STATISTICAL METHODS FOR COMPOSITIONAL AND TREE-STRUCTURED COUNT DATA IN HUMAN MICROBIOME STUDIES

Pixu Shi

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

_____

Hongzhe Li, Professor of Biostatistics

Graduate Group Chairperson

_____

John H. Holmes, Professor of Medical Informatics in Epidemiology

Dissertation Committee

Nandita Mitra, Associate Professor of Biostatistics

Frederic D. Bushman, William Maul Measey Professor in Microbiology

T. Tony Cai, Dorothy Silberberg Professor in Statistics

James D. Lewis, Professor of Medicine

ACKNOWLEDGEMENT

I would like to thank my advisor Dr. Hongzhe Li, who has been an extraodinary mentor during the past four years. His expertise, vision and perspicacity in the field make it possible for me to work on such an exciting and promising research topic. My life as a PhD student has been very smooth given his kindness, generosity and constant support. I feel very fortunate to start my career working with Dr. Li.

I would like to thank my committee chair Dr. Nandita Mitra for watching over me in every step from matriculation to graduation. Her upbeat and warm personality makes it a delightful experience to work with her. I am also very grateful to my committee members Dr. Rick Bushman, Dr. Tony Cai and Dr. James Lewis for providing valuable comments and encouragements.

I would like to express my gratitude to my coauthors Dr. Wei Lin, Dr. Dylan Small, Colin Fogarty and my fellow student Eric Zhang Chen for their help in my research. I am also thankful for having the joyful company of my friends from Penn Biostatistics.

Last of all, I would like to thank my husband Anru Zhang for being supportive in every aspect of my life. He bears with my crankiness when I have difficulties and brings fresh views to my work by discussing my research with me. This dissertation would not be possible without him.

ABSTRACT

STATISTICAL METHODS FOR COMPOSITIONAL AND TREE-STRUCTURED COUNT DATA IN HUMAN MICROBIOME STUDIES

Pixu Shi

Hongzhe Li, PhD

In human microbiome studies, sequencing reads data are often summarized as counts of bacterial taxa at various taxonomic levels. In this thesis, we develop statistical methods for analyzing such counts data. We first consider regression analysis with bacterial counts normalized into compositions as covariates. In order to satisfy the subcompositional coherence of the resulting model, linear models with a set of linear constraints on the regression coefficients are introduced. A penalized estimation procedure for estimating the regression coefficients and for selecting variables under the linear constraints is developed. A method is also proposed to obtain de-biased estimates of the regression coefficients that are asymptotically unbiased and have a joint asymptotic multivariate normal distribution. This provides valid confidence intervals of the regression coefficients and can be used to obtain the p-values. Simulation results have shown the validity of the confidence intervals and smaller variances of the de-biased estimates when the linear constraints are imposed. The proposed methods are applied to a gut microbiome data set and identify four bacterial genera that are associated with the body mass index after adjusting for the total fat and caloric intakes.

We then consider the problem of testing difference between two repeated measurements of microbiome from the same subjects. Multiple microbiome measurements are often obtained from the same subject to assess the difference in microbial composition across body sites or time points. Existing models for analyzing such data are limited in modeling the covariance structure of the counts and in handling paired multinomial data. We propose a new probability distribution for paired multinomial count data, which allows flexible covariance structure of the counts and can be used to model repeatedly measured multivariate counts. Based on this new distribution, a test statistic is developed to test the difference in compositions of paired multinomial count data. The proposed test can be applied to count data observed on taxonomic trees in order to test difference in microbiome compositions and to identify subtrees with different subcompositions. Simulation results

shown that the proposed test has correct type 1 errors and increased power compared to some commonly used methods. An analysis of an upper respiratory tract microbiome data set is used to illustrate the proposed methods.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

1.1. Human microbiome and human health

A typical human body is inhabited by at least 10-100 trillion microbes, outnumbering the human cells by an estimated 10-fold (Turnbaugh et al., 2007). The community formed by microbes, *microbiota*, includes bacteria, fungi, archaea and viruses, and can be found at various human body sites such as gut, skin, oral cavity, vagina, respiratory tract, *etc.*. The collective genome of human microbiota, also known as the *human microbiome*, is estimated to contain $\sim$150 times more genes than the human genome (Qin et al., 2010). Compared with the human genome, the human microbiome has much more diversity. The microbiota found at different body sites of the same individual can differ remarkably, and even at the same body site, human microbiome can display substantial inter-individual variation (Consortium et al., 2012) and temporal variation within the same individual (Flores et al., 2014; Grice et al., 2009).

Many recenyt studies have been investigating the role of microbiome in human health. For example, the gut microbiome has been shown to be associated with many human diseases such as obesity, diabetes and inflammatory bowel disease (Ley et al., 2005, 2006; Manichanh et al., 2012; Qin et al., 2012; Turnbaugh et al., 2006); the skin microbiome has been postulated to have contribution to several skin disorders (Grice et al., 2009; Kong et al., 2012). The health and lifestyles of host have also been shown to affect the composition of microbiome. Studies have found that long-term dietary habits affect the composition of gut microbiota (Wu et al., 2011, 2014). The microbial communities in the upper respiratory tract of cigarette smokers differ between smokers and non-smokers (Charlson et al., 2010; Morris et al., 2013). The host genotypes also have influence on the microbiota compositions (Spor, Koren, and Ley, 2011; Turnbaugh et al., 2006). These links between human microbiome and human health indicate the possibility of designing therapeutic strategies for treatment of complex diseases and conditions by modulating the microbial composition (Cani and Delzenne, 2011; Hsiao et al., 2013; Smits et al., 2013; Virgin and Todd, 2011), and the potential of using microbiome as biomarkers for disease prevention and early diagnostics (Gevers et al., 2014; Kostic et al., 2012; Segata et al., 2011). Several large-scale human microbiome studies such as

the Human Microbiome Project (HMP) (Peterson et al., 2009) and The European Union Project on metagenomics of the human intestinal tract (MetaHIT) (Ehrlich, Consortium, et al., 2011) have provided important data on human microbiome.

1.2. Investigating human microbiome through sequencing

To understand the impact of the human microbiome on human health, it is necessary to characterize and decipher the content of human microbiome. Prior to the invention of Sanger sequencing technology in 1977 (Sanger, Nicklen, and Coulson, 1977), microbiota characterization largely relied on culture-based methods, which are highly biased and time-consuming. The advent of Sanger sequencing allowed for some more thorough view of microbial communities, but is still limited in use by its high cost and low throughput. With the development of next-generation sequencing technology such as Roche (454) pyrosequencing, Illumina Solexa sequencing and Applied Biosystems SOLiD sequencing, researchers are now able to study the human microbiome with much lower cost compared to older Sanger method, yet still achieve a coverage of microbial genes thorough enough to characterize the true microbial population.

Two high-throughput sequencing based approaches have been used for interrogating complex microbial communities. The first approach is based on sequencing the 16S ribosomal RNA (rRNA) amplicons. The 16S rRNA is a structural component of the prokaryotic ribosomes, and thus is presented in all bacteria and archaea cells. The 16S rRNA gene contains highly conserved regions that can be used as primer binding sites in PCR amplification, while its hypervariable regions can be used to identify different bacterial lineages. After sequencing, the 16S sequences are clustered into sequence clusters called *Operational Taxonomic Units* (OTUs) using software pipelines such as Qiime (Caporaso et al., 2010). The representative sequences of each OTU are then compared to the existing 16S databases such as Greengenes (DeSantis et al., 2006), RDP (Cole et al., 2007) and EzTaxon-e (Kim et al., 2012) to obtain taxonomic assignments. The ubiquitous presence of 16S rRNA enables lineage assignments at the levels of kingdom, phylum, class, order, family and genus simultaneously.

Another approach is based on shotgun metagenomic sequencing, which sequences all the microbial genomes presented in the sample, rather than just one marker gene. This approach enables evaluation of both gene abundance and microbial abundance, and can be used to study other

microorganisms such as viruses. The accuracy of this approach in quantifying gene/microbe abundance is highly dependent on the DNA preparation protocols, which demands special cautions for comparative metagenomic studies (Morgan, Darling, and Eisen, 2010). The massive amount of short reads produced also requires efficient computational tools to perform read mapping and assembly, which imposes more challenges in the applications of this approach. Several databases and software tools have been developed to analyze the shotgun metagenomic data (Huson et al., 2007; Meyer et al., 2008; Segata et al., 2012; Seshadri et al., 2007).

1.3. Data structure

1.3.1. Tree-structured count data on taxonomic tree

The taxonomic tree is a tree-structured diagram that illustrates the taxonomic classifications of biological organisms, with each tree node representing a taxon from the taxonomic rank of kingdom to species. For 16S rRNA sequencing, since the same marker gene is used in taxonomic assignments at all ranks, each read compared to the reference database can be subsequently aligned to a node of the taxonomic tree depending on its taxonomic assignment. The resulting data is a set of *tree-structured count data* with each count number representing the number of reads aligned to the corresponding node on the taxonomic tree.

This data format has several features: (1). The total number of reads varies greatly from sample to sample. This can be attributed to the difference in sequencing depth and the amount of DNA yielding materials. (2). The number of reads at each internal node is larger or equal to the sum of read numbers at all its child nodes. This comes from the fact that if a read is assigned a certain taxon, then all the higher ranked taxa along this lineage can also be assigned to this read. (3). There may be a lot of zero counts, which can be caused by the rarity or absence of the corresponding bacteria.

The analysis of tree-structured count data is difficult due to high dimensionality, non-normality and the tree structure underlying the data. Many of the current methods applied to this type of data have been distance-based (Turnbaugh et al., 2006; Wu et al., 2011; Yatsunenko et al., 2012), where a distance measure is defined and computed between each pair of the microbiome samples. One commonly used distance measure is the UniFrac distance (Lozupone and Knight, 2005), upon which many other distance measures are defined, such as the generalized Unifrac distance (Chen

et al., 2012) and the phylogenetic Kantorovich–Rubinstein metric (Evans and Matsen, 2012). The statistical methods based on these distance measures, such as Permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2001) and Principal Coordinate Anlaysis (PCoA) (Gower, 1966; Torgerson, 1958) are widely applied, their results are very dependent on the specific choice of distance measure (Chen et al., 2012).

