is the first p columns of $\mathbf{W}$.

- **Step 6:** Run a Lasso type regression using $\tilde{\mathbf{Y}}$ as the response and $\mathbf{X}^* = [\mathbf{X}\ \tilde{\mathbf{X}}]$ as the design matrix.

- **Step 7:** Follow the procedure of the Knockoff method to get the knockoff selection set of the heterogeneous variables.

The modification of HTE-TO is simple: instead of using the original transformed outcome Y^*, we use the residual $\tilde{Y}$ as the response in model. By doing so, the potential effects caused by adjusted covariates V are reduced, and the working model in consideration is

$$\tilde{Y} = X\beta + \epsilon, \tag{3.13}$$

where ϵ is some random error. Then we apply Knockoff procedure using this working model (3.13).

As we mention in Section 3.3.3, HTE-TO does not deal with survival outcome case, whereas HTE-TO is still applicable to it. The same story holds when it comes to covariate adjustment, as we are not able to compute the transformed outcome Y^* from a survival response Y, not to mention the residual $\tilde{Y}$. So we need to rely on HTE-TC again.

Suppose that the true model with covariate adjustment is

$$\lambda\left(t|X, V, T\right) = \lambda_0(t) \exp\left\{-\left(X\beta + V\alpha + XT\gamma/2\right)\right\}, \tag{3.14}$$

where $\lambda(t|\cdot)$ is the hazard function for survival time S, $\lambda_0(\cdot)$ is a baseline hazard function independent of X, V and T, and V are covariates for adjustment. The working model considered in HTE-TC is

$$\lambda\left(t|X, T\right) = \lambda_0(t) \exp\left\{-\left(XT\gamma/2\right)\right\}, \tag{3.15}$$

where the adjusted covariates V are not involved. Although we mis-specify the model using HTE-TC, Tian et al. [40] has shown that even when the working model is mis-specified, the maximum partial likelihood estimator of the working model (3.15) still converges to a deterministic limit that can be used to characterize the covariate-specific treatment effect. Therefore, HTE-TC can be used without modification for FDR controlled HTE detection even when the true model involves some covariates that need to be adjusted for. A potential cost of neglecting the adjusted covariates in our working model, however, could be a loss of power in HTE detection.

3.4 Simulation

In this section we conduct simulations to compare HTE-TO and HTE-TC in various settings of Gaussian linear regression model. We also illustrate the wide application of HTE-TC by presenting its performance of HTE detection and FDR control in Cox model.

3.4.1 Continuous Outcome

We first consider the case where the response Y is continuous. We generate n observations of X from Gaussian distribution with mean zero. The correlation between X_k and X_l is $\rho^{|l-k|}$, for $1 \leq k, l \leq p$. We randomly sample the treatment indicator T with $P(T = 1) = \frac{1}{2}$. Then we generate n independent samples from two types of linear regression models.

The first true model does not involve the main effects of X, i.e. the linear model in (3.11):

$$Y = Z\beta + \epsilon,$$

where $Z = XT/2$, and ϵ is standard normal error. We set $n = 400$ and $p = 200$, with 30 out of 200 interaction terms to have nonzero coefficients ranging from 0.3 to 0.7. The correlation factor ρ is set to be 0.6.

The second scenario is when the true model involves the main effects of X, i.e. the linear

model in (3.5):

$$Y = \beta^T X + \alpha T + \gamma^T XT/2 + \epsilon.$$

We set $\rho = 0.6$, and choose $n = 200$ and $p = 100$ with the first 30 interaction terms $\gamma_1, \ldots, \gamma_{30}$ having nonzero coefficients. We fix the coefficients of the main effect terms such that the coefficients of the first 30 covariates $\beta_1, \ldots, \beta_{30}$ are 0.5 and the coefficients $\beta_{31}, \ldots, \beta_{40}$ are 0.1. We set the coefficient of the treatment indicator $\alpha = 0.4$. In this case, HTE-TC mis-specifies the model.

All the results in Figures 3.1–3.2 are averaged over 100 replicates. As we can see from the figures that in a Gaussian linear regression model, both methods control FDR under the pre-specified level $q = 0.2$ and they perform quite similar in terms of power.

3.4.2 Survival Outcome

For the survival outcome cases, we first generate $p = 200$ covariates X from Gaussian distribution as previous but with $\rho = 0.3$. And then we generate $n = 400$ independent survival times from the regression model

$$S = \exp\left\{\beta^T X + \alpha T + \gamma^T XT/2 + \epsilon\right\},$$

where ϵ is a normally distributed with mean zero and standard deviation equal to 0.5. We fix α at 0.4, and set $\beta_1, \ldots, \beta_{30}$ to be 0.5, $\beta_{31}, \ldots, \beta_{90}$ to be 0.05, and the rest β to be zeros. For the interaction terms, we set $\gamma_1, \ldots, \gamma_{30}$ to range from 1 to 2, as increasing these signal values leads to increment in detection power. The rest of γ's are set zeros to induce sparsity. The censoring time is generated from the uniform distribution $U(0, A)$, where A is selected to induce a censoring rate around 20%.

Since HTE-TO is not applicable to the survival outcomes, we only present the results of HTE-TC. As Figures 3.3 show, HTE-TC successfully controls the FDR under the pre-specified level 0.2, and at the same time attains enough power for detection.

3.4.3 Covariate Adjustment

In order to see that covariate adjusted version of HTE-TO and HTE-TC control FDR and achieve good enough power when the true model involves covariates for adjustment, we consider generating data from the following two models. One is a Gaussian linear regression model including adjusted covariates:

$$Y = \beta^T X + \alpha T + \gamma^T XT/2 + \omega^T V + \epsilon,$$

and another is a covariate adjusted Cox model with survival time generated as

$$S = \exp\left\{\beta^T X + \alpha T + \gamma^T XT/2 + \omega^T V + \epsilon\right\},$$

where $\alpha, \beta, \gamma, \epsilon$ in two models are chosen same as their non-covariate-adjustment counterparts in previous simulations, and $V \in \mathbb{R}^{10}$ are covariates that should be adjusted in the model with corresponding coefficients $\omega_j = 0.5$ for $j = 1, \ldots, 10$.

Figure 3.4 shows that both HTE-TO and HTE-TC work well when the generating model is Gaussian linear with some covariates adjusted. In addition, as Figure 3.5 show, HTE-TC performs well in controlling FDR and achieving high enough power, even when it mis-specifies the true covariate-adjusted Cox model.

3.5 Real Data

For HIV-1 vaccine or mAb prevention efficacy trials conducted by the HVTN, a common secondary or exploratory objective is assessment of baseline factors as potential modifiers of the level of vaccine/monoclonal antibody efficacy against HIV-1 infection. These baseline factors may include high-dimensional features, such as readouts from transcriptional arrays, microbiome interrogations, or host genetics data. Accordingly, it is of interest to have available statistical methods for assessing a set of baseline covariates as modifiers of vaccine efficacy.

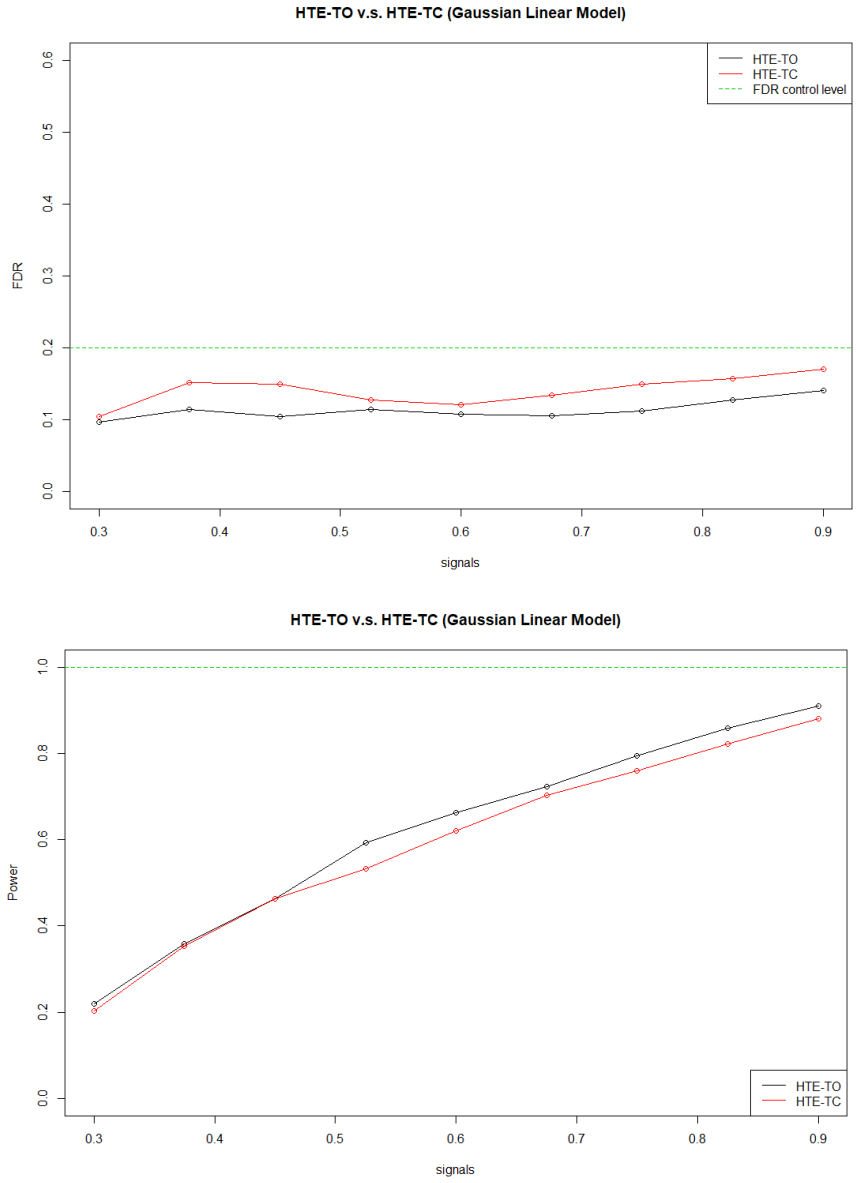

Figure 3.1: The plots of FDR control and Power by HTE-TO and HTE-TC in a Gaussian linear regression without including main effects, with $n = 400$, $p = 200$, and $\rho = 0.6$.

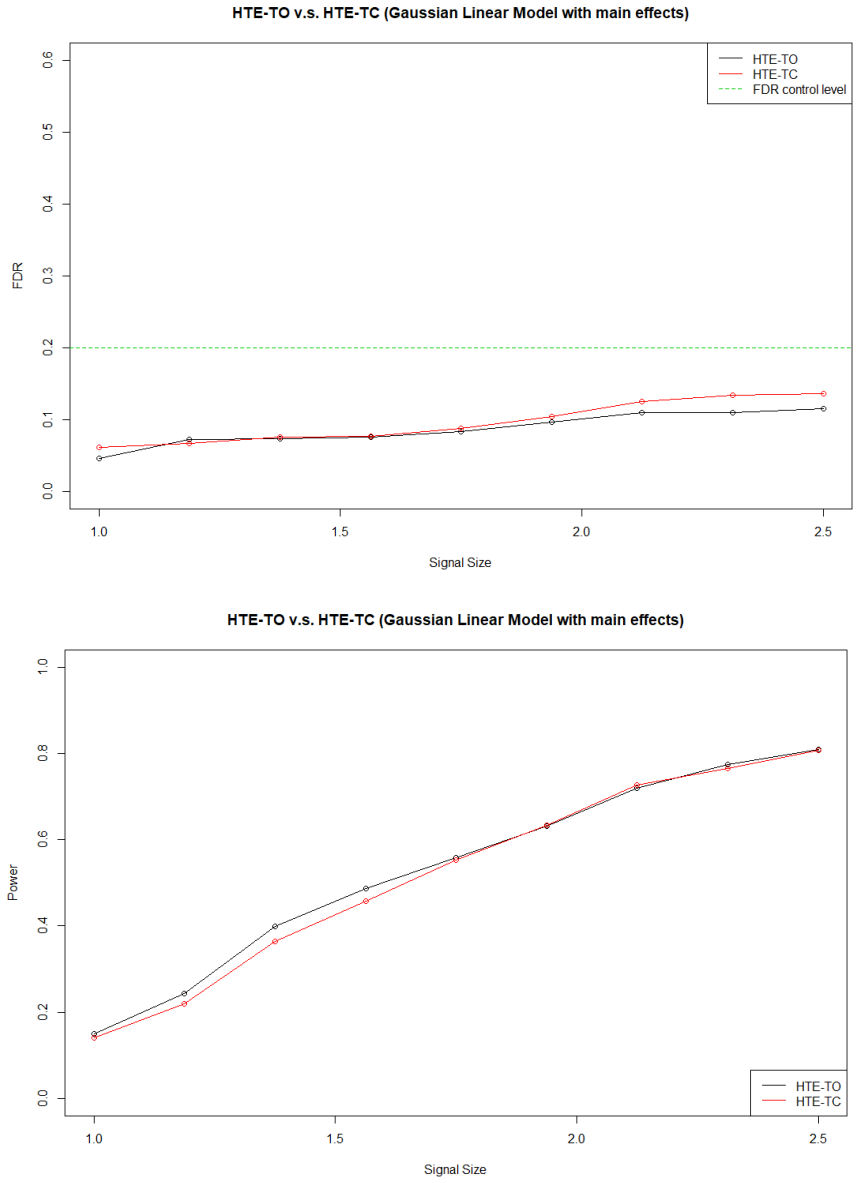

Figure 3.2: The plots of FDR control and Power by HTE-TO and HTE-TC in a Gaussian linear regression including main effects, with $n = 200$, $p = 100$, and $\rho = 0.6$.

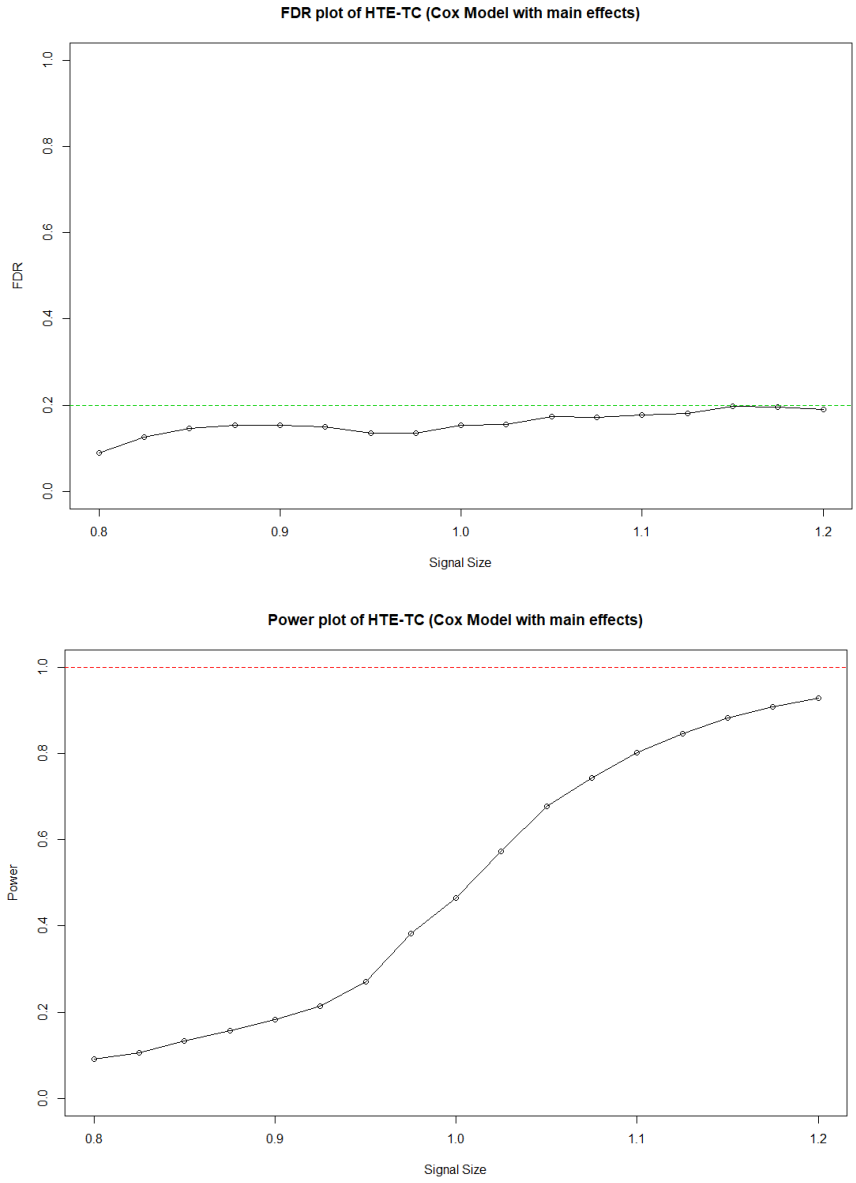

Figure 3.3: The plots of FDR control and Power by HTE-TC in a Cox model including main effects, with $n = 400$, $p = 200$, and $\rho = 0.3$.

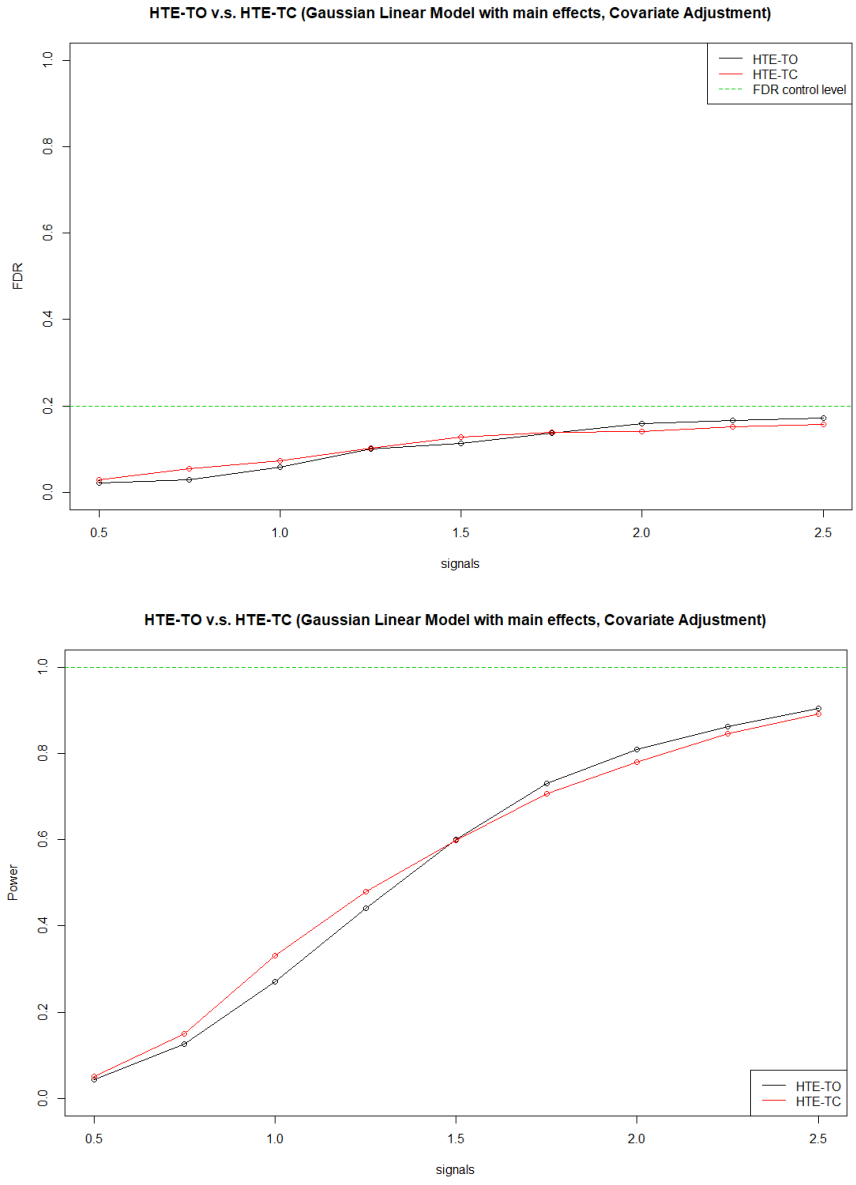

Figure 3.4: The plots of FDR control and Power by HTE-TO and HTE-TC in a Gaussian linear regression including main effects and covariate adjustment, with $n = 200$, $p = 100$, and $\rho = 0.6$.

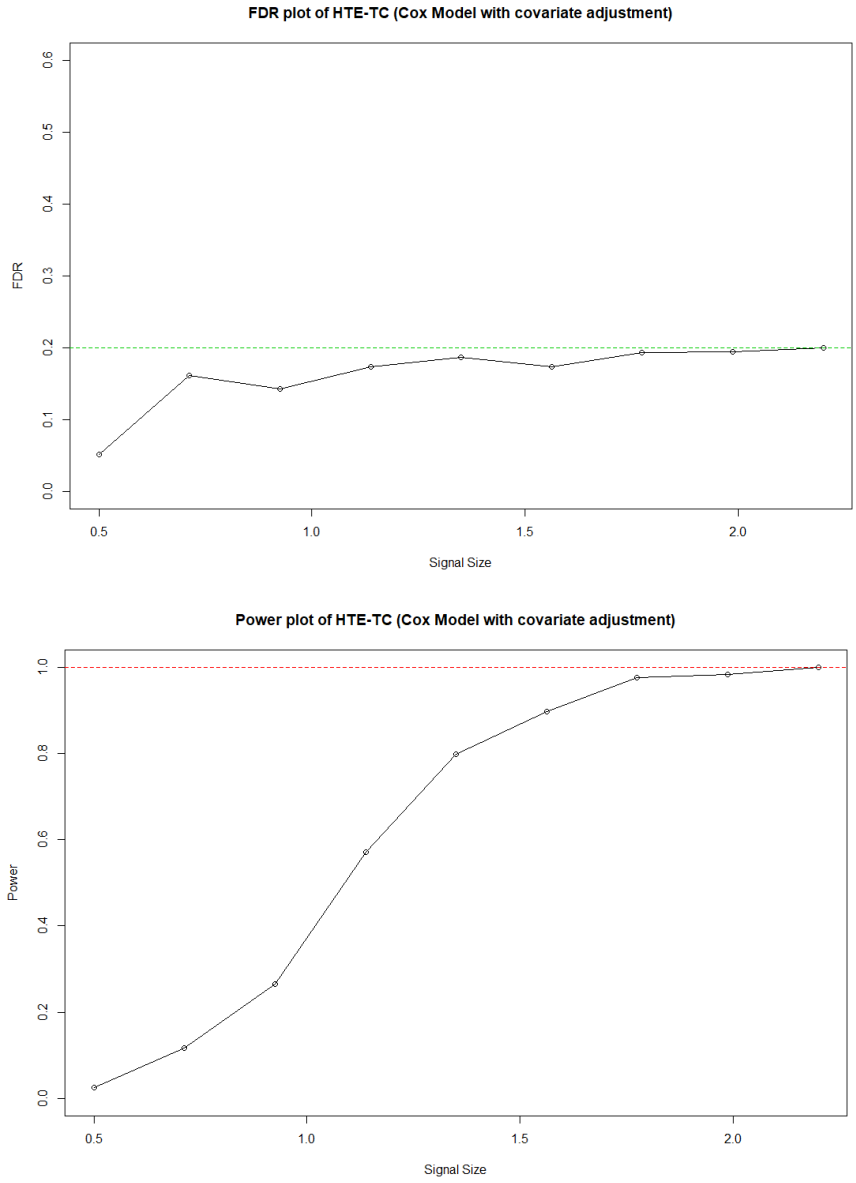

Figure 3.5: The plots of FDR control and Power by HTE-TC in a Cox model including main effects and covariate adjustment, with $n = 400$, $p = 200$, and $\rho = 0.3$

We apply our methods to the HVTN 502 Step HIV-1 vaccine efficacy trial. Previous analyses supported that baseline circumcision status and baseline Ad5 serostatus significantly modified vaccine efficacy in men (Buchbinder et al. [7]; Duerr et al. [14]; Huang et al. [22]). These results may serve as some ground truth for validating our methods in this real data application.

Suggested by researchers in HIV-1 vaccine efficacy trial, the following baseline covariates are assessed as potential effect modifiers of the hazard ratio: circumcision status (yes or no), Ad5 serostatus ($1 = AD5 > 18$, 0 otherwise), baseline Ad5 serotiter, region (North America + Australia or other), race (white or other), age (≤ 30 or > 30), HSV-2 serostatus (positive or negative), participant adjudicated (yes or no), risk behavior in the past 6 months: recreational drug usage (yes or no), unprotected receptive anal sex with male partners (yes or no), unprotected insertive anal sex with HIV+ male partner (yes or no), number of male partners (≤ 4 or > 4). In total there are 12 variables in consideration, with 1748 complete observations.

We consider a Cox survival model as the working model and conduct HTE-TC method on this dataset. Specifically, the time-to-event variable is defined as the first vaccination date to the estimated time of HIV-1 infection (midpoint between date of last RNA negative visit and the date of first evidence of infection). Then we use second-order approximation construction to get the knockoffs for those 12 variables. We choose a target FDR level to be 0.3, such that we expect 70% of the variables selected from our method to be true discoveries on average. Table 3.1 summarizes the selection frequencies of some of the variables over 100 replicates. In addition to the circumcision status that has previously been found to be a modifier of vaccine efficacy, our method selects race, age, and participant adjudicated status with higher or equal frequencies than the circumcision status and baseline Ad5 serostatus variables. This indicates that some of these variables may modify the vaccine efficacy and they may be worth for further investigation.

Variable Names	Selection by HTE-TC
Circumcision status	84%
Race	84%
Age	84%
Participant adjudicated	84%
HSV-2 serostatus	78%
Region	69%
Ad5 serostatus	62%
Unprotected receptive anal sex with male partners	60%

Table 3.1: The table of selection frequencies by HTE-TC for variables being selected at least 60 out of 100 times.

3.6 Conclusion

In this Chapter, we propose the HTE-TO and HTE-TC methods for detecting variables contributing to the heterogeneity in treatment effects. Our methods are based on the Knockoff procedure and its variants (Barber and Candès [3], Candès et al. [9], Fan et al. [16], Xie and Chan [49]) combined with transformed outcome approach (Athey and Imbens [2]) and transformed covariate models (Tian et al. [40]). The use of Knockoff methods help dealing with the false discovery issues that are common to heterogeneous treatment effect detection models. We provide a simple way to construct knockoffs for transformed covariates in HTE-TC methods, which allows us to apply existing Knockoff methods off-the-shelf without modifying their procedures.

We show in our simulations that both proposed methods, when applicable, have good power for detection and in the meantime control FDR under a target level and attain good detection power. While the HTE-TO method has similar performance as the HTE-TC method

in terms of power in Gaussian linear models, HTE-TC has wider application, particularly, as it can also be used to detect heterogeneous variables in Cox models with survival outcomes. Therefore, we would prefer HTE-TC as a more generalized heterogeneous treatment effect detection method.

Chapter 4

An Interpretable Model for Microbioal Composition Estimation from Sparse Count Data

4.1 Introduction

The human microbiome is the totality of all DNA content of microbes inhabiting bodies. Studying the variability in microbiome has been a major topic in microbiome research as it may reveal the association between the microbiome and many complex diseases (Turnbaugh et al. [43], Lewis et al. [25], Lloyd-Price et al. [28]). With the advancement of technology, DNA sequencing has helped researchers to quantify the human microbiome by obtaining counts of sequencing reads for bacterial taxa. Bacterial composition is estimated based on these sequencing read counts in microbiome studies. The count data may possess sparsity (i.e. many zero counts) due to the issue that many rare bacterial taxa are not be captured in the sequencing reads. These zero counts are not truly zeros, thus, simply using count normalization results in many zero proportions which are inaccurate estimates of bacterial abundances.

A handful of work have been proposed to deal with such sparsity issue caused by under-sampling or dropouts in count data. In the compositional data analysis, it is common to replace the zeros by some other positive values. One choice is to replace zeros by the half of the minimum non-zero values (Aitchison [1], Lin et al. [26], Shi et al. [39]). As one of the most recent proposals in compositional matrix estimation, Cao et al. [10] develops a nuclear norm regularized maximum likelihood estimator of the compositional matrix and shows that it is near optimal in Frobenius norm error and average Kullback-Leibler divergence in theory and outperforms the commonly used zero-replacement estimator in various simulation settings. Their work can be regarded as a variant of low-rank Poisson matrix recovery, as they assume low-rank or approximately low-rank structure on compositional matrix and assume that the sparse counts are generated from a Poisson-multinomial model. Most of other existing compositional matrix estimation methods focus on the estimation part, and few pay attention to the interpretability of estimators.