1.3.2. High-dimensional compositional data

Due to the varying number of reads across samples, the read counts from 16S rRNA sequencing and shotgun metagenomic sequencing are often normalized into vectors of proportions at a given taxonomic level. The resulting data are also referred to as *compositional data* (Aitchison, 1982). The unique feature that the components of a composition must sum to one renders many standard multivariate statistical methods inappropriate or inapplicable. Methodological developments for compositional data analysis have resulted in a fruitful line of research, as thoroughly surveyed by Aitchison (2003). While the dimensionality of compositional data from culture based microbial analysis is often very small due to the limited number of cultivable bacteria, high-throughput sequencing technologies make it possible to identify hundreds of genera and species of bacteria within one microbiome sample set. These large compositional data sets, whose dimensionality is comparable to or much larger than the sample size, pose new challenges to existing methodology. However, little formal attempt has been made to develop principled analysis tools for such high-dimensional compositional data.

1.4. Organization of this thesis

My thesis focuses on the analysis of both high-dimensional compositional data and tree-structured count data on taxonomic tree. In Chapter 2[1], we aim to address the variable selection problem in regression analysis with high-dimensional compositional covariates. We propose an ℓ_1 regularization method for the estimation and variable selection in high-dimensional linear log-contrast model that considers the unique features of the compositional data. We formulate the proposed procedure as a constrained convex optimization problem and introduce a coordinate descent method of multipliers for efficient computation. In the high-dimensional setting where the dimensionality grows at most exponentially with the sample size, model selection consistency and ℓ_∞ bounds for

[1] This part of the thesis is based on the published paper Lin et al. (2014).

the resulting estimator are established under conditions that are mild and interpretable for compositional data. The numerical performance of our method is then evaluated by simulation studies and its usefulness is illustrated by an application to a microbiome study relating body mass index to human gut microbiome composition.

In Chapter 3[2], we consider the inference problem of high-dimensional regression analysis with compositional data, where the goal is to identify the bacterial taxa that are associated with a continuous response such as the body mass index (BMI). In order to satisfy the subcompositional coherence of the results, we propose to use linear models with a set of linear constraints on the regression coefficients. Such models allow regression analysis for subcompositions and include the log-contrast model for compositional covariates as a special case. An ℓ_1 penalized estimation procedure for estimating the regression coefficients and for selecting variables under the linear constraints is developed. To provide valid confidence intervals of the regression coefficients and obtain the corresponding p-values, a method is proposed to obtain de-biased estimates of the regression coefficients that are asymptotically unbiased and have a joint asymptotic multivariate normal distribution. Simulation results show the validity of the confidence intervals and smaller variances of the de-biased estimates when the linear constraints are imposed. The proposed methods are applied to the same gut microbiome dataset as Chapter 2 and results are compared.

In Chapter 4[3], we consider the problem of testing difference between repeatedly measured microbiome data quantified as counts on taxonomic tree. Such repeated data often occur when multiple samples are taken from the same subject to assess the difference in microbial composition across body sites or time points. To model the covariance structure of the count data with flexibility, we propose a general class of probability distributions for paired multinomial count data, which allows us to model repeatedly measured multivariate counts. Based on this new distribution, we develop a test statistic to evaluate the difference in the mean parameters of paired multivariate count data. We then provide a procedure that applies the proposed test to count data observed on an taxonomic tree in order to assess difference in microbiome compositions and to identify subtrees with different sub-compositions. Our simulation results indicate that proposed test has correct type 1 errors and increased power compared to some commonly used methods. The proposed methods are illustrated by an analysis of the human pharynx microbiome data, where nasopharynx microbiome is

[2]This part of the thesis is based on the published paper Shi, Zhang, and Li (2016).
[3]This part of the thesis is based on the peer reviewed manuscript Shi and Li (2016)

compared with oropharynx microbiome, and smokers microbiome is compared with non-smokers.

Some related future topics are discussed in Chapter 5.

CHAPTER 2

VARIABLE SELECTION IN REGRESSION WITH COMPOSITIONAL COVARIATES

2.1. Introduction

Compositional data, which consist of the proportions or percentages of a composition, appear frequently in a wide range of applications; examples include geochemical compositions of rocks in geology, household patterns of expenditures in economics, species compositions of biological communities in ecology, and topic compositions of documents in machine learning. The unique feature that the components of a composition must sum to one renders many standard multivariate statistical methods inappropriate or inapplicable. Since the seminal work of Aitchison (1982), methodological developments for compositional data analysis have resulted in a fruitful line of research, as thoroughly surveyed by Aitchison (2003). The recently increasing availability of large compositional data sets, whose dimensionality is comparable to or much larger than the sample size, poses new challenges to existing methodology. However, little formal attempt has been made to develop principled analysis tools for such data. A typical example arises in metagenomic studies of microbial communities based on 16S rRNA gene sequencing, where the relative abundances of hundreds to thousands of bacterial taxa on a few tens to hundreds of individuals are available for analysis; see, for example, Chen and Li (2013).

The aim of this chapter is to address the variable selection problem in high-dimensional regression with compositional covariates. To mitigate the difficulty with high dimensionality, it is crucial to select parsimonious models that tend to improve the performance of statistical procedures and interpretability of the resulting inferences. Regularization methods for simultaneous variable selection and estimation in linear regression and more general contexts have received intense recent interest. In particular, the ℓ_1 regularization or lasso approach (Tibshirani, 1996) has enjoyed widespread popularity, and its theoretical properties in high-dimensional regression are now well understood; see, for example, Bühlmann and Van De Geer (2011) for an overview. Owing to the special nature of compositional data, however, the usual linear regression model is inappropriate for our purposes. In this chapter we consider the linear log-contrast model of Aitchison and Bacon-shone (1984), which is particularly useful for regression analysis with compositional covariates. Under this model

the expected response does not depend on the basis counts from which a composition is obtained. This is the case in our microbiome data example, where the number of sequencing reads varies drastically across samples and should not play a role in predicting the response of interest.

We propose an ℓ_1 regularization methodology for variable selection and estimation in high-dimensional linear log-contrast models. We formulate the proposed procedure as a constrained convex optimization problem, develop efficient algorithms for computation, and provide strong theoretical guarantees. Since the constraint in the problem couples the parameters, coordinate descent methods for solving ℓ_1-regularized least squares problems (Friedman et al., 2007) are not directly applicable. We therefore combine coordinate descent with the method of multipliers to introduce an efficient algorithm for solving the optimization problem. To establish model selection consistency and ℓ_∞ bounds for the resulting estimator, we impose conditions analogous to the irrepresentability condition for linear regression in Zhao and Yu (2006). Our conditions, however, differ from those for linear regression models in important ways, which account for the compositional effect and adapt well to the dependence structure of compositional data.

2.2. Variable selection in the linear log-contrast model

Log-contrast models were originally introduced by Aitchison and Bacon-shone (1984) for modeling experiments with mixtures, and have proved to be useful for a wide variety of regression problems with a composition playing the role of covariate. Suppose that we observe an n-vector y of responses and an $n \times p$ matrix $X = (x_{ij})$ of covariates, with each row of X lying in the $(p-1)$-dimensional positive simplex $S^{p-1} = \{(x_1, \ldots, x_p) \colon x_j > 0, j = 1, \ldots, p, \sum_{j=1}^{p} x_j = 1\}$. Because of the unit-sum constraint, the p components of a composition cannot vary freely, traditional methodology often requires the omission of certain components to ensure identifiability and encounters intrinsic difficulties in providing sensible interpretations for the regression parameters. To resolve the difficulties with the compositional constraint, Aitchison and Bacon-shone (1984) proposed to apply the log-ratio transformation (Aitchison, 1982) to compositional covariates, resulting in the linear log-contrast model

$$y = Z^p \beta^*_{\setminus p} + \varepsilon, \tag{2.1}$$

where $Z^p = \{\log(x_{ij}/x_{ip})\}$ is the $n \times (p-1)$ log-ratio matrix, whose pth component is the reference component, $\beta^*_{\setminus p} = (\beta^*_1, \ldots, \beta^*_{p-1})^\top$ is the corresponding $(p-1)$-vector of regression coefficients, and

ε is an n-vector of independent noise distributed as $N(0,\sigma^2)$. By introducing a new coefficient $\beta_p^* = -\sum_{j=1}^{p-1}\beta_j^*$, model (2.1) can be more conveniently expressed in the symmetric form

$$y = Z\beta^* + \varepsilon, \quad \sum_{j=1}^{p}\beta_j^* = 0, \tag{2.2}$$

where $Z = (z_1,\ldots,z_p) = (\log x_{ij})$ is the $n \times p$ design matrix and $\beta^* = (\beta_1^*,\ldots,\beta_p^*)^\top$ is the p-vector of regression coefficients. We do not include an intercept in the model, since it can be eliminated by centering the response and predictor variables. We are concerned with the high-dimensional sparse setting, where the dimensionality p is comparable to or much larger than the sample size n, while only a small portion of the regression coefficients are nonzero.

Applying the ℓ_1 regularization approach to model (2.2), we consider the constrained convex optimization problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\left(\frac{1}{2n}\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1\right) \quad \text{subject to } \sum_{j=1}^{p}\beta_j = 0, \tag{2.3}$$

where $\beta = (\beta_1,\ldots,\beta_p)^\top$, $\lambda > 0$ is a regularization parameter, and $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the ℓ_2 and ℓ_1 norms, respectively. The zero-sum constraint in problem (2.3) is crucial for the resulting estimator to enjoy interpretive advantages over a standard lasso estimator. Specifically, the proposed estimator possesses the following desirable properties:

(i) Scale invariance: the estimator is unchanged under the transformation $X \mapsto TX$ for an arbitrary diagonal matrix $T = \operatorname{diag}(t_1,\ldots,t_n)$ with all $t_i > 0$;

(ii) Permutation invariance: the estimator is invariant under any permutation π of the p components, meaning that it is unchanged if π is applied to both the columns of X and the components of $\hat{\beta}$;

(iii) Selection invariance: the estimator is unchanged if one knew in advance which components would be estimated as zero and applied the procedure to the subcomposition formed by the remaining components.

Properties (i) and (iii) are due to the zero-sum constraint; they ensure that the inferences are independent of an arbitrary scaling of the basis from which a composition is obtained, and remain

unaffected by correctly excluding some or all of the zero components. Property (ii) is immediately seen from the symmetric formulation of problem (2.3), but would not be guaranteed by first transforming the p components into a $(p-1)$-dimensional feature space and then applying a standard variable selection procedure.

By eliminating the constraint with $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, we can rewrite problem (2.3) as the unconstrained problem

$$\hat{\beta}_{\setminus p} = \underset{\beta_{\setminus p}}{\operatorname{argmin}} \left(\frac{1}{2n} \|y - Z^p \beta_{\setminus p}\|_2^2 + \lambda \|D\beta_{\setminus p}\|_1 \right),$$

where $\beta_{\setminus p} = (\beta_1, \ldots, \beta_{p-1})^\top$, $D = (I_{p-1}, -1_p)^\top \in \mathbb{R}^{p \times (p-1)}$, and I_r and 1_r denote the $r \times r$ identity matrix and the r-vector of 1s, respectively. This asymmetric form can be recognized as an instance of the generalized lasso problem considered by Tibshirani et al. (2011), but existing developments do not specialize in our case to give an appropriate algorithm or theory for several reasons. First, eliminating one arbitrary component and applying a generic algorithm to the $(p-1)$-dimensional problem generally does not yield numerical solutions that are permutation invariant. Second, a coordinate descent algorithm that is fast and applicable to a prespecified set of λ values is not yet available. Third, theory for the generalized lasso problem does not provide useful insights into the compositional constraint and its effect on variable selection. All these limitations call for the development of computational methods and theoretical results that are relevant to the analysis of compositional data.

2.3. Computation

2.3.1. Optimization algorithm

Coordinate descent algorithms have been shown to be very efficient for solving large-scale ℓ_1 regularization problems (Friedman et al., 2007). They are not directly applicable to problem (2.3), however, because the nondifferentiable ℓ_1 terms are inseparable under the zero-sum constraint. Here we propose an efficient, easily implemented algorithm based on an iterative modification of coordinate descent by combining it with the method of multipliers or the augmented Lagrangian method (Bertsekas, 2014) to deal with the constraint.

To derive the algorithm, we first form the augmented Lagrangian for problem (2.3) as

$$L_\mu(\beta,\gamma) = \frac{1}{2n}\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1 + \gamma\sum_{j=1}^{p}\beta_j + \frac{\mu}{2}\left(\sum_{j=1}^{p}\beta_j\right)^2,$$

where γ is the Lagrange multiplier and $\mu > 0$ is a penalty parameter. The method of multipliers for problem (2.3) consists of the iterations

$$\beta^{k+1} \leftarrow \operatorname*{argmin}_{\beta} L_\mu(\beta,\gamma^k), \quad \gamma^{k+1} \leftarrow \gamma^k + \mu\sum_{j=1}^{p}\beta_j^{k+1}.$$

Define by $\alpha = \gamma/\mu$ the scaled Lagrange multiplier. The above iterations can be more conveniently expressed as

$$\beta^{k+1} \leftarrow \operatorname*{argmin}_{\beta}\left\{\frac{1}{2n}\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1 + \frac{\mu}{2}\left(\sum_{j=1}^{p}\beta_j + \alpha^k\right)^2\right\}, \tag{2.4}$$

$$\alpha^{k+1} \leftarrow \alpha^k + \sum_{j=1}^{p}\beta_j^{k+1}. \tag{2.5}$$

Now the ℓ_1 terms in (2.4) are separable and the subproblem can be solved by coordinate descent. With the other components held fixed, the jth component of β is updated by

$$\beta_j^{k+1} \leftarrow \frac{1}{v_j + \mu} S_\lambda\left\{\frac{1}{n}z_j^\top\left(y - \sum_{i\neq j}\beta_i^{k+1}z_i\right) - \mu\left(\sum_{i\neq j}\beta_i^{k+1} + \alpha^k\right)\right\}, \tag{2.6}$$

where $v_j = \|z_j\|_2^2/n$ and $S_\lambda(t) = \operatorname{sgn}(t)(|t| - \lambda)_+$ is the soft thresholding operator. Combining (2.4)–(2.6) yields the following coordinate descent method of multipliers for solving problem (2.3).

Input: y, Z, and λ.
Output: $\hat{\beta}$

1: Initialize β^0 with 0 or a warm start, $\alpha^0 = 0$, $\mu > 0$, and $k = 0$.
2: For $j = 1,\ldots,p,1,\ldots,p,\ldots$, update β_j^{k+1} by (2.6) until convergence.
3: Update α^{k+1} by (2.5).
4: $k \leftarrow k+1$ and repeat Steps 2 and 3 until convergence. Output $\hat{\beta} = \beta^{k+1}$.

Algorithm 1: *Coordinate descent method of multipliers.*

Minimization of subproblem (2.4), which is carried out in Step 2 of Algorithm 1, need not be exact; it suffices to adopt a stopping criterion such that the minimization is asymptotically exact. This results in a more efficient algorithm for which convergence is still ensured. We have the following result

regarding the convergence of Algorithm 1 with inexact minimization.

Proposition 1. *Assume that Step 2 of Algorithm 1 finds at iteration k an approximate minimizer β^{k+1} such that $L_\mu(\beta^{k+1},\gamma^k) \le \min_\beta L_\mu(\beta,\gamma^k)+\delta_k$ for all k, where $\delta_k \ge 0$ and $\sum_{k=0}^{\infty}\sqrt{\delta_k}<\infty$. Then the sequence $\{\beta^k\}$ generated by Algorithm 1 is bounded. Moreover, every cluster point of $\{\beta^k\}$ is an optimal solution of problem* (2.3).

2.3.2. Tuning parameter selection

The regularization parameter λ can be selected by the generalized information criterion for high-dimensional penalized likelihood proposed by Fan and Tang (2013). They showed that the criterion with a uniform choice of the model complexity penalty identifies the true model with probability tending to 1 when the dimensionality p grows at most exponentially with the sample size n. For model (2.2) and our regularization method, we define

$$\mathrm{GIC}(\lambda) = \log \hat{\sigma}^2_\lambda + (s_\lambda - 1)\frac{\log\log n}{n}\log(p \vee n),$$

where $\hat{\sigma}^2_\lambda = \|y - Z\hat{\beta}_\lambda\|_2^2/n$, $\hat{\beta}_\lambda$ is the regularized estimator, $p \vee n$=max(p,n), and s_λ is the number of nonzero coefficients in $\hat{\beta}_\lambda$. Because of the zero-sum constraint, the effective number of free parameters is $s_\lambda - 1$ for $s_\lambda \ge 2$. We then select the optimal λ by minimizing $\mathrm{GIC}(\lambda)$. Alternatively, one can apply K-fold crossvalidation with $K = 5$ or 10 to choose λ, which tends to select a larger model and trades off between model selection consistency and prediction accuracy. Although crossvalidation is computationally more expensive, it is less parsimonious and can often yield a more satisfactory performance in practice.

The penalty parameter μ that is needed to enforce the zero-sum constraint does not affect the convergence of Algorithm 1 as long as $\mu > 0$, and we take $\mu = 1$ in all computations.

2.4. Theoretical properties

We establish model selection consistency and ℓ_∞ bounds for the proposed estimator under deterministic designs. We first introduce some notation. Let Z^r denote the log-ratio matrix with the rth component taken as the reference component, and $C^r = n^{-1}(Z^r)^\top Z^r$ the corresponding sample log-ratio covariance matrix. Let $S = \{j\colon \beta^*_j \ne 0\}$ denote the support of β^*, and $s = |S|$ the cardinality of S. For any subset $J \subset \{1,\ldots,p\}$ and $j \in J$, denote by J^c the complement of J, and $J_{\setminus j} = J \setminus \{j\}$.

We will use subsets to index a vector or matrix; for example, $C^r_{S^c S_{\setminus r}}$ is the submatrix formed by the (i,j)th entries of C^r with $i \in S^c$ and $j \in S_{\setminus r}$. Define by $\beta_{\min} = \min_{j \in S} |\beta^*_j|$ the minimum signal. Let $\|\cdot\|_\infty$ denote the ℓ_∞ or matrix ∞-norm, i.e., $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ for a matrix $A = (a_{ij})$.

We assume without loss of generality that $p \in S$. Central to guaranteed support recovery through our ℓ_1 regularization method is the following condition.

Condition 1. *There exists some $\xi \in (0,1]$ such that*

$$\|C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} \{\operatorname{sgn}(\beta^*_{S_{\setminus p}}) - \operatorname{sgn}(\beta^*_p) 1_{s-1}\} + \operatorname{sgn}(\beta^*_p) 1_{p-s}\|_\infty \le 1 - \xi. \tag{2.7}$$

Also, our assumption for the minimum signal threshold involves the quantity φ defined by

$$\varphi = \|D_{SS_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} (D_{SS_{\setminus p}})^\top\|_\infty. \tag{2.8}$$

Although the definitions of ξ and φ seem to depend on the choice of the reference component, we show that this is not the case. Let D^r denote the matrix formed by interchanging the rth and pth rows of D. The following proposition states the permutation invariance of ξ and φ.