The low-rank structure assumption on compositional matrix is reasonable since it is indicated by recent observations on co-occurrence pattern and symbiotic relationships in microbial communities (Faust et al. [18], Chaffron et al. [11]). Motivated by this low-rank structure assumption on compositional matrix, we propose in this chapter to get an estimator of the compositional matrix by using non-negative matrix factorization (NMF) methods, which naturally results in low-rank matrix estimation. In addition to the potential improvement in estimation accuracy, the other important part of using NMF methods to study compositional matrix estimation is to increase the interpretability of the model. It allows us to consider many important modeling such as clustering while estimating the compositions, which enhances the interprebility in contrast to the aforementioned compositional data analysis proposals.

The rest of this chapter is organized as follows. In Section 4.2, we introduce notations and formulate the NMF problems for estimating compositional matrix. We also present algorithms for solving our NMF problems based on some previous work in study of constrained NMF. In addition, we establish theoretical properties which studies an upper bound of ex-

pected reconstruction error. We present simulation study results in Section 4.3, and a real data application is shown in Section 4.4. We close with a discussion in Section 4.5.

4.2 Non-negative Matrix Factorization Estimator

4.2.1 Motivations

In microbiome studies, the sequencing read data is usually summarized as a matrix $\mathbf{W} = [w_{ij}]$ for $1 \leq i \leq n, 1 \leq j \leq p$, where w_{ij} is the read count of j-th taxon of i-th individual. Define $N_i = \sum_{j=1}^{p} w_{ij}$, the total counts for i-th individual. Given N_i, we model the stratified count data over all taxa as a multinomial distribution: $w_{i1}, \ldots, w_{ip} | N_i \sim Mult(N_i; p_{i1}, \ldots, p_{ip})$, where $\mathbf{P} = [p_{ij}]$ for $1 \leq i \leq n$ and $1 \leq j \leq p$ is the unknown taxon composition matrix satisfying $\mathbf{P}\mathbf{1}_p = \mathbf{1}_n$ and $p_{ij} \geq 0 \ \forall i, j$. Unlike in Cao et al. [10], we do not impose Poisson assumption on the total counts N_i, which allows more flexibility in modeling.

A naive approach for estimating $\mathbf{P}$ is to simply normalize $\mathbf{W}$. However, this may lead to excessive zeros due to the sparsity of $\mathbf{W}$ caused by dropouts or under-sampling. We aim to estimate $\mathbf{P} \in \mathbb{R}^{n \times p}$ based on observed count matrix $\mathbf{W} \in \mathbb{R}^{n \times p}$ using non-negative matrix factorizations. Following the terminology in NMF studies, we estimate $\mathbf{P} \in \mathbb{R}^{n \times p}$ by a product of a representation matrix $\mathbf{H} \in \mathbb{R}^{n \times r}$ and a base matrix $\mathbf{X} \in \mathbb{R}^{r \times p}$, where $r < \min\{n, p\}$, and both $\mathbf{H}$ and $\mathbf{X}$ are low-rank non-negative matrices. The benefits of using NMF are twofold: not only zero proportions in estimated $\mathbf{P}$ can be avoided by putting some lower bound constraint in NMF problem, but also NMF estimators of compositional matrix are naturally low-ranked. Moreover, the intrinsic low-rank property of NMF estimator can provide concise and elegant interpretations of the compositional estimation results. For

example, if we get

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

and

$$\mathbf{X} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \end{bmatrix},$$

it indicates that these $n = 6$ patients are from $r = 3$ groups which are exclusive to one another in terms of the bacterial diversity for $p = 5$ taxa. This is a type of cluster model allowed by NMF. In this case $\mathbf{H}$ serves as a weight matrix for $r = 3$ clusters, and $\mathbf{X}$ serves as a row-dimension reduction version of the composition matrix $\mathbf{P}$ where each row of $\mathbf{X}$ is a vector of bacterial composition probabilities. In real life, $\mathbf{H}$ may not be exactly binary as clusters are not likely to be totally independent to one another, and the compositions always vary across individuals even when the patients are in a same group. In such case, $\mathbf{H}$ still provides information about which row of probabilities in $\mathbf{X}$ dominates the bacterial composition for each patient.

One way to estimate $\mathbf{P}$ using NMF is to first estimate $\mathbb{E}[\mathbf{W}]$ by a product of $\mathbf{H}$ and $\mathbf{X}$ and then estimate $\mathbf{P}$ as $\hat{\mathbf{P}}$ such that $\hat{\mathbf{P}}_{ij} = \frac{\mathbf{H}_{i\cdot}\mathbf{X}_{\cdot j}}{\sum_{j=1}^{r}\mathbf{H}_{ij}}$. Another way is to first normalize $\mathbf{W}$ and then directly apply NMF on the normalized $\mathbf{W}$. In rest of the paper, we focus on the later one and use $\mathbf{W}$ to refer the normalized version since we do not require distributional assumption on the total counts N_i and the observations with large N_i may be up-weighted if not normalized.

4.2.2 Proposal 1: Constrained Euclidean Distance NMF

Given a normalized non-negative matrix $\mathbf{W}$, we propose to get an estimator of compositional matrix $\mathbf{P}$ by solving

$$\begin{aligned} \min_{\mathbf{H}\in\mathbb{R}^{n\times r},\mathbf{X}\in\mathbb{R}^{r\times p}} & \sum_{i=1}^{n} ||\mathbf{W}_{i\cdot} - \mathbf{H}_{i\cdot}\mathbf{X}||_2^2 \\ s.t.\ & \mathbf{H}, \mathbf{X} \geq \mathbf{0}, \\ & \mathbf{H}\mathbf{1}_r = \mathbf{1}_n \\ & \mathbf{X}\mathbf{1}_p = \mathbf{1}_r, \end{aligned} \tag{4.1}$$

where $\sum_{j=1}^{p} w_{ij} = 1$ after normalization.

The two constraints in 4.1 serve for two purposes. One is to make sure that the row-sum of the representation $\mathbf{H}$ is 1, so that we can interpret the NMF estimate for each individual's composition as a weighted mixture of the probability bases in $\mathbf{X}$. Another is to enforce the row-sum of the row-dimensional reduction version of the compositional matrix, i.e. $\mathbf{X}$, to be 1, such that each row of $\mathbf{X}$ is a vector of valid probabilities for bacterial composition. In addition, the constraints in 4.1 imply $\sum_{j=1}^{p} (\mathbf{HX})_{ij} = 1$ for all i, which ensures the row-sum of estimated compositional matrix to be 1, thus indicating $\mathbf{HX}$ a valid estimator of the compositional matrix.

Theoretical Analysis

Following the terminology of NMF studies, for fixed bases $\mathbf{X}$, the representation $\mathbf{H}$ is determined by a convex problem. Therefore, in order to analyze the performance of the NMF problem in (4.1), we can choose to study the choices of bases $\mathbf{X}$. Motivated by statistical learning theory in NMF (Poggio and Shelton [33], Vapnik [44], Liu and Tao [27]), we first define the expected reconstruction error for our NMF problems and then derive upper bounds for this error term.

In our constrained Euclidean distance NMF formulation, for any row dimension reduced compositional matrix $\mathbf{X} \in \mathbb{R}^{r\times p}$, we define the reconstruction error of an individual sample

w as:

$$f_w(\mathbf{X}) = \min_{h \in \mathbb{R}_+^r, \sum_j h_j = 1} \left|\left| w - h^T \mathbf{X} \right|\right|_2^2 . \tag{4.2}$$

The expected reconstruction error of the compositional matrix $\mathbf{X}$ is $F(\mathbf{X}) = \mathbb{E}_{\mathbf{W}}\left[f_w(\mathbf{X})\right]$. And the empirical reconstruction error is $F_n(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n f_{w_i}(\mathbf{X})$. In practice, we do not know the true sub-compositional matrix $\mathbf{X}_{target} = \arg\min_{\mathbf{X} \in \mathbb{R}^{r \times p}} F(\mathbf{X})$, but we can get the learned compositional matrix $\hat{\mathbf{X}} = \arg\min_{\mathbf{X} \in \mathbb{R}^{r \times p}} F_n(\mathbf{X})$ from the NMF. So our goal is to analyze the expected reconstruction error of the learned $\hat{\mathbf{X}}$. One strategy in literature for analyzing such expected reconstruction error is to decompose it into the estimation error and approximation error [27]:

$$F(\hat{\mathbf{X}}) = \Big[F(\hat{\mathbf{X}}) - F(\mathbf{X}_{target})\Big] + F(\mathbf{X}_{target}), \tag{4.3}$$

where the first part in bracket is the estimation error and the rest is the approximation error.

We first analyze the estimation error. With some algebra (see the details in Liu and Tao [27]), we can upper bound the estimation error by the generalization error $\sup_{\mathbf{X} \in \mathbb{R}^{r \times p}} |F(\mathbf{X}) - F_n(\mathbf{X})|$ using the inequality

$$F(\hat{\mathbf{X}}) - F(\mathbf{X}_{target}) \leq 2 \sup_{\mathbf{X} \in \mathbb{R}^{r \times p}} |F(\mathbf{X}) - F_n(\mathbf{X})| . \tag{4.4}$$

Therefore, if we can get an upper bound of the right-hand side in (4.4), then we have an upper bound for $F(\hat{\mathbf{X}}) - F(\mathbf{X}_{target})$. Motivated by this observation, we derive the following theorem.

Theorem 4. *Consider the NMF problem in (4.1). For any $\mathbf{X} \in \mathbb{R}^{r \times p}$ and any $\epsilon > 0$, with probability at least $1 - \epsilon$, we have*

$$|F(\mathbf{X}) - F_n(\mathbf{X})| \leq \frac{2\sqrt{2\pi} p r^{\frac{3}{2}}}{\sqrt{n}} + p\sqrt{\frac{\log(\frac{1}{\epsilon})}{2n}} \tag{4.5}$$

.

Theorem 4 helps us to get an upper bound for the estimation error. This bound result is asymptotic and it is dimensionality-dependent as the right-hand-side (RHS) term in (4.5)

involves p. This is not desirable for high dimensional setting with $p >> n$, however, in typical microbiome studies the dimensionality p is usually not very large. And we show in our simulations that our proposals are also competitive to some existing methods such zero-replacement estimator in high dimensional setting across various simulation setups.

Now we provide another theorem for the approximation error which gives an asymptotic approximation error bound. This theorem directly follows the Theorem 11 in Liu and Tao [27] so we omit the proof.

Theorem 5. *For the NMF problem in* (4.1)*, when the reduced dimensionality* $r \to p$*, we have*

$$F(\mathbf{X}_{target}) \leq \mathbf{O}\left(r^{-\frac{2}{p}}\right). \tag{4.6}$$

Theorem 4 and 5 together provide an upper bound of the estimation error $F(\hat{\mathbf{X}})$.

Algorithm

We propose to solve the constrained optimization (4.1) by using an iterative algorithm based on the method developed in Peng et al. [32]. We first define the following terms.

$$\begin{aligned} C^{(1)}(\mathbf{H}, \mathbf{X}) &= \mathbf{H}\mathbf{1}_r - \mathbf{1}_n \\ C^{(2)}(\mathbf{H}, \mathbf{X}) &= \mathbf{X}\mathbf{1}_p - \mathbf{1}_r \\ J(\mathbf{H}, \mathbf{X}) &= \sum_{i=1}^{n} ||\mathbf{W}_{i\cdot} - \mathbf{H}_{i\cdot}\mathbf{X}||_2^2 . \end{aligned} \tag{4.7}$$

Then taking the derivatives of them with respect to $\mathbf{H}$ and $\mathbf{X}$, we get

$$\begin{aligned} \nabla_{\mathbf{H}} \left|\left|C^{(1)}\right|\right|_2^2 &= \left(\mathbf{H}\mathbf{1}_r - \mathbf{1}_n\right)\mathbf{1}_r^T \\ \nabla_{\mathbf{X}} \left|\left|C^{(2)}\right|\right|_2^2 &= \left(\mathbf{X}\mathbf{1}_p - \mathbf{1}_r\right)\mathbf{1}_p^T \\ \nabla_{\mathbf{H}} ||J(\mathbf{H}, \mathbf{X})||_2^2 &= \left(\mathbf{H}\mathbf{X} - \mathbf{W}\right)\mathbf{X}^T \\ \nabla_{\mathbf{X}} ||J(\mathbf{H}, \mathbf{X})||_2^2 &= \mathbf{H}^T\left(\mathbf{H}\mathbf{X} - \mathbf{W}\right). \end{aligned} \tag{4.8}$$

The iterative algorithm is as following:

1. Initialize $\mathbf{H}$ and $\mathbf{X}$ as $\mathbf{H}^{(0)}$ and $\mathbf{X}^{(0)}$.

2. For the $(k+1)$-th iteration, update

   - $\mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} \times \frac{\mathbf{W}\left(\mathbf{X}^{(k)}\right)^T + \lambda_1 \mathbf{1}_n \mathbf{1}_r^T}{\mathbf{H}^{(k)}\mathbf{X}^{(k)}\left(\mathbf{X}^{(k)}\right)^T + \lambda_2 \mathbf{H}^{(k)} \mathbf{1}_r \mathbf{1}_r^T}$,
   - $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} \times \frac{\left(\mathbf{H}^{(k+1)}\right)^T \mathbf{W} + \lambda_3 \mathbf{1}_r \mathbf{1}_p^T}{\left(\mathbf{H}^{(k+1)}\right)^T \mathbf{H}^{(k+1)}\mathbf{X}^{(k)} + \lambda_4 \mathbf{X}^{(k+1)} \mathbf{1}_p \mathbf{1}_p^T}$,

   where '$\times$' and '$+$' are meant to be element-wise operators, and λ_i's are tuning parameters.