Proposition 2. *For every $r \in S_{\setminus p}$, we have*

$$\begin{aligned} &C^r_{S^c S_{\setminus r}} (C^r_{S_{\setminus r} S_{\setminus r}})^{-1} \{\operatorname{sgn}(\beta^*_{S_{\setminus r}}) - \operatorname{sgn}(\beta^*_r) 1_{s-1}\} + \operatorname{sgn}(\beta^*_r) 1_{p-s} \\ &\quad = C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} \{\operatorname{sgn}(\beta^*_{S_{\setminus p}}) - \operatorname{sgn}(\beta^*_p) 1_{s-1}\} + \operatorname{sgn}(\beta^*_p) 1_{p-s}, \end{aligned} \tag{2.9}$$

and

$$D^r_{SS_{\setminus r}} (C^r_{S_{\setminus r} S_{\setminus r}})^{-1} (D^r_{SS_{\setminus p}})^\top = D_{SS_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} (D_{SS_{\setminus p}})^\top. \tag{2.10}$$

Condition 1 is in the spirit of the irrepresentability condition for linear regression in Zhao and Yu (2006), though important differences exist. It is worthwhile to compare Condition 1 with its counterparts for two usual lasso estimators:

(i) the condition

$$\|C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} \operatorname{sgn}(\beta^*_{S_{\setminus p}})\|_\infty \le 1 - \xi \tag{2.11}$$

for the lasso problem

$$\hat{\beta}_{\setminus p}^{(\mathrm{i})} = \underset{\beta_{\setminus p}}{\operatorname{argmin}} \left(\frac{1}{2n} \|y - Z^p \beta_{\setminus p}\|_2^2 + \lambda \|\beta_{\setminus p}\|_1 \right), \tag{2.12}$$

which is a direct application of lasso to model (2.1);

(ii) the condition

$$\|C_{S^c S}(C_{SS})^{-1} \operatorname{sgn}(\beta_S^*)\|_\infty \leq 1 - \xi, \tag{2.13}$$

where $C = n^{-1} Z^\top Z$, for the lasso problem

$$\hat{\beta}^{(\mathrm{ii})} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2n} \|y - Z\beta\|_2^2 + \lambda \|\beta\|_1 \right), \tag{2.14}$$

which simply ignores the zero-sum constraint in problem (2.3).

Expression (2.11) lacks the permutation invariance of Condition 1, reflecting the fact that the pth component is not regularized in problem (2.12) and hence no recovery guarantees can be provided. Condition (2.13) is ideally suited to nearly orthogonal designs, but would be problematic for designs with generally negative correlations such as those common in compositional data analysis. In contrast, the extra term $\operatorname{sgn}(\beta_p^*) 1_{p-s}$ in Condition 1 allows it to adapt well to the negative correlations resulting from the compositional constraint.

To develop further intuition for Condition 1, we consider the illustrative example where the covariate matrix X is generated from an orthogonal design $W = (w_{ij})$ with $W^\top W = nI$ by the transformation $x_{ij} = e^{w_{ij}} / \sum_{k=1}^p e^{w_{ik}}$. This represents an extreme case where the dependence among the components is purely due to the unit-sum constraint. In this example, we have $C^p = n^{-1}(Z^p)^\top Z^p = n^{-1} D^\top W^\top W D = D^\top D = I_{p-1} + 1_{p-1} 1_{p-1}^\top$, and then

$$C^p_{S^c S_{\setminus p}} = 1_{p-s} 1_{s-1}^\top, \quad (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} = (I_{s-1} + 1_{s-1} 1_{s-1}^\top)^{-1} = I_{s-1} - s^{-1} 1_{s-1} 1_{s-1}^\top.$$

Some straightforward calculation yields that the left-hand side of (2.7) equals

$$s^{-1} |1_s^\top \operatorname{sgn}(\beta_S^*)| \leq (s-2)/s < 1,$$

where the first inequality is due to the constraint $1_s^\top \beta_S^* = 0$. This implies that Condition 1 holds, and

ξ can be taken close to 1 provided that the signals are nearly evenly divided between positive and negative signs.

We are now ready to state our main result regarding the model selection consistency of the proposed estimator. We assume without loss of generality that the columns of Z are normalized such that $\max_j \|z_j\|_2 \leq \sqrt{n}$.

Theorem 1. *Assume that Condition 1 holds, the regularization parameter λ satisfies $\lambda \geq c_1 \sigma\{(\log p)/n\}^{1/2}/\xi$ for some constant $c_1 > 2\sqrt{2}$, and the minimum signal satisfies $\beta_{\min} > 3\varphi\lambda/2$. Then, with probability at least $1 - p^{-c_2}$ for some constant $c_2 > 0$, problem* (2.3) *has an optimal solution $\hat{\beta}$ that satisfies the following properties:*

(i) sign consistency*: $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^*)$;*

(ii) ℓ_∞ loss*: $\|\hat{\beta}_S - \beta_S^*\|_\infty \leq 3\varphi\lambda/2$.*

To understand the asymptotic implications of Theorem 1, assume for simplicity that ξ and φ are constants. Then Theorem 1 implies that the proposed estimator is model selection consistent and uniformly estimation consistent as long as $\log p = o(n)$. Taking the smallest possible λ, we have the convergence rate $\|\hat{\beta}_S - \beta_S\|_\infty = O_P[\{(\log p)/n\}^{1/2}]$. These rates parallel those for the usual lasso estimator (Wainwright, 2009), but are established here under a different form of the irrepresentability condition, which explicitly takes the zero-sum constraint into account.

2.5. Numerical studies

2.5.1. Simulations

We conducted simulation studies to compare the numerical performance of the proposed method with two usual lasso estimators defined in (2.12) and (2.14), which we refer to as lasso (i) and lasso (ii), respectively. In lasso (i), the reference component is taken at random from the p components, and after $\hat{\beta}_{\setminus p}^{(\mathrm{i})}$ is obtained, we let $\hat{\beta}_p^{(\mathrm{i})} = -1^\top \hat{\beta}_{\setminus p}^{(\mathrm{i})}$. Note that both lasso (i) and the proposed estimator satisfy the zero-sum constraint, whereas lasso (ii) does not.

We generated the covariate data in the following way. We first generated an $n \times p$ data matrix $W = (w_{ij})$ from a multivariate normal distribution $N_p(\theta, \Sigma)$, and then obtained the covariate matrix $X = (x_{ij})$ by the transformation $x_{ij} = e^{w_{ij}} / \sum_{k=1}^{p} e^{w_{ik}}$. The covariates generated thus follow a logistic normal

Table 2.1: Means and standard errors (in parentheses) of various performance measures for three methods based on 100 simulations

(n,p)	Method	PE	ℓ_1 loss	ℓ_2 loss	ℓ_∞ loss	FP	FN
			$\rho = 0.2$				
(50, 30)	Lasso (i)	0.43 (0.01)	1.16 (0.03)	0.19 (0.01)	0.25 (0.01)	5.44 (0.29)	0.00 (0.00)
	Lasso (ii)	0.42 (0.01)	1.10 (0.03)	0.19 (0.01)	0.25 (0.01)	4.15 (0.28)	0.00 (0.00)
	Proposed	0.42 (0.01)	1.05 (0.03)	0.18 (0.01)	0.24 (0.01)	3.57 (0.23)	0.00 (0.00)
(100, 200)	Lasso (i)	0.45 (0.01)	1.25 (0.03)	0.24 (0.01)	0.27 (0.01)	4.94 (0.28)	0.00 (0.00)
	Lasso (ii)	0.42 (0.01)	1.12 (0.02)	0.21 (0.01)	0.26 (0.01)	2.96 (0.23)	0.00 (0.00)
	Proposed	0.41 (0.01)	1.07 (0.02)	0.19 (0.01)	0.24 (0.01)	3.03 (0.24)	0.00 (0.00)
(100, 1000)	Lasso (i)	0.82 (0.05)	2.01 (0.07)	0.69 (0.06)	0.42 (0.02)	5.18 (0.27)	0.28 (0.08)
	Lasso (ii)	0.66 (0.03)	1.68 (0.05)	0.52 (0.03)	0.38 (0.01)	2.84 (0.21)	0.13 (0.04)
	Proposed	0.61 (0.02)	1.57 (0.04)	0.43 (0.03)	0.34 (0.01)	3.10 (0.22)	0.04 (0.02)
			$\rho = 0{\cdot}5$				
(50, 30)	Lasso (i)	0.46 (0.01)	1.55 (0.05)	0.35 (0.03)	0.33 (0.01)	6.70 (0.32)	0.08 (0.04)
	Lasso (ii)	0.43 (0.01)	1.40 (0.04)	0.30 (0.02)	0.31 (0.01)	5.00 (0.30)	0.02 (0.01)
	Proposed	0.42 (0.01)	1.32 (0.04)	0.28 (0.02)	0.30 (0.01)	4.81 (0.27)	0.02 (0.01)
(100, 200)	Lasso (i)	0.62 (0.03)	2.11 (0.07)	0.75 (0.06)	0.48 (0.02)	7.16 (0.39)	0.33 (0.08)
	Lasso (ii)	0.47 (0.01)	1.60 (0.04)	0.45 (0.03)	0.37 (0.01)	4.61 (0.27)	0.09 (0.05)
	Proposed	0.45 (0.01)	1.54 (0.03)	0.40 (0.02)	0.36 (0.01)	4.60 (0.29)	0.01 (0.01)
(100, 1000)	Lasso (i)	1.51 (0.08)	3.70 (0.08)	2.32 (0.11)	0.81 (0.02)	3.39 (0.23)	2.55 (0.12)
	Lasso (ii)	0.94 (0.05)	2.72 (0.08)	1.40 (0.08)	0.62 (0.02)	2.44 (0.20)	1.29 (0.13)
	Proposed	0.91 (0.07)	2.59 (0.08)	1.25 (0.09)	0.59 (0.02)	3.73 (0.29)	0.99 (0.13)

PE, prediction error; FP, number of false positives; FN, number of false negatives.

distribution (Atchison and Shen, 1980). To reflect the fact that the components of a composition in metagenomic data often differ by orders of magnitude, we let $\theta = (\theta_j)$ with $\theta_j = \log(0{\cdot}5p)$ for $j = 1,\ldots,5$ and $\theta_j = 0$ otherwise. To describe different levels of correlations among the components, we let $\Sigma = (\rho^{|i-j|})$ with $\rho = 0{\cdot}2$ or $0{\cdot}5$. We generated the responses according to model (2.2) with $\beta^* = (1,-0{\cdot}8,0{\cdot}6,0,0,-1{\cdot}5,-0{\cdot}5,1{\cdot}2,0,\ldots,0)^\top$ and $\sigma = 0{\cdot}5$, so that three of the six nonzero coefficients were among the five major components and the rest among the minor components.

We set $(n,p) = (50,30)$, $(100,200)$, and $(100,1000)$, and repeated 100 simulations for each setting. The tuning parameter λ was selected by GIC as described in § 2.3.2. We used six performance measures for comparisons. The prediction error $\|y - Z\hat{\beta}\|_2^2/n$ was computed from an independent test sample of size n. The estimation accuracy was assessed by the ℓ_q losses $\|\hat{\beta} - \beta^*\|_q$ with $q = 1,2$, and ∞. Two variable selection measures were the number of false positives and the number of false negatives, where positives and negatives refer to nonzero and zero coefficients, respectively. The means and standard errors of these performance measures for three methods are reported in Table 2.1.

As seen from Table 2.1, the lasso (i) estimator has inferior performance in almost all settings, since the reference component is not regularized and is always included in the selected model. The lasso (ii) estimator performs better than lasso (i), but always violates the zero-sum constraint in finite samples. The proposed estimator performs slightly better than lasso (ii) in terms of prediction and estimation. The variable selection performance of the proposed estimator is comparable to lasso (ii) with low to moderate dimensionality, but it tends to select fewer false negatives at the cost of slightly increased false positives in high dimensions. This is reasonable because missing important variables is more influential than including unimportant variables with shrunk coefficients. A potential remedy for the violation of the zero-sum constraint in the lasso (ii) estimator would be to refit the unpenalized linear log-contrast model with the constraint using the selected variables, which is also useful for reducing the bias caused by the ℓ_1 penalty. In the Supplementary Material, we compare the performance of the two-step procedures formed by adding a refitting step to lasso (ii) or the proposed method, confirming the advantages of our method in the more challenging settings.

2.5.2. Application to gut microbiome data

Gut microbiome composition is considered an important factor that affects energy extraction from the diet and contributes to human health and diseases such as obesity. We illustrate the proposed method by an application to the data set reported in Wu et al. (2011), where a cross-sectional study of 98 healthy volunteers was carried out at the University of Pennsylvania for investigating long-term dietary effect on gut microbiome composition. Stool samples were collected on these subjects and DNA samples were analyzed by 454/Roche pyrosequencing of 16S rRNA gene segments of the V1–V2 region. The pyrosequences were denoised to yield an average of 9265, with standard deviation 3864, reads per sample. After taxonomic assignment of the denoised sequences, 3068 operational taxonomic units were combined into 87 genera that appeared in at least one sample. Since the number of sequencing reads varies greatly across samples, these count data should not be used directly in a standard regression analysis, and we transformed them into compositional data after replacing zero counts by the maximum rounding error 0·5 (Aitchison, 2003 § 11·5). Demographic information including body mass index (BMI) was also collected on these subjects. We are interested in identifying a subset of important genera whose subcomposition is associated with BMI.

We applied the proposed method to this data set with BMI as the response, and used a refitted version of tenfold crossvailidation to choose the tuning parameter, where the prediction error for each sample split was computed with the refitted coefficients obtained after model selection and without penalization. To obtain stable selection results, we generated 100 bootstrap samples and applied the same crossvalidation procedure to select the genera. The selection probabilities of 87 genera with bootstrapped crossvalidation are shown in Fig. 2.1(a). Four genera were selected over 70 times out of the 100 bootstrap replicates. We also followed the approach of stability selection (Meinshausen and Bühlmann, 2010) to assess the stability of the selected genera, where 100 subsamples of size $n/2$ were taken to compute the selection probabilities. All four genera had a selection probability greater than 0·85, indicating that the selection results are quite stable. These four genera along with their selection probabilities and refitted coefficients are presented in Table 2.2. A plot of the fitted versus observed values of BMI in Fig. 2.1(b) shows that the model with four selected genera fits the data reasonably well except for five obese subjects.

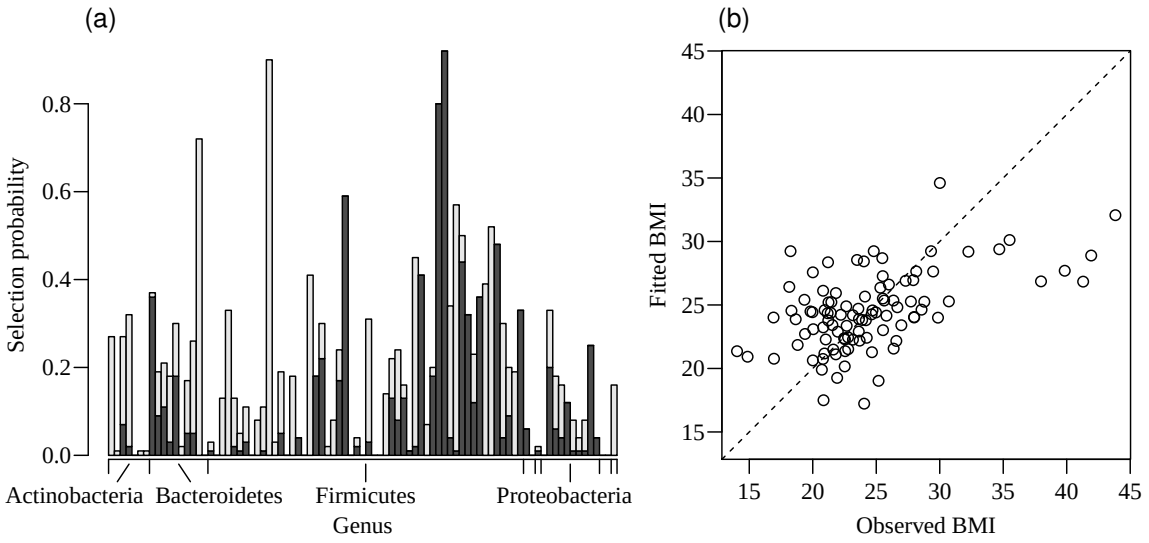


Figure 2.1: Analysis of gut microbiome data. (a) Selection probabilities with bootstrapped crossvalidation for 87 genera that belong to eight phyla. Selections with a positive sign and a negative sign are shown by dark grey blocks and light grey blocks, respectively; only four major phyla are indicated. (b) Fitted versus observed values of BMI.

Since our simulations have demonstrated that the lasso (i) estimator is inferior in all respects, we compare our method only with the lasso (ii) estimator. With selection probabilities above the cutoff value of 0·7, three genera were selected by lasso (ii) with bootstrapped crossvalidation, which coincide with three of the four previously selected genera except alistipes. To compare the prediction

Table 2.2: Selection probabilities and refitted coefficients of four selected genera in the gut microbiome data

		Selection probability		Refitted
Phylum	Genus	Boot. CV	Stab. Sel.	coefficient
Bacteroidetes	Alistipes	0.72	0.89	-0.76
Firmicutes	Clostridium	0.90	0.96	-1.35
Firmicutes	Acidaminococcus	0.80	0.92	-0.61
Firmicutes	Allisonella	0.92	0.87	-1.50

Boot. CV, bootstrapped crossvalidation; Stab. Sel., stability selection.

performance of the two methods, we randomly divided the data into a training set of 70 subjects and a test set of 28 subjects, and used the fitted model chosen by crossvalidation based on the training set to evaluate the prediction error on the test set. The prediction error averaged over 100 replicates was 30·30 for the proposed method and 30·55 for lasso (ii), with standard errors 0·97 and 1·04, respectively, suggesting that the prediction performance of the proposed method is slightly better than or similar to that of lasso (ii).

It is interesting to contrast the variable selection results at the phylum level: the proposed method selected both bacteroidetes and firmicutes as associated with BMI, whereas lasso (ii) selected only the firmicutes. Thus, our method seems more consistent with the previous finding that the relative proportion of bacteroidetes to firmicutes is decreased in obese mice and humans by comparison with lean subjects (Ley et al., 2005, 2006). One biological explanation for the finding as suggested by metagenomic and biochemical analyses is that the firmicutes-enriched microbiome holds a greater metabolic potential than the bacteroidetes-enriched microbiome for more efficient energy harvest from the diet, which in turn contributes to changes in energy balance and subsequent weight gain (Turnbaugh et al., 2006). Furthermore, our selection results at the genus level indicate that obesity may be associated with changes in gut microbiome composition at a finer taxonomic level than previously thought.

2.6. Discussion

The linear log-contrast model assumes that the absolute amounts of the covariate components have no effect on the response. We have adopted this modeling approach in the microbiome data analysis because the total amount of the microbiome cannot be reliably measured in experiments. Nevertheless, if such measurements are available, it would be worthwhile to assume a more flexible

model in which the total amount also plays a role in affecting the response. To this end, one may consider the semiparametric varying-coefficient log-contrast model

$$y_i = \beta_0(a_i) + \sum_{j=1}^{p} \beta_j(a_i) \log x_{ij} + \varepsilon_i, \quad \sum_{j=1}^{p} \beta_j(a_i) = 0,$$

with a_i being the total amounts. This reduces to model (2.2) when all the coefficients $\beta_0, \ldots, \beta_p$ are constants. A regularized estimation procedure for this model could be developed by combining the ideas of our approach and the kernel lasso method in Wang and Xia (2012).

Another possible extension of our method for microbiome data analysis would be to take into account the phylogenetic relationships among the bacterial taxa. Under the biologically plausible assumption that phylogenetically close taxa tend to have similar effects on the clinical trait, one can combine the ℓ_1 penalty in our regularization problem with a Laplacian penalty that encourages smoothness among the regression coefficients of closely related taxa on the phylogenetic tree (Chen et al., 2013). Such an extension is likely to increase the power of identifying important taxa that are relatively rare but phylogenetically close.

CHAPTER 3

REGRESSION ANALYSIS FOR MICROBIOME COMPOSITIONAL DATA

3.1. Introduction

The human microbiome includes all microorganisms in and on the human body. These microbes play important roles in human metabolism, nutrient intake and energy generation and thus are essential in human health. The gut microbiome has been shown to be associated with many human diseases such as obesity, diabetes and inflammatory bowel disease (Manichanh et al., 2012; Qin et al., 2012; Turnbaugh et al., 2006). Next generation sequencing technologies make it possible to study the microbial compositions without the need for culturing the bacterial species. There are, in general, two approaches to quantify the relative abundances of bacteria in a community. One approach is based on sequencing the 16S ribosomal RNA (rRNA) gene, which is ubiquitous in all bacterial genomes. The resulting sequencing reads provide information about the bacterial taxonomic composition. Another approach is based on shotgun metagenomic sequencing, which sequences all the microbial genomes presented in the sample, rather than just one marker gene. Both 16S rRNA and shotgun sequencing approaches provide bacterial taxonomic composition information and have been widely applied to human microbiome studies, including the Human Microbiome Project (HMP) (Turnbaugh et al., 2007) and the Metagenomics of the Human Intestinal Tract (MetaHIT) project (Qin et al., 2010).