3. Repeat the procedure until convergence.

4.2.3 Proposal 2: Constrained Hellinger Distance NMF

In addition to the Euclidean distance NMF, We consider a Hellinger distance NMF problem, for which we expect to be more robust to outliers than the Euclidean distance NMF. In this case, we also first conduct normalization so that $\mathbf{W} \in [0,1]^{n \times p}$. Then we consider solving the optimization problem

$$\begin{aligned} \min_{\mathbf{H} \in \mathbb{R}^{n \times r}, \mathbf{X} \in \mathbb{R}^{r \times p}} & \sum_{i=1}^{n} \left|\left| \sqrt{\mathbf{W}_{i\cdot}} - \sqrt{\mathbf{H}_{i\cdot}\mathbf{X}} \right|\right|_2^2 \\ s.t.\ & \mathbf{H}, \mathbf{X} \geq \mathbf{0}, \\ & \mathbf{H}\mathbf{1}_r = \mathbf{1}_n \\ & \mathbf{X}\mathbf{1}_p = \mathbf{1}_r \\ & \mathbf{X}_{kj} \geq a \end{aligned} \tag{4.9}$$

where $\sum_{j=1}^{p} \mathbf{W}_{ij} = 1$ after normalization, and a is some pre-specified constant that controls the lower bound of entries in estimated $\mathbf{X}$. It is reasonable to have such lower bound because the derivative of the objective function can be very large if there are too small entries. We do not put constraint on the lower bound of entries in $\mathbf{H}$ because the sparsity in $\mathbf{H}$ would allow

important modeling such as clustering, for example the $\mathbf{H}$ in Section 4.2.1. Furthermore, the square root term in Helling distance may help in estimating matrix when there are extreme values (e.g. very close to zeros) involved.

Theoretical Analysis

Similar to the previous analysis for Euclidean distance NMF, we study the choices of $\mathbf{X}$ to evaluate the performance of the NMF problem in (4.9). For any sub-compositional matrix $\mathbf{X} \in \mathbb{R}^{r\times p}$, we define the reconstruction error of an individual sample w as:

$$f_w^{HL}(\mathbf{X}) = \min_{h\in\mathbb{R}_+^r, \sum_{j=1}^p h_j=1, h_j\geq a} \left|\left|\sqrt{w} - \sqrt{h^T\mathbf{X}}\right|\right|_2^2. \tag{4.10}$$

We denote the expected reconstruction error of the compositional matrix $\mathbf{X}$ to be $F^{HL}(\mathbf{X}) = \mathbb{E}_{\mathbf{W}}\left[f_w^{HL}(\mathbf{X})\right]$. And the empirical reconstruction error is defined to be $F_n^{HL}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n f_{w_i}^{HL}(\mathbf{X})$. We can derive a upper bound of $\left|F^{HL}(\mathbf{X}) - F_n^{HL}(\mathbf{X})\right|$ using the result in (4.11).

Theorem 6. *Consider the NMF problem in (4.9). For any $\mathbf{X} \in \mathbb{R}^{r\times p}$ satisfying the constraints in 4.9 and any $\epsilon > 0$, with probability at least $1-\epsilon$, we have*

$$\left|F^{HL}(\mathbf{X}) - F_n^{HL}(\mathbf{X})\right| \leq \frac{\sqrt{2\pi}r^{\frac{3}{2}}p}{n\sqrt{a}} + p\sqrt{\frac{\log(\frac{1}{\epsilon})}{2n}}. \tag{4.11}$$

.

The result is quite similar to the one in Theorem 4 except that there is a in the denominator of the first term in RHS of (4.11). This implies that the choice of a should not be too small.

Algorithm

We again propose to solve the constrained optimization (4.9) by using an iterative algorithm based on the method developed in Peng et al. [32]. First, we note that the Hellinger distance can be expressed in another form i.e.

$$\sum_{i=1}^n \left|\left|\sqrt{\mathbf{W}_{i\cdot}} - \sqrt{\mathbf{H}_{i\cdot}\mathbf{X}}\right|\right|_2^2 = 2 - 2\sum_{i=1}^n\sum_{j=1}^p \left(\sqrt{\mathbf{W}_{ij}\left(\mathbf{HX}\right)_{ij}}\right).$$

We first define the following terms.

$$
\begin{aligned}
C^{(1)}(\mathbf{H},\mathbf{X}) &= \mathbf{H}\mathbf{1}_r - \mathbf{1}_n \\
C^{(2)}(\mathbf{H},\mathbf{X}) &= \mathbf{X}\mathbf{1}_p - \mathbf{1}_r \\
C^{(3)}(\mathbf{H},\mathbf{X}) &= \mathbf{X} - b + \mathbf{S}^{(3)} \\
J(\mathbf{H},\mathbf{X}) &= 2 - 2\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\sqrt{\mathbf{W}_{ij}(\mathbf{H}\mathbf{X})_{ij}}\right). \qquad (4.12)
\end{aligned}
$$

Then taking the derivatives of them with respect to $\mathbf{H}$ and $\mathbf{X}$, we get

$$
\begin{aligned}
\nabla_{\mathbf{H}}\left|\left|C^{(1)}\right|\right|_2^2 &= (\mathbf{H}\mathbf{1}_r - \mathbf{1}_n)\,\mathbf{1}_r^T \\
\nabla_{\mathbf{X}}\left|\left|C^{(2)}\right|\right|_2^2 &= (\mathbf{X}\mathbf{1}_p - \mathbf{1}_r)\,\mathbf{1}_p^T \\
\nabla_{\mathbf{X}}\left|\left|C^{(3)}\right|\right|_2^2 &= \mathbf{X} - b + \mathbf{S}^{(3)} \\
\left(\nabla_{\mathbf{H}} J(\mathbf{H},\mathbf{X})\right)_{ik} &= \sum_{j=1}^{p}\frac{-\sqrt{\mathbf{W}_{ij}}\mathbf{X}_{kj}}{(\mathbf{H}\mathbf{X})_{ij}} \\
\left(\nabla_{\mathbf{X}} J(\mathbf{H},\mathbf{X})\right)_{kj} &= \sum_{i=1}^{n}\frac{-\sqrt{\mathbf{W}_{ij}}\mathbf{H}_{ik}}{(\mathbf{H}\mathbf{X})_{ij}}. \qquad (4.13)
\end{aligned}
$$

The iterative algorithm is as following:

1. Initialize $\mathbf{H}$ and $\mathbf{X}$ as $\mathbf{H}^{(0)}$ and $\mathbf{X}^{(0)}$.

2. For the $(k+1)$-th iteration, update

   - $\mathbf{H}^{(k+1)} = \mathbf{H}^{(k)} \times \frac{\nabla_{\mathbf{H}} J(\mathbf{H},\mathbf{X})^- + \lambda_1 \mathbf{1}_n \mathbf{1}_r^T}{\nabla_{\mathbf{H}} J(\mathbf{H},\mathbf{X})^+ + \lambda_2 \mathbf{H}^{(k)} \mathbf{1}_r \mathbf{1}_r^T}$,
   - $\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} \times \frac{\nabla_{\mathbf{X}} J(\mathbf{H},\mathbf{X})^- + \lambda_3 \mathbf{1}_r \mathbf{1}_p^T}{\nabla_{\mathbf{X}} J(\mathbf{H},\mathbf{X})^+ + \lambda_4 \mathbf{X}^{(k+1)} \mathbf{1}_p \mathbf{1}_p^T + \lambda_5\left(\mathbf{X} - b + \mathbf{S}^{(3)}\right)}$,

   where '$\times$' and '$+$' are meant to be element-wise operators, $(\cdot)^+$ is the positive part of $(\cdot)$, $(\cdot)^-$ is the negative part of $(\cdot)$, and λ_i's are tuning parameters.

3. Repeat the procedure until convergence.

4.3 Simulation

In this section, we compare the performance of our NMF estimators $\widehat{\mathbf{P}}_{Frob}, \widehat{\mathbf{P}}_{HL}$ with the nuclear norm regularized maximum likelihood estimator $\widehat{\mathbf{P}}_{pg}$ proposed in Cao et al. [10] and the standard zero replacement estimator $\widehat{\mathbf{P}}_{zr}$. We use the R code attached in Cao et al. [10] to get $\widehat{\mathbf{P}}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$. The choice of the lower bound a in Hellinger distance NMF can be selected based on some data-driven procedure. The estimated rank r in both NMF methods by evaluating the stability of clustering associated with a given rank r. In order to do so, we use the cophenetic correlation coefficient [6] to select the value of r, which has been widely applied in the field of biostatistics for use as a test for nested clusters.

We consider several setups in simulation studies. The first setup is a Poisson-Multinomial model, which is considered in Cao et al. [10] as a way to simulate correlated compositional data arising from metagenomics. The second setup is a family of cluster models, in which the true compositional matrix is generated as a product of a weighted matrix and a row-dimension reduction version of the compositional matrix. In both setups, the true compositional matrix has low-rank structure. The third setup is to compare the two NMF estimators, $\widehat{\mathbf{P}}_{Frob}$ and $\widehat{\mathbf{P}}_{HL}$, in order to see which one is more robust to extreme values or outliers.

4.3.1 Setup 1: Poisson-Multinomial Modeling

We follow a similar simulation setup in Cao et al. [10] here. For i-th individual, we let the total count $N_i \sim \text{Poisson}(\nu_i)$, where $\nu_i \sim \text{Uniform}[\gamma p, 10\gamma p]$ to reflect the heterogeneity of total count across n individuals, where $\gamma = 1, 2, 3, 4$. Let $U \in \mathbb{R}^{n\times r}$ be the absolute values of an independent and identically distributed standard normal matrix. Let $V = V_1 + V_2 \in \mathbb{R}^{p\times r}$ be a random spike matrix, where

$$(V_1)_{ij} = \begin{cases} 1, & i = j; \\ 1, & i \neq j \text{ with probability } 0{\cdot}3; \\ 0, & i \neq j \text{ with probability } 0{\cdot}7, \end{cases} \qquad (V_2)_{ij} \sim N(0, 0.001).$$

The true composition matrix is generated as $\mathbf{P}^*_{ij} = Z_{ij}/\sum_{k=1}^{p} Z_{ik}$, where $Z = UV^\top$. We repeat the generating procedure until a positive matrix $\mathbf{P}^*$ is generated.

We set the sample size to be $n = 100$ and vary the number of taxa to be $p = 50, 100$. Since we assume the low rank structure in $\mathbf{P}^*$, we set $r = 20$. These parameters are chosen similarly as the ones in Cao et al. [10] to mimic the microbiome data in their real data analysis. The read counts W are generated from the Poisson-multinomial model, i.e. $N_i \sim \text{Pois}(\nu_i)$, $W_i \sim \text{Mult}(N_i; \mathbf{P}^*_{i\cdot})$. The estimation performances are evaluated by the average loss in squared Frobenius norm $\|\widehat{\mathbf{P}} - \mathbf{P}^*\|_F^2$, the averaged Kullback-Leibler (KL) divergence, and the averaged Hellinger distance $\sum_{i=1}^{n} \|\sqrt{\widehat{\mathbf{P}}_{i\cdot}} - \sqrt{\mathbf{P}^*_{i\cdot}}\|_2^2$. We compare our proposed NMF estimators $\widehat{\mathbf{P}}_{Frob}$ and $\widehat{\mathbf{P}}_{HL}$ with the nuclear norm regularized maximum likelihood estimator $\widehat{\mathbf{P}}_{pg}$ in Cao et al. [10] and the standard zero replacement estimator $\widehat{\mathbf{P}}_{zr}$. All the results are averaged over 100 replicates and are summarized in Table 4.1.

As the results in Table 4.1 show, $\widehat{\mathbf{P}}_{Frob}$ and $\widehat{\mathbf{P}}_{HL}$ outperform $\widehat{\mathbf{P}}_{zr}$ for all choices of γ and perform better than $\widehat{\mathbf{P}}_{pg}$ when γ is large. Our simulation results agree with the findings in Cao et al. [10] that $\widehat{\mathbf{P}}_{pg}$ is much less sensitive to the choice of γ. On the other hand, the advantage of $\widehat{\mathbf{P}}_{Frob}$ and $\widehat{\mathbf{P}}_{HL}$ against $\widehat{\mathbf{P}}_{pg}$ is more apparent as γ increases.

4.3.2 Setup 2: Clustering Modeling

In addition to the Poisson-Multinomial model, we also consider two types of clustering models. For i-th individual, we again let the total count $N_i \sim \text{Poisson}(\nu_i)$, where $\nu_i \sim \text{Uniform}[\gamma p, 10\gamma p]$ and $\gamma = 2, 3, 4$. Unlike the previous setup, we construct the true composition matrix $\mathbf{P}^*$ to reflect cluster properties among the individuals. The observed count matrix $\mathbf{W}$ is generated from a multinomial distribution based on N_i's and $\mathbf{P}^*$.

Clustering 1

In the first cluster model setup, we fix the sample size $n = 50$ and vary the number of taxa p to be $20, 50$, and 100. We divide these 100 individuals equally into $r = 10$ clusters. The

compositions for individuals within a cluster are same. Specifically, Let $\mathbf{H} \in \mathbb{R}^{50\times 10}$ be the weighted matrix, where

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 1 \end{bmatrix}.$$

And let $\mathbf{X} \in \mathbb{R}^{10\times p}$ be a random composition matrix

$$\mathbf{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{10,p} & X_{10,2} & \dots & X_{10,p} \end{bmatrix},$$

where we first generate $X_{k,j} \sim Unif(0,1)$, for $k = 1,\dots,r$ and $j = 1,\dots,p$, and then normalize $\mathbf{X}$ in each row. The true composition matrix $\mathbf{P}^*$ is generated as the product of $\mathbf{H}$ and $\mathbf{X}$.

Clustering 2

In the second cluster model setup, we keep the same setting as previous but change the weighted matrix $\mathbf{H}$ to be

$$\mathbf{H} = \begin{bmatrix} 0.55 & 0.05 & \dots & 0.05 \\ \vdots & \vdots & \vdots & \vdots \\ 0.55 & 0.05 & \dots & 0.05 \\ 0.05 & 0.55 & \dots & 0.05 \\ \vdots & \vdots & \vdots & \vdots \\ 0.05 & 0.55 & \dots & 0.05 \\ \vdots & \vdots & \vdots & \vdots \\ 0.05 & \dots & 0.05 & 0.55 \\ \vdots & \vdots & \vdots & \vdots \\ 0.05 & \dots & 0.05 & 0.55 \end{bmatrix}.$$

And we generate $\mathbf{X}$ and $\mathbf{P}^*$ in a same way as before.

In both clustering models, we see from Tables 4.2–4.7 that our NMF proposals have better performance compared to other two methods in almost all settings.