Several methods are available for quantifying the microbial relative abundances based on the sequencing data, which typically involve aligning the reads to some known database (Segata et al., 2012). Since the DNA yielding materials are different across different samples, the resulting numbers of sequencing reads vary greatly from sample to sample. In order to make the microbial abundance comparable across samples, the abundances in read counts are usually normalized to the relative abundances of all bacteria observed. This results in high-dimensional compositional data with a unit sum. Some of the most widely used metagenomic processing softwares such as MEGAN (Huson et al., 2007) and MetaPhlAn (Segata et al., 2012) only output the relative abundances of the bacterial taxa at different taxonomic levels.

This chapter considers regression analysis of microbiome compositional data, where the goal is to

identify the bacterial taxa that are associated with a continuous response such as the body mass index (BMI). Compositional data are strictly positive and multivariate that are constrained to have a unit sum. Such data are also referred to as mixture data (Aitchison and Bacon-shone, 1984; Cornell, 2011; Snee, 1973). Regression analysis with compositional covariates needs to account for the intrinsic multivariate nature and the inherent interrelated structure of such data. For compositional data, it is impossible to alter one proportion without altering at least one of the other proportions. Linear log-contrast model (Aitchison and Bacon-shone, 1984) has been proposed for compositional data regression where logarithmic-transformed proportions are treated as covariates in a linear regression model with the constraint of the sum of the regression coefficients being zero. Lin et al. (2014) proposed a variable selection procedure for such models in high-dimensional settings and derived the weak oracle property of the resulting estimates. In analysis of microbiome data, it is also of biological interest to study the subcompositions of bacteria taxa within higher taxonomic levels, such as subcompositions of species under a given genus or phylum, or subcompositions of genera within a phylum. In subcompositional data, the proportions of species have been calculated relative to total proportions of the species under a given genus; that is, the values in the subcomposition have been re-closed to add up to 1. Regression analysis of such subcompositional data is also considered in this chapter.

One of the founding principles of compositional data analysis is that of subcompositional coherence (Aitchison, 1982): any compositional data analysis should be done in a way that we obtain the same results in a subcomposition, regardless of whether we analyze only that subcomposition or a larger composition containing other parts. This is especially relevant in high-dimensional regression analysis with compositional covariates, where the goal is to select the bacteria whose compositions are associated with the response. Once such bacteria are identified, it is desirable to recalculate the subcomposition only within those identified. However, these subcompositions have different values from those calculated based on a larger set of bacterial taxa. The log-contrast model of Aitchison and Bacon-shone (1984) and Lin et al. (2014) satisfies this principal by imposing a linear constraint on the regression coefficients. This chapter extends this model for analysis of microbiome subcompositions, where multiple linear constraints are imposed in order to achieve the subcompositional coherence.

Penalized and constrained regression, including constrained Lasso regression, has been studied

by James, Paulson, and Rusmevichientong (2015), where the regression coefficients are subject to a set of linear constraints. A computational algorithm through reformulating the problem as an unconstrained optimization problem was proposed and non-asymptotic error bounds of the estimates were derived. Different from James, Paulson, and Rusmevichientong (2015), this chapter presents an efficient computational algorithm based on the coordinate descent method of multipliers and augmented Lagrange of optimization problem. Since the resulting estimates are often biased due to ℓ_1 penalty imposed on the coefficients, variance estimation and statistical inference of the resulting estimates are difficult to derive. In order to make the statistical inference on the regression coefficients and to obtain the confidence intervals, asymptoticly unbiased estimates of the regression coefficients are first obtained through a de-biased procedure and their joint asymptotic distribution is derived. The proposed de-biased procedure extends that of Javanmard and Montanari (2014) to take into account the linear constraints on regression coefficients. However, due to the linear constraints on the regression coefficients, the theoretical developments are different from Javanmard and Montanari (2014).

Section 3.2 presents linear regression models with linear constraints for compositional covariates. Section 3.3 presents an efficient coordinate descent method of multipliers to implement the penalized estimation of the regression coefficients under linear constraints. Section 3.4 provides an algorithm to obtain de-biased estimates of the coefficients and derives their joint asymptotic distribution. Section 3.5 presents results from an analysis of gut microbiome data set in order to identify the bacterial genera that are associated with BMI. Methods are evaluated in Section 3.6 through simulations.

3.2. Regression Models for Compositional Data

3.2.1. Linear log-contrast model

Linear log-contrast model (Aitchison and Bacon-shone, 1984) has been proposed for compositional data regression. Specifically, suppose an $n \times p$ matrix $\mathbf{X}$ consists of n samples of the composition of mixture with p components, and suppose Y is a response variable depending on $\mathbf{X}$. The nature of composition makes each row of $\mathbf{X}$ lie in a $(p-1)$-dimensional positive simplex $S^{p-1} = \{(x_1, \ldots, x_p) : x_j > 0, j = 1, \ldots, p \text{ and } \sum_{j=1}^{p} x_j = 1\}$. Based on this nature, Aitchison and Bacon-

shone (1984) introduced a linear log-contrast model as follows:

$$Y = \mathbf{Z}^p \beta_{\setminus p} + \varepsilon, \tag{3.1}$$

where $\mathbf{Z}^p = \{\log(x_{ij}/x_{ip})\}$ is $n \times (p-1)$ log-ratio matrix with the pth component as the reference component, $\beta_{\setminus p} = (\beta_1, \ldots, \beta_{p-1})$ is the regression coefficient vector, and noise ε is independently distributed as $N(0, \sigma^2)$. An intercept term is not included in the model, since it can be eliminated by centering the response and predictor variables.

The selection of reference component is crucial to analysis, especially in high-dimensional settings. To avoid choosing an arbitrary reference component, Lin et al. (2014) reformulated model (3.1) as a regression problem with a linear constraint on the coefficients by letting $\beta_p = -\sum_{j=1}^{p-1} \beta_j$,

$$Y = \mathbf{Z}\beta + \varepsilon, \quad 1_p^\top \beta = 0, \tag{3.2}$$

where $1_p = (1, \ldots, 1)^\top \in \mathbb{R}^p$, $\mathbf{Z} = (z_1, \ldots, z_p) = (\log x_{ij}) \in \mathbb{R}^{n \times p}$, and $\beta = (\beta_1, \ldots, \beta_p)^\top$.

3.2.2. Subcompositional regression model

In analysis of microbiome data, the relative abundances of taxa are often obtained at different taxonomic ranks, including species, genus, family, class and phylum. It is of interest to study whether the composition of taxa that belong to a given taxon at a higher rank is associated with the response, in which case subcompositions of taxa (e.g., all the genera that belong to a given phylum) are calculated. Suppose r taxa at a given rank are considered with m_g taxa at the lower rank that belong to taxon g. Let X_{gs} be the relative abundance of the sth taxon that belong to the gth taxon at a higher rank, for $g = 1, \cdots, r$, $s = 1, \cdots, m_g$ such that

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

Let $n \times m_g$ matrix $\mathbf{X}_g$ represents n samples of the subcomposition of m_g taxa. The following model can be used to link the subcompositions to a response Y,

$$Y = \sum_{g=1}^{r} \mathbf{Z}_g \beta_g + \varepsilon, \tag{3.3}$$

where $\mathbf{Z}_g = (Z_{g1}, \ldots, Z_{gm_g}) = (\log X_{g1}, \ldots, \log X_{gm_g}) \in \mathbb{R}^{n \times m_g}$, and $\beta_g = (\beta_{g1}, \cdots, \beta_{gm_g})^\top$. To make the model subcompositional coherence, the following r linear constraints are imposed,

$$\mathbf{1}_{m_g}^\top \beta_g = \sum_{s=1}^{m_g} \beta_{gs} = 0 \text{ for } g = 1 \cdots, r.$$

This set of linear constraints can be written as $\mathbf{C}^\top \beta = 0$, where $\beta = (\beta_1^\top, \cdots, \beta_r^\top)^\top$, and

$$\mathbf{C}^\top = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}_{r \times p}$$

Models (3.2) and (3.3) belong to a more general high-dimensional linear model with r linear constraints on the coefficients,

$$Y = \mathbf{Z}\beta + \varepsilon, \quad \mathbf{C}^\top \beta = 0, \tag{3.4}$$

where the rows of $\mathbf{Z} \in \mathbb{R}^p$ are independently and identically distributed with mean zero, $\mathbf{C}$ is a $p \times r$ matrix of the constraint coefficients, $\beta = (\beta_1, \ldots, \beta_p)^\top$, and $\varepsilon \sim N_n(0, \sigma^2 \mathbf{I})$. Without loss of generality, $\mathbf{C} = (c_1, \ldots, c_r)$ is assumed to be orthonormal. In high-dimensional settings, β is assumed to be s-sparse, where $s = \#\{i : \beta_i \neq 0\}$ and $s = o(\sqrt{n}/\log p)$.

This chapter considers estimation and inference of Model (3.4) under the general linear constraints. Lin et al. (2014) proposed a procedure for variable selection and estimation for Model (3.2) and derived the weak oracle property of the resulting estimates. James, Paulson, and Rusmevichientong (2015) considered a more general model and provided non-asymptotic bounds on estimation errors. However, variances of the estimates and statistical inference are lacking. In this chapter, an algorithm to perform variable selection for Model (3.4) based on ℓ_1 penalized estimation is first proposed based on coordinate descent method of multipliers. An inference procedure for the penalized estimator of the regression coefficients is then introduced. The proposed approach parallels to that of Javanmard and Montanari (2014) by first obtaining de-biased estimates of the coefficients for high-dimensional linear model with linear constraints, $\hat{\beta}^u$, which are shown to be asymptotically Gaussian, with mean β and covariance $\sigma^2(\widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}})/n$, where $\widehat{\Sigma}$ is the empirical covariance and $\widetilde{\mathbf{M}}$

is determined by solving a convex program. Based on this asymptotic result, the corresponding confidence intervals and p-values are constructed and used for statistical inference.