4.3.3 Setup 3: Extreme Values

We have shown in previous two setups that $\widehat{\mathbf{P}}_{Frob}$ and $\widehat{\mathbf{P}}_{HL}$ outperforms $\widehat{P}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$ in various settings. We also note that these two NMF estimators have quite similar performance in previous simulations. To see the difference between these two estimators, we adopt the second cluster model in this simulation setup, but construct some of the entries in the true compositional matrix to be extreme by first generating $\log(X_{k,j}) \sim Unif(-8, 8)$, for $k = 1, \dots, r$ and $j = 1, \dots, p$, and then normalizing $\mathbf{X}$ in each row.

Again, we set $n = 50$ and p varies over $20, 50, 100$. In order to make some sparsity in the observed count table $\mathbf{W}$, we set the total read counts to be low for some individuals. Specifically, we set $N_i = 400$ for $i = 1, \dots, 40$ and $N_i = 4000$ for $i = 41, \dots, 50$. In order

for a fair comparison, we calculated the averaged KL divergence of the two estimators. The results are summarized in Table 4.8.

We see from Table 4.8 that $\widehat{\mathbf{P}}_{HL}$ has smaller averaged KL divergence errors than $\widehat{\mathbf{P}}_{Frob}$. We expect to see this result as the Hellinger distance tends to be more robust to extreme values than the Euclidean distance.

4.4 Real Data

We apply our NMF methods on a breast milk dataset which contains 58 microbiome samples taken from lactating Caucasian Canadian women. We use non-negative double singular value decomposition (Boutsidis and Gallopoulos [4]) to initialize our NMF algorithms. The number of Operational Taxonomic Units (OTUs) is 115, and there are a total of 5.3 million reads across these OTUs. Note that there is one sample with around 2.8 million reads; this comes from a patient with an infection, and the next largest number of reads per sample is 282485, which is ten times less. Therefore, this infected sample provides extreme values in this count table.

The plot 4.1 indicates the low rank structure of the compositional matrix. Combined the plot of eigenvalues with the cophenetic correlation coefficient plot 4.2, we choose $r = 5$ in the NMF methods. Since there is an obvious outlier in this data set (the one with infection), we consider the Hellinger distance NMF method to be a more reliable choice than the Euclidean distance NMF, as Hellinger distance is more robust to outliers than Euclidean distance. Figure 4.3 shows the estimated weighted matrix $\mathbf{H}$ resulted from using the Hellinger distance NMF. We can see from the heat map in Figure 4.3 that the patients are roughly clustered into 5 groups, indexed by the row number of $\mathbf{X}$. Based on the results in Figure 4.4, the compositions of Cluster 1 are dominated by the OTU 0 (Pasteurella), the compositions of Cluster 2 are dominated by OTU 1 (Staphylococcus) and 10 (Lactobacillus), the compositions of Cluster 3 are dominated by OTU 2 (Pseudomonas), the compositions of Cluster 4 are dominated by OTU 3 (Shigella), and the compositions of Cluster 5 are

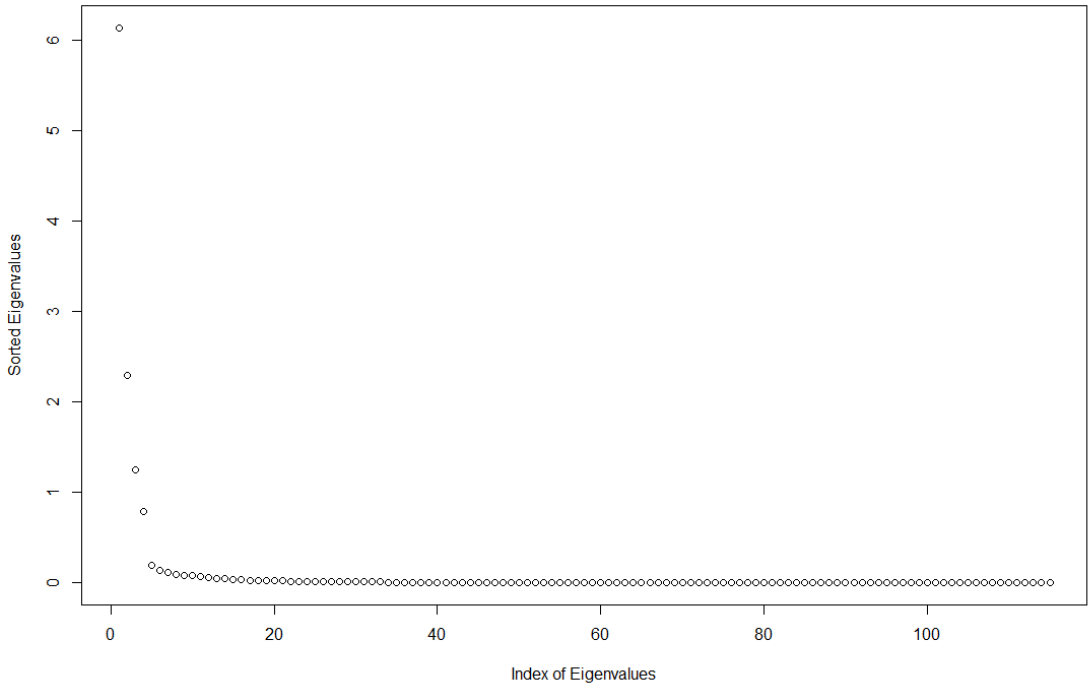

Figure 4.1: The plot of eigenvalues for $\mathbf{W}^T\mathbf{W}$, indicating that the estimated rank of $\mathbf{W}$ could be 5.

dominated by OTU 1 (Staphylococcus).

4.5 Discussion

In this chapter, we propose NMF based estimators for composition matrix estimation. We show in simulation studies that our estimators are very competitive to estimators from other existing methods in terms of Frobenius norm error, averaged KL divergence and Hellinger distance. We also illustrate the interpretability of our estimators in the real data application. The weighted matrix $\mathbf{H}$ allows us to identify potential clusters among patients, and the row dimension reduced composition matrix $\mathbf{X}$ tells which taxa have the dominant compositions

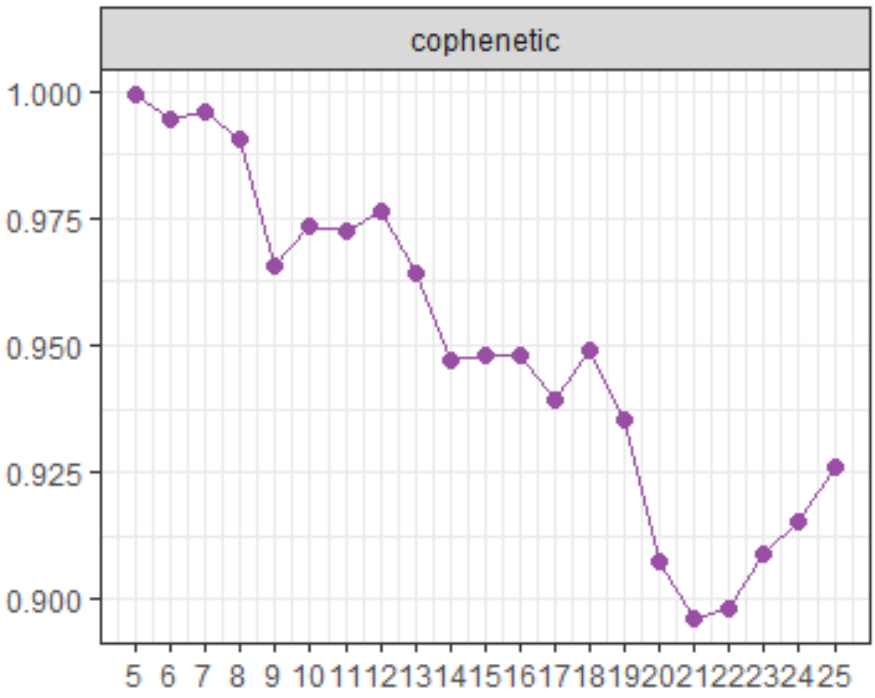

Figure 4.2: The plot of cophenetic correlation coefficients over $r = 5$ to $r = 25$ for the normalized count table $\mathbf{W}$ in the breast milk dataset.

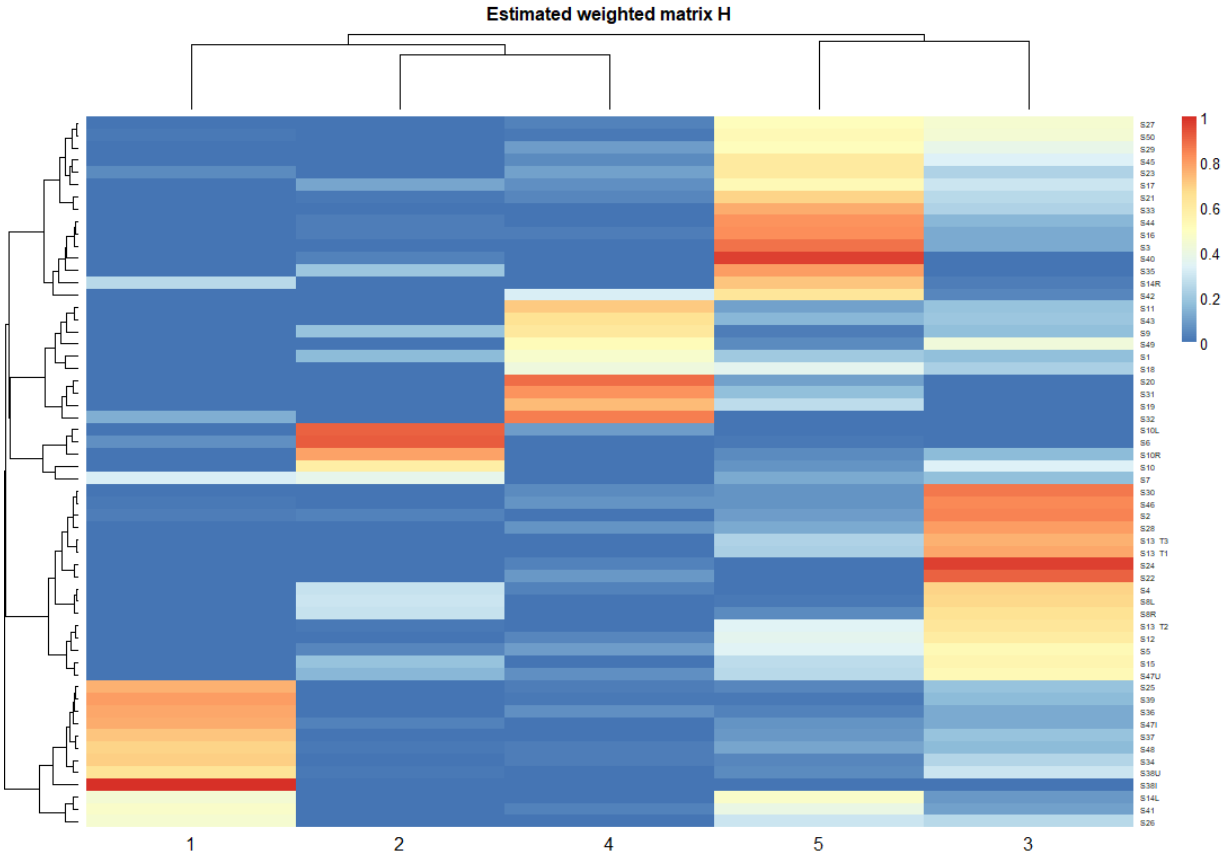

Figure 4.3: The heat map of the weighted matrix **H** resulted from applying our NMF method on the breast milk dataset.

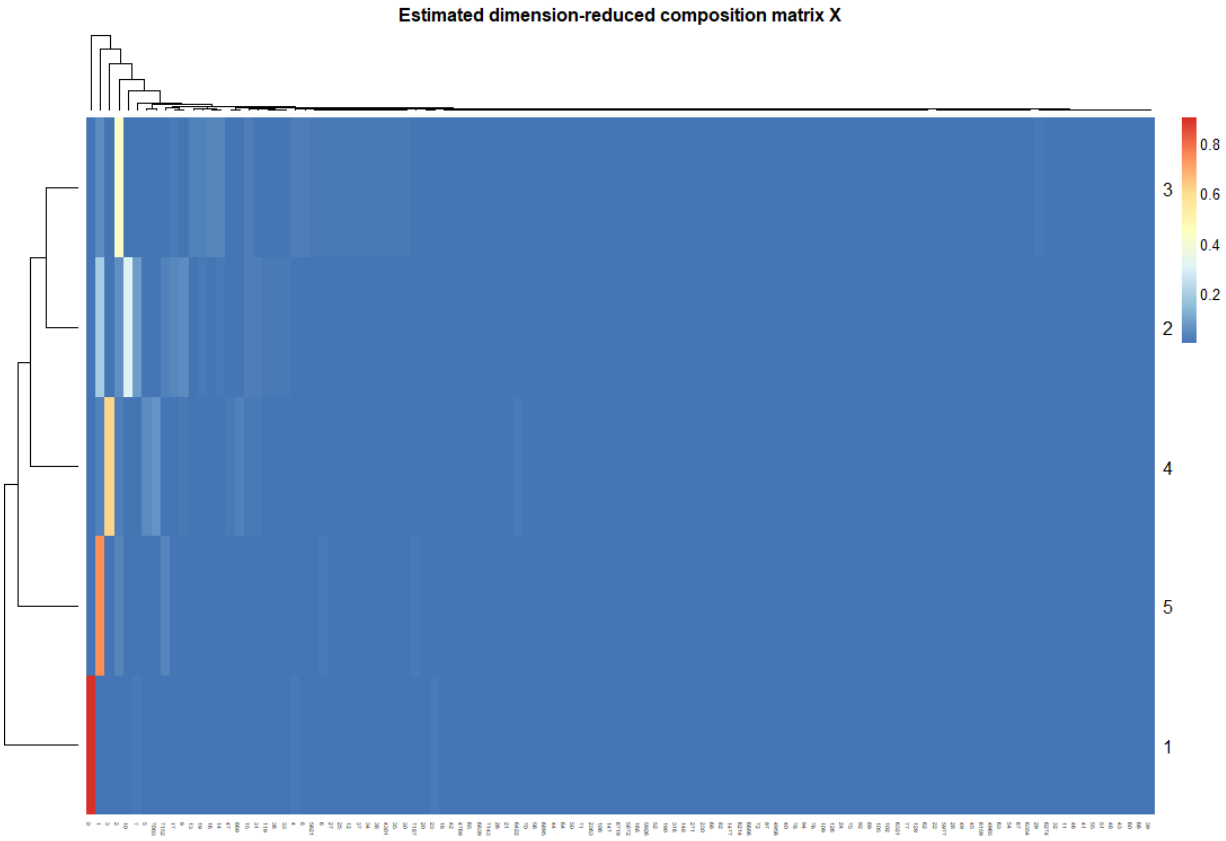

Figure 4.4: The heat map of the dimension reduced compositional matrix $\mathbf{X}$ resulted from applying our NMF method on the breast milk dataset.

of a patient.