3.3. Penalized Estimation

In this following presentation, for a matrix $\mathbf{A}_{m\times n}$, $||\mathbf{A}||_p$ is the ℓ_p operator norm defined as $||\mathbf{A}||_p = \sup_{||x||_p=1} ||\mathbf{A}x||_p$, where $||v||_p$ is the standard ℓ_p norm of a vector v. In particular, $||\mathbf{A}||_\infty = \max_{1\le i\le m}\sum_{j=1}^n |a_{ij}|$. We also define $|\mathbf{A}|_\infty = \max_{i,j}|a_{ij}|$.

Consider model (3.4). Define $\mathbf{P_C} = \mathbf{C}\mathbf{C}^\top$ as the projection onto the space spanned by the columns of $\mathbf{C}$. Two basic regularity conditions on $\mathbf{C}$ are assumed:

Condition 2. $||\mathbf{I}_p - \mathbf{P_C}||_\infty \le k_0$ *for a constant* k_0 *that is free of* p*.*

Condition 3. *The diagonal elements of* $\mathbf{I}_p - \mathbf{P_C}$ *are greater than zero.*

Condition 1 is equivalent to that $||c_j||_1||c_j||_\infty, j = 1,\dots,r$ are all bounded by a constant that is free of p. Condition 2 means that the group of constraints do not indicate simple constraint such as $\beta_j = 0$. If $(\mathbf{I}_p - \mathbf{P_C})_{j,j} = 0$, then the j^{th} row and column of $\mathbf{I}_p - \mathbf{P_C}$ are all zeros, and thus $(\mathbf{I}_p - \mathbf{P_C})e_j = 0$, which means that e_j lies in the space spanned by the columns of $\mathbf{C}$. It is easy to verify that the constraint matrix $\mathbf{C}$ in the log-contrast model (3.2) or the subcompositional model (3.3) satisfies both conditions. For example, in the log-contrast model (3.2), $k_0 = 2$ for $\mathbf{C} = 1_p/\sqrt{p}$ since

$$||(\mathbf{I}_p - 1_p 1_p^\top/p)a||_\infty = ||a - \frac{1}{p}\sum_{j=1}^p a_j 1||_\infty \le ||a||_\infty + |\frac{1}{p}\sum_{j=1}^p a_j| \le 2||a||_\infty.$$

Define $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - \mathbf{P_C})$. Since $\mathbf{P_C}\beta = 0$, model (3.4) can be rewritten as

$$Y = \widetilde{\mathbf{Z}}\beta + \varepsilon, \quad \mathbf{C}^\top\beta = 0. \tag{3.5}$$

The regression coefficients can be estimated using ℓ_1 penalized estimation with linear constraints,

$$\widehat{\beta}^n = \underset{\beta}{\operatorname{argmin}}\left(\frac{1}{2n}||Y - \widetilde{\mathbf{Z}}\beta||_2^2 + \lambda||\beta||_1\right) \text{ subject to } \mathbf{C}^\top\beta = 0, \tag{3.6}$$

where λ is a tuning parameter.

A coordinate descent method of multipliers can be used to implement the constrained optimization

problem (3.6). First, the augmented Lagrange of optimization problem (3.6) (Bertsekas, 2014) is formed as,

$$L_\mu(\beta,\eta) = \frac{1}{2n}||y-\widetilde{\mathbf{Z}}\beta||_2^2 + \lambda||\beta||_1 + \eta^\top \mathbf{C}^\top\beta + \frac{\mu}{2}||\mathbf{C}^\top\beta||_2^2,$$

where $\eta \in \mathbb{R}^r$ is the Lagrange multiplier, and $\mu > 0$ is a penalty parameter. Problem (3.6) can be solved by iterations

$$\beta^{k+1} \leftarrow \underset{\beta}{\operatorname{argmin}} L_\mu(\beta,\eta^k), \quad \eta^{k+1} \leftarrow \eta^k + \mu\mathbf{C}^\top\beta^{k+1}.$$

Define $\xi = \eta/\mu$, the iterations become

$$\beta^{k+1} \leftarrow \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n}||y-\widetilde{\mathbf{Z}}\beta||_2^2 + \lambda||\beta||_1 + \frac{\mu}{2}||\mathbf{C}^\top\beta + \xi^k||_2^2 \right\}, \tag{3.7}$$

$$\xi^{k+1} \leftarrow \xi^k + \mathbf{C}^\top\beta^{k+1}. \tag{3.8}$$

The iteration of β can be further detailed as

$$\beta_j^{k+1} \leftarrow \frac{1}{\frac{||\tilde{z}_j||_2^2}{n} + \mu||C_j||_2^2} S_\lambda \left[\frac{1}{n}\tilde{z}_j^\top (y - \sum_{i\neq j}\beta_i^{k+1}\tilde{z}_i) - \mu(\sum_{i\neq j}\beta_i^{k+1}C_i^\top C_j + C_j^\top\xi^k) \right], \tag{3.9}$$

where $C_i, i = 1,\ldots,p$ are the rows of $\mathbf{C}$, $\tilde{z}_i, i = 1,\ldots,p$ are columns of $\widetilde{\mathbf{Z}}$, and $S_\lambda(t) = \text{sgn}(t)(|t|-\lambda)_+$. Combining (3.7)-(3.9) yields the following algorithm for solving problem (3.6).

Input: Y, $\widetilde{\mathbf{Z}}$, and λ.
Output: $\widehat{\beta}^n$

1: Initialize β^0 with 0 or a warm start, $\xi^0 = 0$, $\mu > 0$ and $k = 0$.
2: For $j = 1,\ldots,p,1,\ldots,p,\ldots$, update β_j^{k+1} by (3.9) until convergence.
3: Update ξ^{k+1} by (3.8).
4: $k \leftarrow k+1$ and repeat the two steps above until convergence.

Algorithm 2: *Coordinate descent method of multipliers for solving problem* (3.6)

The penalty parameter μ that is needed to enforce the zero-sum constraints does not affect the convergence of Algorithm 1 as long as $\mu > 0$. It can however affect the convergence rate of the algorithm. In this chapter, $\mu = 1$ is taken in all the computations.

3.4. A De-biased Estimator and Its Asymptotic Distribution

3.4.1. A De-biased estimator

The asymptotic distribution of ℓ_1 regularized estimator $\widehat{\beta}^n$ is not manageable and $\widehat{\beta}^n$ is biased due to regularization. Javanmard and Montanari (2014) proposed a procedure to construct a de-biased version of the unconstrained LASSO estimator that has a tractable asymptotic distribution, which can be used to obtain the confidence intervals of the regression coefficients. Similar de-biased procedures were also developed by Zhang and Zhang (2014) and Bühlmann et al. (2013).

Adapting the de-biased procedure of Javanmard and Montanari (2014), the following algorithm can be used to obtain de-biased estimates of the regression coefficients, $\widehat{\beta}^u$.

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

Let $\widehat{\beta}^n$ be the regularized estimator from optimization problem (3.6).
Set $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - \mathbf{P}_\mathbf{C})$.
Set $\widehat{\Sigma} \equiv (\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}})/n$.
for $i = 1, 2, \ldots, p$ **do**:
Let m_i be a solution of the convex program:

$$\begin{aligned} &\text{minimize } m^\top \widehat{\Sigma} m \\ &\text{subject to } ||\widehat{\Sigma} m - (\mathbf{I}_p - \mathbf{P}_\mathbf{C}) e_i||_\infty \leq \gamma. \end{aligned} \tag{3.10}$$

end for
Set $\mathbf{M} = (m_1, \ldots, m_p)^\top$, set

$$\widetilde{\mathbf{M}} = (\mathbf{I}_p - \mathbf{P}_\mathbf{C}) \mathbf{M} (\mathbf{I}_p - \mathbf{P}_\mathbf{C}). \tag{3.11}$$

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

$$\widehat{\beta}^u = \widehat{\beta}^n + \frac{1}{n} \widetilde{\mathbf{M}} \widetilde{\mathbf{Z}}^\top (Y - \widetilde{\mathbf{Z}} \widehat{\beta}^n). \tag{3.12}$$

Algorithm 3: *Constructing a de-biased estimator*

To solve problem (3.10), Matlab package CVX is used for specifying and solving convex programs (Grant and Boyd, 2013). To briefly explain the logic behind this algorithm, denote $\Sigma = \mathbb{E}\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}}/n$, and suppose that $\Sigma = \mathbf{V} \Lambda \mathbf{V}^\top$ is the eigenvalue/eigenvector decomposition of Σ, where

$\Lambda = diag(\lambda_1, \ldots, \lambda_{p-r})$. Note that $(\mathbf{V}, \mathbf{C})$ is full rank and orthonormal, and

$$\Sigma = (\mathbf{V}, \mathbf{C}) \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{V}, \mathbf{C})^\top.$$

Define

$$\Omega = (\mathbf{V}, \mathbf{C}) \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{V}, \mathbf{C})^\top,$$

then

$$\Sigma\Omega = (\mathbf{V}, \mathbf{C}) \begin{pmatrix} \mathbf{I}_{p-r} & 0 \\ 0 & 0 \end{pmatrix} (\mathbf{V}, \mathbf{C})^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}_p - \mathbf{P}_\mathbf{C},$$

where Ω is the inverse of Σ in the perpendicular space of the column space of $\mathbf{C}$. The de-biased algorithm first finds an approximation of Ω by rows, denoted by $\widetilde{\mathbf{M}}$, and then corrects the bias based on $\widetilde{\mathbf{M}}$. At the last step of this algorithm, $\widehat{\beta}^u$ is the de-biased version of $\widehat{\beta}^n$. It is easy to check that $\mathbf{C}^\top \widehat{\beta}^u = 0$, which is guaranteed by (3.11).

The feasibility of the optimization (3.10) is presented in Lemma 1 under the following assumptions on matrix $\widetilde{\mathbf{Z}} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_n)^\top$:

Condition 4. *There exist uniform constants $C_{\min}, C_{\max}$ such that $0 < C_{\min} \leq \sigma_{\min}(\Sigma) \leq \sigma_{\max}(\Sigma) \leq C_{\max} < \infty$, where $\sigma_{\max}(\mathbf{A})(\sigma_{\min}(\mathbf{A}))$ is the largest (smallest) non-zero eigenvalue of matrix $\mathbf{A}$.*

Condition 5. *There exists a uniform constant $\kappa \in (0, \infty)$ such that the rows of $\widetilde{\mathbf{Z}}\Omega^{1/2}$ are sub-Gaussian with $||\Omega^{1/2}\widetilde{Z}_1||_{\psi_2} \leq \kappa$, where the sub-Gaussian norm of a random vector $Z \in \mathbb{R}^n$ is defined as*

$$||Z||_{\psi_2} = \sup\{||Z^\top x||_{\psi_2} : x \in \mathbb{R}^n, ||x||_2 = 1\},$$

with $||X||_{\psi_2}$ defined as $||X||_{\psi_2} = \sup_{q \geq 1} q^{-1/2}(\mathbb{E}|X|^q)^{1/q}$ for a random variable X.