Although the theoretical results of our estimators in this chapter are asymptotic, we show in synthetic data that our proposals perform well in high dimensional settings. It is worth pointing out that in simulations we calculate the error terms using the true composition matrix $\mathbf{P}^*$, while the theoretical analyses are based on the normalized $\mathbf{W}$ observed from count table. Therefore, the theoretical results in Cao et al. [10] are not directly comparable to ours as the upper bounds derived in our project are for estimation errors of $\mathbf{W}$ while the upper bounds derived in their paper are for estimation errors of $\mathbf{P}^*$. We focus on theoretical analysis for the constrained NMF estimator because to the best of our knowledge there is no previous literature providing theoretical results on Hellinger distance NMF estimator. Our results close this gap by providing an upper bound of estimation error for the constrained Hellinger distance NMF estimator.

	$p = 50$				$p = 100$			
γ	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$
				Squared Frobenius norm error ($\times 10^{-2}$)				
1	29.1	31.7	12.7	43.8	12.9	12.0	5.8	21.7
2	15.7	15.8	12.7	23.1	6.5	6.5	6.3	11.7
3	10.7	10.9	12.6	15.9	4.8	4.9	6.1	8.5
4	9.2	9.4	12.5	13.0	4.0	4.1	6.4	7.1
				Averaged KL divergence ($\times 10^{-3}$)				
1	77.8	81.3	33.5	120.4	73.2	61.4	30.4	119.9
2	40.6	39.6	33.9	64.8	33.7	33.6	33.8	65.3
3	27.9	27.5	33.7	43.8	25.3	25.4	32.6	47.2
4	23.7	23.9	33.9	36.1	21.0	21.1	34.1	39.5
				Squared Hellinger distance				
1	3.61	3.83	1.74	5.55	3.26	2.92	1.57	5.51
2	1.96	1.94	1.76	3.02	1.64	1.65	1.75	3.04
3	1.35	1.35	1.76	2.07	1.24	1.25	1.69	2.22
4	1.17	1.18	1.77	1.70	1.04	1.05	1.78	1.86

Table 4.1: Squared Frobenius norm error for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$, $\widehat{\mathbf{P}}_{pg}$, and $\widehat{\mathbf{P}}_{zr}$ under the low rank assumption with $r = 20$ and $n = 100$.

$p = 20$

γ	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$
Squared Frobenius norm error ($\times 10^{-2}$)				
2	21.5	24.4	75.5	30.0
3	14.5	15.8	75.1	19.8
4	11.4	13.2	75.2	15.4
Averaged KL divergence ($\times 10^{-3}$)				
2	54.4	54.5	172.8	63.8
3	39.4	37.9	171.9	42.7
4	31.1	31.6	172.1	32.8
Squared Hellinger Distance ($\times 10^{-2}$)				
2	133.5	137.5	502.2	150.2
3	99.3	99.1	500.0	101.4
4	80.9	83.2	500.4	78.5

Table 4.2: Estimation errors for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$, $\widehat{\mathbf{P}}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$ in clustering modeling 1, under the low rank assumption with $r = 10$ and $n = 50$, $p = 20$.

	$p = 50$			
γ	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$
Squared Frobenius norm error ($\times 10^{-2}$)				
2	5.3	5.2	30.7	9.4
3	5.3	5.2	30.6	8.6
4	3.8	3.8	30.6	5.5
Averaged KL divergence ($\times 10^{-3}$)				
2	36.9	35.1	178.9	49.6
3	36.6	35.2	178.1	45.2
4	26.7	26.1	178.5	29.0
Squared Hellinger Distance ($\times 10^{-2}$)				
2	99.9	94.7	524.9	117.6
3	98.9	95.5	522.5	107.3
4	73.5	71.8	523.6	69.6

Table 4.3: Estimation errors for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$, $\widehat{\mathbf{P}}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$ in clustering modeling 1, under the low rank assumption with $r = 10$ and $n = 50$, $p = 50$.

	$p = 100$			
γ	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$
Squared Frobenius norm error ($\times 10^{-2}$)				
2	1.7	1.7	14.5	5.5
3	1.6	1.5	14.6	4.3
4	0.7	0.7	15.2	2.6
Averaged KL divergence ($\times 10^{-3}$)				
2	22.1	21.6	170.7	57.9
3	20.9	20.4	171.2	45.5
4	10.0	9.7	178.3	26.8
Squared Hellinger Distance ($\times 10^{-2}$)				
2	57.9	56.4	502.5	136.9
3	55.9	54.3	504.0	107.7
4	26.8	26.0	524.3	64.5

Table 4.4: Estimation errors for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$, $\widehat{\mathbf{P}}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$ in clustering modeling 1, under the low rank assumption with $r = 10$ and $n = 50$, $p = 100$.

	$p = 20$			
γ	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$
	Squared Frobenius norm error ($\times 10^{-2}$)			
2	20.9	23.7	19.4	28.2
3	14.8	16.8	19.4	19.9
4	11.0	12.3	19.3	15.2
	Averaged KL divergence ($\times 10^{-3}$)			
2	48.1	50.9	40.0	63.3
3	34.2	36.6	40.1	44.5
4	24.6	26.3	39.9	33.5
	Squared Hellinger Distance ($\times 10^{-2}$)			
2	110.6	119.7	102.2	148.0
3	78.7	86.1	102.6	104.8
4	58.7	63.6	102.0	79.9

Table 4.5: Estimation errors for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$, $\widehat{\mathbf{P}}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$ in clustering modeling 2, under the low rank assumption with $r = 10$ and $n = 50$, $p = 20$.

	$p = 50$			
γ	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$
Squared Frobenius norm error ($\times 10^{-2}$)				
2	4.8	5.1	7.9	9.7
3	4.2	4.4	7.9	8.4
4	3.1	3.3	7.9	6.3
Averaged KL divergence ($\times 10^{-3}$)				
2	27.1	28.0	41.3	54.6
3	22.9	23.8	41.3	46.7
4	16.8	17.9	41.2	34.3
Squared Hellinger Distance ($\times 10^{-2}$)				
2	64.7	66.9	105.5	128.2
3	55.7	57.9	105.7	109.9
4	41.2	43.7	105.4	81.8

Table 4.6: Estimation errors for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$, $\widehat{\mathbf{P}}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$ in clustering modeling 2, under the low rank assumption with $r = 10$ and $n = 50$, $p = 50$.

	$p = 100$			
γ	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{pg}$	$\widehat{\mathbf{P}}_{zr}$
Squared Frobenius norm error ($\times 10^{-2}$)				
2	2.4	2.4	3.8	5.7
3	1.7	1.8	3.8	4.4
4	1.1	1.1	3.8	2.8
Averaged KL divergence ($\times 10^{-3}$)				
2	26.5	26.8	39.9	64.9
3	18.8	19.3	39.8	49.5
4	12.2	12.2	40.2	30.8
Squared Hellinger Distance ($\times 10^{-2}$)				
2	63.3	64.6	102.1	151.1
3	45.7	46.8	102.0	116.0
4	29.9	30.0	102.9	73.6

Table 4.7: Estimation errors for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$, $\widehat{\mathbf{P}}_{pg}$ and $\widehat{\mathbf{P}}_{zr}$ in clustering modeling 2, under the low rank assumption with $r = 10$ and $n = 50$, $p = 100$.

$p = 20$		$p = 50$		$p = 100$	
$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$	$\widehat{\mathbf{P}}_{Frob}$	$\widehat{\mathbf{P}}_{HL}$
Averaged KL divergence ($\times 10^{-4}$)					
135	118	91	62	70	54

Table 4.8: Estimation errors for $\widehat{\mathbf{P}}_{Frob}$, $\widehat{\mathbf{P}}_{HL}$ in the extreme value setting, under the low rank assumption with $r = 10$ and $n = 50$.

Bibliography

[1] Aitchison, J. (2003). The statistical analysis of compositional data. *Caldwell*.

[2] Athey, S. and Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *PNAS*, 113(27):7353–7360.

[3] Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085.

[4] Boutsidis, C. and Gallopoulos, E. (2008). Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362.

[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[6] Brunet, J.-P., Tamayo, P., Golub, T., and Mesirov, J. (2004). Metagenes and molecular pattern discovery using matrix factorization. *In Proceedings of the National Academy of Sciences of the USA*, 101(12):4164–4169.

[7] Buchbinder, S., Mehrotra, D., Duerr, A., Fitzgerald, D., Mogg, R., Li, D., Gilbert, P., Lama, J., Marmor, M., del Rio, C., McElrath, M., Casimiro, D., Gottesdiener, K., Chodakewitz, J., Corey, L., and Robertson, M. (2008). Efficacy assessment of a cell-mediated immunity hiv-1 vaccine (the step study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *Lancet*, 372(9653):1881–1893.

[8] Bühlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data: Methods, theory and applications.

[9] Candès, E. J., Fan, Y., Jason, L., and Lv, J. (2017). Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *arXiv*.

[10] Cao, Y., Zhang, A., and Li, H. (2018). Multi-sample estimation of bacterial composition matrix in metagenomics data. *arXiv*.

[11] Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome research*, 20:947–959.

[12] Dai, B., Ding, S., and Wahba, G. (2013). Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483.

[13] Deng, A., Zhang, P., Chen, S., Kim, D. W., and Lu, J. (2013). Concise summarization of heterogeneous treatment effect using total variation regularized regression. *arXiv*.

[14] Duerr, A., Huang, Y., Buchbinder, S., Coombs, R., Sanchez, J., del Rio, C., Casapia, M., Santiago, S., Gilbert, P., Corey, L., Robertson, M., and Step, H. (2012). Extended follow-up confirms early vaccine-enhanced risk of hiv acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus hiv vaccine (step study). *Journal of Infectious Diseases*, 206(2):258–266.

[15] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

[16] Fan, Y., Demirkaya, E., Li, G., and Lv, J. (2017). Rank: Large-scale inference with graphical nonlinear knockoffs. *arXiv*.

[17] Fan, Y., Kong, Y., Li, D., and Lv, J. (2016). Interaction pursuit with feature screening and selection. *arXiv*.

[18] Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and

Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*.

[19] Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N., Müller, M., Herman, T., Giladi, N., Kalinin, A., Spino, C., Dauer, W., Hausdorff, J., and Dinov, I. (2018). Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in parkinsons disease. *Scientific Report*, (1):7129.

[20] Glonek, G. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, (83):15–28.

[21] Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association*, 81(396):945–960.

[22] Huang, Y., Follmann, D., Nason, M., Zhang, L., Huang, Y., Mehrotra, D., Moodie, Z., Metch, B., Janes, H., Keefer, M., Churchyard, G., Robb, M., Fast, P., Duerr, A., McElrath, M., Corey, L., Mascola, J., Graham, B., Sobieszczyk, M., Kublin, J., Robertson, M., Hammer, S., Gray, G., Buchbinder, S., and Gilbert, P. (2015). Effect of rad5-vector hiv-1 preventive vaccines on hiv-1 acquisition: a participant-level meta-analysis of randomized trials. *PLoS One*, 10(9).

[23] Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

[24] Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Z. Physik*, 31(4):253–258.

[25] Lewis, J., Chen, E., Baldassano, R., Otley, A., Griffiths, A., Lee, D., Bittinger, K., Bailey, A., Friedman, E., Hoffmann, C., Albenberg, L., Sinha, R., Compher, C., Gilroy, E., Nessel, L., Grant, A., Chehoud, C., Li, H., Wu, G., and Bushman, F. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohn's disease. *Cell Host and Microbe*, 18:489–500.

[26] Lin, W., Shi, P., Feng, R., and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797.

[27] Liu, T. and Tao, D. (2016). On the performance of manhattan nonnegative matrix factorization. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, 27(9).

[28] Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome Med*, 8(51).

[29] Martinez-Cajas, J., Wainberg, M., Oliveira, M., Asahchop, E., Doualla-Bell, F., Lisovsky, I., Moisi, D., Mendelson, E., Grossman, Z., and Brenner, B. (2012). The role of polymorphisms at position 89 in the hiv-1 protease gene in the development of drug resistance to hiv-1 protease inhibitors. *Journal of Antimicrobial Chemotherapy*, 67(4):988–994.

[30] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*.

[31] Mittal, S., Bandaranayake, R., King, N., Prabu-Jeyabalan, M., Nalam, M., Nalivaika, E., Yilmaz, N., and Schiffer, C. (2013). Structural and thermodynamic basis of amprenavir/darunavir and atazanavir resistance in hiv-1 protease with mutations at residue 50. *Journal of Virology*, 87(8):4176–4184.

[32] Peng, C., Wong, K.-C., Rockwood, A., Zhang, X., Jiang, J., and Keyes, D. (2012). Multiplicative algorithms for constrained non-negative matrix factorization. *IEEE 12th International Conference on Data Mining*.

[33] Poggio, T. and Shelton, C. (2002). On the mathematical foundations of learning. *Ameri. Math. Soc.*, 39(1):1–49.

[34] Qaqish, B. and Ivanova, A. (2006). Multivariate logistic models. *Biometrika*, (4):1011–1017.

[35] Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319.

[36] Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D., and Shafer, R. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 103(46):17355–17360.

[37] Sesia, M. and Candès, E. J. (2018). Gene hunting with hidden markov model knockoffs. *Biometrika*.

[38] Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494.

[39] Shi, P., Zhang, A., and Li, H. (2016). Variable selection in regression with compositional covariates. *Annals of Applied Statistics*, 10:1019–1040.

[40] Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of American Statistical Association*, 109(508):1517–1532.

[41] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.

[42] Tie, Y., Wang, Y.-F., Boross, P., Chiu, T.-Y., Ghosh, A., Tozser, J., Louis, J., Harrison, R., and Weber, I. (2012). Critical differences in hiv-1 and hiv-2 protease specificity for clinical inhibitors. *Protein Science*, 21(3):339–350.

[43] Turnbaugh, P., Hamady, M., Yatsunenko, T., Cantarel, B., Duncan, A., Ley, R., Sogin, M., Jones, W., Roe, B., Affourtit, J., Egholm, M., Henrissat, B., Heath, A., Knight, R., and Gordon, J. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484.

[44] Vapnik, V. (2000). *The Nature of Statistical Learning Theory*.

[45] Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *arXiv*.

[46] Weinstein, A., Barber, R., and Candès, E. J. (2018). A power and prediction analysis for knockoffs with lasso statistics. *arXiv*.

[47] Wu, M., Cai, T., and Lin, X. (2010). Testing for regression coefficienets in lasso regularized regression. *Technical report, Harvard University*.

[48] Xiao, Y., Angulo, T., Friedman, J., Waldor, M., Weiss, S., and Liu, Y. (2017). Mapping the ecological networks of microbial communities from steady-state data. *bioRxiv*, page 150649.

[49] Xie, Y. and Chan, G. (2019). False discovery rate controlled binary variable selection in regression framework. *arXiv*.

[50] Xie, Y., Chen, N., and Shi, X. (2018). False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 876–885.

[51] Xue, L., Zou, H., and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *Annals of Statistics*, 40(3):1403–1429.

[52] Yu, Y. and Feng, Y. (2014). Model selection via multifold cross validation. *Journal of Computational and Graphical Statistics*, 23(4):1009–1027.

[53] Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 21(1):299–313.

[54] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320.

Appendix A

Proof

A.1 Proof of Theorem 1

Proof. Since the Binary Knockoffs $\tilde{X}$ satisfy the Model-X knockoffs conditions, by Lemma 2 and Lemma 3 in Candès et al. [9], the signs of the null statistics $\{W_j : \beta_j = 0\}$ for $j = 1, \ldots, p$ are distributed as random coin flips. Hence, following the same arguments in the proof of Theorems 1 and 2 in Barber and Candès [3], our Knockoff procedure controls the false discovery rate at a pre-specified level. □

A.2 Proof of Theorem 2

In order to prove Theorem 2, we need the following Lemma 2 and Lemma 3.

Lemma 2. *Assume that* $\mathbf{X}^* \in \mathbb{R}^{n \times 2p}$ *has independent rows with all values being 0 or 1, and* $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ *are i.i.d sub-Gaussian components. Then we have*

$$Pr\left(\left|\left|\frac{1}{n}\left(\mathbf{X}^*\right)^T \epsilon\right|\right|_\infty \leq C_2\sqrt{(\log p)/n}\right) \geq 1 - p^{-C_3} \tag{A.1}$$

for large enough constant $C_2 > 0$ *and some constant* $C_3 > 0$.

Proof. (Lemma 2) Since $X^*_{ij} = 0$ or 1 and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are i.i.d sub-Gaussian components by assumption, for $t > 0$ we have

$$Pr\left(\left|\epsilon_i X^*_{ij}\right| > t\right) \leq Pr\left(|\epsilon_i| > t\right) \leq C_1 \exp\left(-C_1^{-1} t^2\right). \tag{A.2}$$

Thus by Lemma 6 in Fan et al. [17], we have

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i X^*_{ij}\right| > v\right) \leq \tilde{C}_1 \exp\left(-\tilde{C}_1 n v^2\right) \tag{A.3}$$

for some $\tilde{C}_1 > 0$ and all $0 < v < 1$. Hence

$$1 - Pr\left(\left|\left|\frac{1}{n}\left(\mathbf{X}^*\right)^T \epsilon\right|\right|_\infty \leq v\right) = Pr\left(\left|\left|\frac{1}{n}\left(\mathbf{X}^*\right)^T \epsilon\right|\right|_\infty > v\right) \tag{A.4}$$

$$= Pr\left(\max_{1\leq j\leq 2p}\left|\frac{1}{n}\epsilon^T \mathbf{X^*}_j\right| > v\right) \tag{A.5}$$

$$\leq 2p\tilde{C}_1 \exp\left(-\tilde{C}_1 n v^2\right). \tag{A.6}$$

Substituting $v = C\sqrt{(\log p)/n}$ into the above inequality and taking large enough C_2, we have the stated result in Lemma 2. □

Note that our knockoff construction method ensures that X^* is binary. In addition, Condition 1 implies the inequality of the error term are sub-Gaussian as assumed in Lemma 2. Therefore, Lemma 2 always holds in our setup.

Based on the result of Lemma 2, combining with the basic inequality of Lasso regression, we can derive $\left|\left|\left(\hat{\beta}(\lambda) - \beta_T\right)_{\mathcal{S}^C}\right|\right|_1 \leq 3\left|\left|\left(\hat{\beta}(\lambda) - \beta_T\right)_{\mathcal{S}}\right|\right|_1$ with high probability.

Lemma 3. *With high probability, Condition 4 implies the compatibility condition for* $\Sigma_1 = X^{*T}X^*$ *with some constant* $\phi_{\Sigma_1} > 0$

Proof. (Lemma 3) Note that $Z := \Sigma_1 - \Sigma_0 = \Sigma_1 - \mathbb{E}\left[\Sigma_1\right]$, so $Z_{jk} = \frac{1}{n}\left(\sum_{i=1}^n Z^{(i)}_{jk}\right)$ where each $Z^{(i)}_{jk}$ is zero-mean and bounded (since $|Z^{(i)}_{jk}| \leq 2$). By the Azuma-Hoeffding bound,

$$P\left((Z_{jk})^2 \geq \lambda^2\right) = P\left(\left|\frac{1}{n}\left(\sum_{i=1}^{n} Z^{(i)}_{jk}\right)\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2 n}{32}\right). \tag{A.7}$$

Therefore, $||\Sigma_1 - \Sigma_0||_\infty \leq \lambda$ holds with high probability.

Given $||\Sigma_1 - \Sigma_0||_\infty \leq \lambda$, by Bühlmann and van de Geer [8] Lemma 6.17, for all α s.t. $||\alpha_{\mathcal{S}^C}||_1 \leq 3||\alpha_{\mathcal{S}}||_1$ and Σ_0-compatibility condition holds, we have

$$\left|\frac{\alpha^T\Sigma_1\alpha}{\alpha^T\Sigma_0\alpha} - 1\right| \leq \frac{16\lambda s}{\phi^2_{\Sigma_0}}. \tag{A.8}$$

By Bühlmann and van de Geer [8] Corollary 6.8, then Σ_1-compatibility condition holds with $\phi^2_{\Sigma_1} \geq \phi^2_{\Sigma_0}/2$. □

Since Σ_1-compatibility condition is implied by Condition 4 with high probability, then by Theorem 6.1 in Bühlmann and van de Geer [8] and the result from Lemma 2 that $\left|\left|\left(\hat{\beta}(\lambda) - \beta_T\right)_{S^c}\right|\right|_1 \leq 3\left|\left|\left(\hat{\beta}(\lambda) - \beta_T\right)_{S}\right|\right|_1$ with high probability, we have $\left|\left|\hat{\beta}(\lambda) - \beta_T\right|\right|_1 = \mathcal{O}(s\lambda)$ where $\lambda = C\sqrt{(\log p)/n}$ with high probability, for some constant $C > 0$, which will be used in the proof of Theorem 2.

Proof. (Theorem 2) Now we start proving the main result in Theorem 2 by mimicking the way of proof of Theorem 3 in Fan et al. [16]. Denote W_j to be the LCD based on $\hat{\beta}(\lambda)$, and let $\left|W_{(1)}\right| \geq \ldots \geq \left|W_{(p)}\right|$ be the ordered knockoff statistics according to absolute size. Denote j^* the index such that $|W_{j^*}| = T$, where T is the threshold defined in (2.3). It holds that $-T < |W_{j^*+1}| \leq 0$.

Case 1: For the case of $W_{(j^*+1)} = 0$, we have $W_k = 0$ for $k = j^* + 1, \ldots, p$. Then the index set $\{j : W_j \neq 0\}$ is same as the index set of $\hat{\mathcal{S}}$ selected by the Knockoff procedure. We have

$$\{1, \ldots, p\} \setminus S_1 \subset \hat{\mathcal{S}}, \tag{A.9}$$

where $S_1 = \left\{1 \leq j \leq p : \hat{\beta}_j(\lambda) = 0\right\}$.

We have shown in Lemma 2 and Lemma 3 that with high probability $\left|\left|\hat{\beta}(\lambda) - \beta_T\right|\right|_1 = \mathcal{O}(s\lambda)$. Then we have

$$\begin{aligned}\mathcal{O}(s\lambda) = \left|\left|\hat{\beta}(\lambda) - \beta_T\right|\right|_1 &\geq \sum_{j\in S_1\cap\mathcal{S}} \left|\hat{\beta}_j(\lambda) - \beta_{T,j}\right| = \sum_{j\in S_1\cap\mathcal{S}} |\beta_{T,j}| \\ &\geq |S_1 \cap \mathcal{S}| \min_{j\in\mathcal{S}} |\beta_{T,j}|\,. \end{aligned} \tag{A.10}$$

Since $\beta_{0,j} = \beta_{T,j}$ for $1 \leq j \leq p$, by Condition 2 and $\lambda = \mathcal{O}(\sqrt{\frac{\log p}{n}})$, we can derive from (A.10) that $|S_1 \cap \mathcal{S}| = o(s)$, where $s = |\mathcal{S}|$. Also note that $|(\{1, \ldots, p\} \setminus S_1) \cap \mathcal{S}| \geq |\mathcal{S}| - |S_1 - \mathcal{S}| = (1 - o(1))\, s$. Together with (A.9), we obtain

$$\left|\hat{\mathcal{S}} \cap \mathcal{S}\right| \geq (1 - o(1))\, s. \tag{A.11}$$

Therefore, with asymptotic probability one, we have $\frac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{s} \geq 1 - o(1)$.

Case 2: For the case of $-T < |W_{j^*+1}| < 0$, we first note that

$$\frac{|\{j : W_j \leq -T\}| + 2}{|\{j : W_j \geq T\}|} > q, \tag{A.12}$$

where q is the pre-specified FDR control level. Then by Condition 3 together with (A.12), we have $|\{j : W_j \leq -T\}| > q\,|\{j : W_j \geq T\}| - 2 \geq qcs - 2$ with asymptotic probability one. In addition, in this case, $\left|\hat{\beta}_{j+p}(\lambda)\right| \geq T$ for all j such that $W_j \leq -T$. Again, using the result of $\left|\left|\hat{\beta}(\lambda) - \beta_T\right|\right|_1$ from Lemma 2–3, we obtain

$$\mathcal{O}(s\lambda) = \left|\left|\hat{\beta}(\lambda) - \beta_T\right|\right|_1 \geq \sum_{j : W_j \leq -T} \left|\hat{\beta}_{j+p}(\lambda)\right| \geq T\,|\{j : W_j \leq -T\}|\,. \tag{A.13}$$

Therefore, $\mathcal{O}(s\lambda) \geq T(qcs - 2)$, thus $T \leq \mathcal{O}(\lambda)$.

On the other hand, we have

$$\begin{aligned} \mathcal{O}(s\lambda) = \left|\left|\hat{\beta}(\lambda) - \beta_T\right|\right|_1 &= \sum_{j=1}^{p} \left(\left|\hat{\beta}_j(\lambda) - \beta_{T,j}\right| + \left|\hat{\beta}_{j+p}(\lambda)\right|\right) \\ &\geq \sum_{j \in \mathcal{S} \cap \left(\hat{\mathcal{S}}\right)^C} \left(\left|\hat{\beta}_j(\lambda) - \beta_{T,j}\right| + \left|\hat{\beta}_j(\lambda)\right| - T\right). \end{aligned} \tag{A.14}$$

By Condition 2, we have $\min_{j \in \mathcal{S}} |\beta_{0,j}| \geq \tau_n \lambda$ for some $\tau_n \to \infty$. Therefore, by (A.14) and triangle inequality, we get

$$\mathcal{O}(s\lambda) \geq \sum_{j \in \mathcal{S} \cap \left(\hat{\mathcal{S}}\right)^C} (|\beta_{0,j} - T|) \geq (\lambda\tau_n - T) \left|\left\{j \in \mathcal{S} \cap \left(\hat{\mathcal{S}}\right)^C\right\}\right|. \tag{A.15}$$

With some algebra, we can conclude that $\frac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{s} \geq 1 - o(1)$.

By combining the results in two cases, we complete the proof of Theorem 2. □

A.3 Proof of Theorem 3

Proof. By Theorem 1, Lemma 3 and the proof of Proposition 1 in Ravikumar et al. [35], if Conditions 5–6 are satisfied by the population Fisher information matrix Q^*, and $\lambda \geq \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$ for α in Condition 6, then

$$||\hat{\beta}(\lambda) - \beta_T||_2 \leq \frac{5}{C_{\min}}\sqrt{s}\lambda \tag{A.16}$$

with probability greater than $1-2exp\{-c\lambda^2 n\}$ for some positive constant c. Therefore, with high probability $||\hat{\beta}(\lambda) - \beta_T||_1 = \mathcal{O}(s\lambda)$. Then following same arguments as in the proof of Theorem 2 with the use of Conditions 2–3, we have asymptotic power equal to one in this case as well. □

A.4 Proof of Theorem 4

We use the following three lemmas from existing literature to prove Theorem 4.

Lemma 4 (Slepian's Lemma)**.** *A Gaussian process of X is defined as $\mathbb{G}_X = \sum_{i=1}^n \alpha_i x_i$. Let Ω and Ψ be mean zero, separable Gaussian processes index by a common set $\mathcal{S}$, such that*

$$\mathbb{E}\left[(\Omega_{s_1} - \Omega_{s_2})^2\right] \leq \mathbb{E}\left[(\Psi_{s_1} - \Psi_{s_2})^2\right]$$

for all $s_1, s_2 \in \mathcal{S}$. Then

$$\mathbb{E}\left[\sup_{s\in\mathcal{S}} \Omega_s\right] \leq \mathbb{E}\left[\sup_{s\in\mathcal{S}} \Psi_s\right].$$

Lemma 5. *Let $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$ be the independent Rademacher variables and independent standard normal variables, respectively. Let $w_1, \ldots, w_n$ be i.i.d samples and $\mathcal{F}$ a function class. Then*

$$\mathbb{E}_w\mathbb{E}_\alpha \sup_{f\in\mathcal{F}} \frac{2}{n}\sum_{i=1}^n \alpha_i f(w_i) \leq \sqrt{\frac{\pi}{2}}\mathbb{E}_w\mathbb{E}_\beta \sup_{f\in\mathcal{F}} \frac{2}{n}\sum_{i=1}^n \beta_i f(w_i).$$

Lemma 6 (Theorem 8 in Liu and Tao [27]). *Let $\alpha_1, \ldots, \alpha_n$ be the independent Rademacher variables. Let $\mathcal{F}$ be an $[a, b]$-valued function class on $\mathcal{W}$ and $W = (w_1, \ldots, w_n) \in \mathcal{W}^n$. For any $\sigma > 0$, with probability at least 1-σ, we have*

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}_w f(w) - \frac{1}{n} \sum_{i=1}^{n} f(w_i) \right) \leq \mathbb{E}_w \mathbb{E}_\alpha \sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \alpha_i f(w_i) + (b-a) \sqrt{\frac{\log(1/\sigma)}{2n}}.$$

Now we start proving the main result in Theorem 4.

Proof. Let α_k's and β_{kij}'s be the independent standard norm variables, and e_k be a vector with kth element equal to 1 and other elements equal to 0. Let $w_1, \ldots, w_n$ be i.i.d samples. Define

$$\Omega_{\mathbf{X}} = \sum_{k=1}^{n} \alpha_k \min_{h \in \mathbb{R}_+^r, \sum_{j=1}^{r} h_j = 1} \left|\left| w_k - h^T \mathbf{X} \right|\right|_2^2, \tag{A.17}$$

and

$$\Psi_{\mathbf{X}} = 2\sqrt{p} \sum_{k=1}^{n} \sum_{i=1}^{r} \sum_{j=1}^{p} \beta_{kij} e_i^T \mathbf{X} e_j \tag{A.18}$$

We have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\Omega_{\mathbf{X}_1}-\Omega_{\mathbf{X}_2}\right)^2\right] \\
&=\sum_{k=1}^{n}\left(\min_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\left|\left|w_k-h^T\mathbf{X}_1\right|\right|_2^2-\min_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\left|\left|w_k-h^T\mathbf{X}_2\right|\right|_2^2\right)^2 \\
&=\sum_{k=1}^{n}\left(\min_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\max_{g\in\mathbb{R}_+^r,\sum_{j=1}^r g_j=1}\left(\left|\left|w_k-h^T\mathbf{X}_1\right|\right|_2^2-\left|\left|w_k-g^T\mathbf{X}_2\right|\right|_2^2\right)\right)^2 \\
&\leq\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\left(\left|\left|w_k-h^T\mathbf{X}_1\right|\right|_2^2-\left|\left|w_k-h^T\mathbf{X}_2\right|\right|_2^2\right)\right)^2 \\
&=\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\sum_l\left(\left(w_{kl}-h^T\left(\mathbf{X}_1\right)_l\right)^2-\left(w_{kl}-h^T\left(\mathbf{X}_2\right)_l\right)^2\right)\right)^2 \\
&=\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\sum_l\left(2w_{kl}-h^T\left(\mathbf{X}_1\right)_l-h^T\left(\mathbf{X}_2\right)_l\right)\left(h^T\left(\mathbf{X}_2\right)_l-h^T\left(\mathbf{X}_1\right)_l\right)\right)^2 \\
&\leq\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\left(\max_l\left|2w_{kl}-h^T\left(\mathbf{X}_1\right)_l-h^T\left(\mathbf{X}_2\right)_l\right|\left|\left|h^T\mathbf{X}_2-h^T\mathbf{X}_1\right|\right|_1\right)\right)^2 \\
&\leq 4\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\left|\left|h^T\mathbf{X}_2-h^T\mathbf{X}_1\right|\right|_1\right)^2 \\
&=4\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\sum_{j=1}^{r}h_j\left|\left|e_j^T\left(\mathbf{X}_1-\mathbf{X}_2\right)\right|\right|_1\right)^2 \\
&\leq 4\sum_{k=1}^{n}\max_{h\in\mathbb{R}_+^r,\sum_{j=1}^r h_j=1}\sum_{j=1}^{r}h_j^2\sum_{j=1}^{r}\left|\left|e_j^T\left(\mathbf{X}_1-\mathbf{X}_2\right)\right|\right|_1^2 \\
&\leq 4p\sum_{k=1}^{n}\sum_{j=1}^{r}\left|\left|e_j^T\left(\mathbf{X}_1-\mathbf{X}_2\right)\right|\right|_2^2 \\
&=\mathbb{E}\left[\left(\Psi_{\mathbf{X}_1}-\Psi_{\mathbf{X}_2}\right)^2\right]. \qquad \text{(A.19)}
\end{aligned}
$$

The second inequality is because of the dual norm inequality. The third inequality is due to the fact that $\max_l\left|2w_{kl}-h^T\left(\mathbf{X}_1\right)_l-h^T\left(\mathbf{X}_2\right)_l\right|\leq 2$. The fourth and fifth inequalities are due to the constraint on h and Cauchy-Schwarz inequality.

Given the results in (A.19), we now have the condition in Lemma 4 hold. Let $f_{\mathbf{X}}(w)=$

$\min_{h\in\mathbb{R}^r_+,\sum_{j=1}^r h_j=1} \left|\left|w - h^T\mathbf{X}\right|\right|_2^2$. Then

$$
\begin{aligned}
&\mathbb{E}_w\mathbb{E}_\alpha \sup_{\mathbf{X}} \frac{2}{n}\sum_{i=1}^n \alpha_i f_{\mathbf{X}}(w_i) \\
&\leq \frac{\sqrt{2\pi}}{n}\mathbb{E}\sup_{\mathbf{X}} \Omega_{\mathbf{X}} \\
&\leq \frac{\sqrt{2\pi}}{n}\mathbb{E}\sup_{\mathbf{X}} \Psi_{\mathbf{X}} \\
&\leq \frac{2\sqrt{2\pi p}}{n}\mathbb{E}\sup_{\mathbf{X}} \sqrt{\sum_{i=1}^r\sum_{j=1}^p\left(\sum_{k=1}^n \beta_{kij}\right)^2}\sum_{i=1}^r\sum_{j=1}^p \left|e_i^T\mathbf{X}e_j\right| \\
&= \frac{2\sqrt{2\pi pr}}{n}\mathbb{E}\sup_{\mathbf{X}} \sqrt{\sum_{i=1}^r\sum_{j=1}^p\left(\sum_{k=1}^n \beta_{kij}\right)^2} \\
&\leq \frac{2\sqrt{2\pi pr}}{n}\sqrt{\sum_{i=1}^r\sum_{j=1}^p n} \\
&= \frac{2\sqrt{2\pi}pr^{\frac{3}{2}}}{\sqrt{n}}
\end{aligned}
\tag{A.20}
$$

The first inequality is due to Lemma 5. The second inequality is due to Lemma 4. The third inequality is derived by using the definition of $\Psi_{\mathbf{X}}$ and Cauchy-Schwarz Inequality. The first equality is using the fact that $||\mathbf{X}_i||_1 = 1$ by the constraint. The fourth inequality is using Jensen's inequality and the fact that β_{kij} are independent standard normal.

Since $f_{\mathbf{X}}(w) \leq \left|\left|w - \frac{1}{r}\mathbf{1}_r^T\mathbf{X}\right|\right|_2^2 \leq p$, it is a $[0,p]$-valued function. Then apply Lemma 6 together with the result from (A.19), we complete the proof of Theorem 4. □

A.5 Proof of Theorem 6

We will use Lemma 4–6 again in this proof.

Proof. Let α_k's and β_{kij}'s be the independent standard norm variables, and e_k be a vector with kth element equal to 1 and other elements equal to 0. Let $w_1, \ldots, w_n$ be i.i.d samples.

Define

$$\Omega_{\mathbf{X}} = \sum_{k=1}^{n} \alpha_k \min_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1, h_j\geq a} \left|\left|\sqrt{w_k} - \sqrt{h^T\mathbf{X}}\right|\right|_2^2, \tag{A.21}$$

and

$$\Psi_{\mathbf{X}} = \sqrt{\frac{p}{a}} \sum_{k=1}^{n}\sum_{i=1}^{r}\sum_{j=1}^{p} \beta_{kij} e_i^T \mathbf{X} e_j. \tag{A.22}$$

We have

$$\begin{aligned}
&\mathbb{E}\left[(\Omega_{\mathbf{X}_1} - \Omega_{\mathbf{X}_2})^2\right] \\
&= \sum_{k=1}^{n}\left(\min_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \left|\left|\sqrt{w_k} - \sqrt{h^T\mathbf{X}_1}\right|\right|_2^2 - \min_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \left|\left|\sqrt{w_k} - \sqrt{h^T\mathbf{X}_2}\right|\right|_2^2\right)^2 \\
&= \sum_{k=1}^{n}\left(\min_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \max_{g\in\mathbb{R}_+^r, \sum_{j=1}^r g_j=1} \left(\left|\left|\sqrt{w_k} - \sqrt{h^T\mathbf{X}_1}\right|\right|_2^2 - \left|\left|\sqrt{w_k} - \sqrt{g^T\mathbf{X}_2}\right|\right|_2^2\right)\right)^2 \\
&\leq \sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \left(\left|\left|\sqrt{w_k} - \sqrt{h^T\mathbf{X}_1}\right|\right|_2^2 - \left|\left|\sqrt{w_k} - \sqrt{h^T\mathbf{X}_2}\right|\right|_2^2\right)\right)^2 \\
&= \sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \sum_{l}\left(\left(\sqrt{w_{kl}} - \sqrt{h^T\left(\mathbf{X}_1\right)_l}\right)^2 - \left(\sqrt{w_{kl}} - \sqrt{h^T\left(\mathbf{X}_2\right)_l}\right)^2\right)\right)^2 \\
&= \sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \sum_{l}\left(2\sqrt{w_{kl}} - \sqrt{h^T\left(\mathbf{X}_1\right)_l} - \sqrt{h^T\left(\mathbf{X}_2\right)_l}\right)\left(\sqrt{h^T\left(\mathbf{X}_2\right)_l} - \sqrt{h^T\left(\mathbf{X}_1\right)_l}\right)\right)^2 \\
&\leq \sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \left(\max_{l}\left|2\sqrt{w_{kl}} - \sqrt{h^T\left(\mathbf{X}_1\right)_l} - \sqrt{h^T\left(\mathbf{X}_2\right)_l}\right| \left|\left|\sqrt{h^T\mathbf{X}_2} - \sqrt{h^T\mathbf{X}_1}\right|\right|_1\right)\right)^2 \\
&\leq 4\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \left|\left|\sqrt{h^T\mathbf{X}_2} - \sqrt{h^T\mathbf{X}_1}\right|\right|_1\right)^2 \\
&\leq 4\sum_{k=1}^{n}\left(\max_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \left|\left|h^T\mathbf{X}_1 - h^T\mathbf{X}_2\right|\right|_1 \frac{1}{2\sqrt{a}}\right)^2 \\
&\leq \frac{1}{a}\sum_{k=1}^{n} \max_{h\in\mathbb{R}_+^r, \sum_{j=1}^r h_j=1} \sum_{j=1}^{r} h_j^2 \sum_{j=1}^{r}\left|\left|e_j^T\left(\mathbf{X}_1 - \mathbf{X}_2\right)\right|\right|_1^2 \\
&\leq \frac{p}{a}\sum_{k=1}^{n}\sum_{j=1}^{r}\left|\left|e_j^T\left(\mathbf{X}_1 - \mathbf{X}_2\right)\right|\right|_2^2 \\
&= \mathbb{E}\left[(\Psi_{\mathbf{X}_1} - \Psi_{\mathbf{X}_2})^2\right].
\end{aligned} \tag{A.23}$$

The second inequality is because of the dual norm inequality. The third inequality is due to

the fact that $\max_l \left|2\sqrt{w_{kl}} - \sqrt{h^T (\mathbf{X}_1)_l} - \sqrt{h^T (\mathbf{X}_2)_l}\right| \leq 2$. The fourth inequality is due to the constraint that $\mathbf{X}_{kj} \geq a$. The last few inequalities are due to the constraint on h and Cauchy-Schwarz inequality.

Now we have the condition in Lemma 4 hold. Let $f_w^{HL}(\mathbf{X}) = \min_{h \in \mathbb{R}_+^r, \sum_{j=1}^r h_j = 1} \left|\left|\sqrt{w} - \sqrt{h^T \mathbf{X}}\right|\right|_2^2$. Then

$$
\begin{aligned}
&\mathbb{E}_w \mathbb{E}_\alpha \sup_{\mathbf{X}} \frac{2}{n} \sum_{i=1}^n \alpha_i f_{w_i}^{HL}(\mathbf{X}) \\
&\leq \frac{\sqrt{2\pi}}{n} \mathbb{E} \sup_{\mathbf{X}} \Omega_{\mathbf{X}} \\
&\leq \frac{\sqrt{2\pi}}{n} \mathbb{E} \sup_{\mathbf{X}} \Psi_{\mathbf{X}} \\
&\leq \frac{\sqrt{2\pi p}}{n\sqrt{a}} \mathbb{E} \sup_{\mathbf{X}} \sqrt{\sum_{i=1}^r \sum_{j=1}^p \left(\sum_{k=1}^n \beta_{kij}\right)^2} \sum_{i=1}^r \sum_{j=1}^p \left|e_i^T \mathbf{X} e_j\right| \\
&= \frac{\sqrt{2\pi p r}}{n\sqrt{a}} \mathbb{E} \sup_{\mathbf{X}} \sqrt{\sum_{i=1}^r \sum_{j=1}^p \left(\sum_{k=1}^n \beta_{kij}\right)^2} \\
&\leq \frac{\sqrt{2\pi p r}}{n\sqrt{a}} \sqrt{\sum_{i=1}^r \sum_{j=1}^p n} \\
&= \frac{\sqrt{2\pi} r^{\frac{3}{2}} p}{n\sqrt{a}}
\end{aligned}
\tag{A.24}
$$

The first inequality is due to Lemma 5. The second inequality is due to Lemma 4. The third inequality is derived by using the definition of $\Psi_{\mathbf{X}}$ and Cauchy-Schwarz Inequality. The first equality is using the fact that $||\mathbf{X}_i||_1 = 1$ by the constraint. The fourth inequality is using Jensen's inequality and the fact that β_{kij} are independent standard normal.

Since $f_w^{HL}(\mathbf{X}) \leq p$, it is a $[0, p]$-valued function. Then apply Lemma 6 together with the result from (A.23), we complete the proof of Theorem 6. □