These two conditions are imposed on $\widetilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{I}_p - \mathbf{P}_\mathbf{C})$, not on the original log-ratio matrix $\mathbf{Z}$. For the subcompositional model (3.3), it is easy to see that $\widetilde{\mathbf{Z}}$ is the matrix of the centered log-ratio (CLR) transformation of the original taxonomic composition (Aitchison, 1982), where

$$\widetilde{\mathbf{Z}}_{gs} = \log \frac{X_{gs}}{\sqrt[m_g]{\prod_{s=1}^{m_g} X_{gs}}}.$$

CLR has been shown to be effective in transforming compositional data to approximately multivariate normal in many real compositional and microbiome data (Aitchison, 1982; Kurtz et al., 2015). Conditions 4 and 5 are therefore reasonable assumptions in our setting.

The following Lemma shows that if $\gamma = c\sqrt{\log p/n}$ in Algorithm 2 is properly chosen, Ω is in the feasible set of the optimization problem (3.10) with a large probability.

Lemma 1. *Let $\widehat{\Sigma} \equiv (\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}})/n$ be the empirical covariance. For any constant $c > 0$, the following holds true,*

$$\mathbb{P}\left\{ \left| \Omega\widehat{\Sigma} - (\mathbf{I}_p - \mathbf{P}_\mathbf{C}) \right|_\infty \geq c\sqrt{\frac{\log p}{n}} \right\} \leq 2p^{-c''},$$

where $c'' = (c^2 C_{\min})/(24e^2\kappa^4 C_{\max}) - 2$.

3.4.2. Asymptotic distribution and inference

To obtain the asymptotic distribution of the de-biased estimator $\widehat{\beta}^u$, an additional assumption on $\widetilde{\mathbf{Z}}$ is required.

Condition 6. *The inequality $(3\tau-1)\delta_{2s}^-(\widetilde{\mathbf{Z}}/\sqrt{n}) - (\tau+1)\delta_{2s}^+(\widetilde{\mathbf{Z}}/\sqrt{n}) \geq 4\tau\phi_0$ holds for a constant $\phi_0 > 0$, where for any matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\delta_k^+(\mathbf{A})$ and $\delta_k^-(\mathbf{A})$ are the upper and lower restricted isometry property (RIP) constants of order k defined as*

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

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

Condition 6 means that $\delta_{2s}^-(\widetilde{\mathbf{Z}}/\sqrt{n})$ and $\delta_{2s}^+(\widetilde{\mathbf{Z}}/\sqrt{n})$ should be close, that is, any $2s$ columns of the CLR transformed compositional data matrix $\widetilde{\mathbf{Z}}/\sqrt{n}$ should be close to orthonormal.

The following theorem gives the asymptotic distribution of the de-biased estimates of the regression coefficients.

Theorem 2. *Consider the linear model* (3.5) *with β as an s-sparse vector, and let $\widehat{\beta}^u$ be defined as in equation* (3.12) *in Algorithm 3. Then,*

$$\sqrt{n}(\widehat{\beta}^u - \beta) = B + \Delta, \quad B|\mathbf{Z} \sim N(0, \sigma^2 \widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}}^\top), \quad \Delta = \sqrt{n}(\widetilde{\mathbf{M}}\widehat{\Sigma} - (\mathbf{I}_p - \mathbf{P}_\mathbf{C}))(\beta - \widehat{\beta}^n).$$

Further, assume the Conditions (2)-(6) *hold. Then setting* $\lambda = r\tilde{c}\sigma\sqrt{(\log p)/n}$ *in optimization problem* (3.6) *and* $\gamma = c\sqrt{(\log p)/n}$ *in Algorithm 3, the following holds true:*

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

where $K = \max_i\sqrt{\widehat{\Sigma}_{i,i}}$ *and constants* c' *and* c'' *are given by*

$$c' = \frac{\tilde{c}^2}{2K^2}-1, \quad c'' = \frac{c^2 C_{\min}}{24e^2\kappa^4 C_{\max}}-2.$$

Theorem 2 says that $N(0,\sigma^2\widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}}^\top)$ can be used to approximate the distribution of $\widehat{\beta}^u$ with proper choices of c and $\tilde{c}$ (or equivalently γ and λ). This leads to the following corollary that can be used to construct asymptotic confidence intervals and p-values for β in high-dimensional linear model with linear constraints (3.4).

Corollary 1. *Let* $\widehat{\sigma}$ *be a consistent estimator of* σ*.*

(1) Define $\delta_i(\alpha,n) = \Phi^{-1}(1-\alpha/2)\widehat{\sigma}n^{-1/2}[\widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}}^\top]_{i,i}^{1/2}$.
Then $I_i = [\widehat{\beta}_i^u - \delta_i(\alpha,n), \widehat{\beta}_i^u + \delta_i(\alpha,n)]$ *is an asymptotic two-sided level* $1-\alpha$ *confidence interval for* β_i*.*

(2) For individual hypothesis $H_{0,i}:\beta_i = 0$ *versus* $H_{0,i}:\beta_i \neq 0$*, an asymptotic p-value can be constructed as follows:*

$$P_i = 2\left[1-\Phi\left(\frac{n^{1/2}|\widehat{\beta}_i^u|}{\widehat{\sigma}[\widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}}^\top]_{i,i}^{1/2}}\right)\right].$$

The following lemma shows that with Condition 2, the diagonal elements of $\widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}}^\top$ are nonzero with a γ that is not too large.

Lemma 2. *Let* $\widetilde{\mathbf{M}}$ *be the matrix obtained by equation* (3.11)*. Then for* $\gamma < (1-(\mathbf{P}_\mathbf{C})_{i,i})/k_0$ *and all* $i = 1,\ldots,p$*,*

$$[\widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}}^\top]_{i,i} \ge \frac{(1-(\mathbf{P}_\mathbf{C})_{i,i}-k_0\gamma)^2}{\widehat{\Sigma}_{i,i}}.$$

3.4.3. Selection of the tuning parameters

In real applications, the estimator $\widehat{\beta}^n$, tuning parameter λ and estimation of noise level $\widehat{\sigma}$ are obtained through scaled LASSO (Sun and Zhang, 2012). Specifically, the following two steps are iterated until convergence:

$$\widehat{\beta}^n \leftarrow \underset{\mathbf{C}^\top \beta = 0}{\operatorname{argmin}} \left\{ ||Y - \widetilde{\mathbf{Z}}\beta||_2^2 + 2n\lambda_0 \widehat{\sigma} ||\beta||_1 \right\},$$

$$\widehat{\sigma}^2 \leftarrow ||Y - \widetilde{\mathbf{Z}}\widehat{\beta}||_2^2 / n,$$

where $\lambda_0 = \sqrt{2} L_n(k/p)$, $L_n(t) = n^{-1/2}\Phi^{-1}(1-t)$, Φ^{-1} is the quantile function for standard normal and k is the solution of $k = L_1^4(k/p) + 2L_1^2(k/p)$. Then $\widehat{\lambda} = \lambda_0 \widehat{\sigma}$, and $\gamma = a\hat{\lambda}/\hat{\sigma}$ are used in Algorithm 2, where $a = 1/3$ is used in all simulations and real data analysis in this chapter.

3.5. Association Between Body Mass Index and Gut Microbiome

Gut microbiome plays an important role in food digestion and nutrition absorption. Wu et al. (2011) reported a cross-sectional study to examine the relationship between micronutrients and gut microbiome composition, where the fecal samples of 98 healthy volunteers from the University of Pennsylvania were collected, together with demographic data such as body mass index, age and sex. The DNAs from the fecal samples were analyzed by 454/Roche pyrosequencing of 16S rRNA gene segments of the V1-V2 region. After the pyrosequences were denoised, a total of about 900,000 16S reads were obtained with an average of 9165 reads per sample and 3068 operational taxonomic units (OTUs) were obtained. These OTUs were combined into 87 genera that appeared in at least one sample. Out of these 87 genera, 42 genera have zero counts in more than 90% of the samples and were removed from our analysis. The remaining 45 relatively common genera belong to four phyla, *Actinobacteria, Bacteroidetes, Firmicutes* and *Proteobacteria*. Since dysbiosis of gut microbiome has been shown to be associated with obesity (Ley et al., 2005, 2006; Turnbaugh et al., 2006), it is interesting to identify the bacterial genera that are associated with BMI after adjusting for total fat and caloric intakes. In the following analysis, zero count was replaced by the maximum rounding error of 0.5, commonly used in compositional and microbiome data analysis (Aitchison, 2003; Kurtz et al., 2015). Since the number of reads is very large, replacing zero with other very small counts does not affect our results. These read counts are then converted into compositions

of the genera or subcompositions of the genera within phylum.

3.5.1. Analysis of the data at the genus-level

The proposed method was first applied to perform regression analysis with BMI as the response and the log-transformed compositions of the 45 genera as the covariates. In addition, total fat intake and total caloric intake were also included as the covariates in the model. The model was fit with the constraint that the sum of the coefficients corresponding to the 45 genera is zero, assuming

$$E(\text{BMI}) = \sum_{g=1}^{45} \beta_g \log(X_g) + \gamma_1 \text{FAT} + \gamma_2 \text{CALORIE},$$

where $\sum_{g=1}^{45} \beta_g = 0$, and $\log(X_g)$ is the logarithm of the relative abundance of the gth genus. The goal of this analysis is to identify the bacteria genera that are associated with BMI.

Figure 3.1 shows the estimated regression coefficients from LASSO with one constraint and their de-biased estimates together with the 95% confidence intervals of the regression coefficients. Four genera were statistically significant with p-value of 0.0251 for *Alistipes*, 0.0031 for *Clostridium*, 0.0031 for *Acidaminococcus*, and 0.0042 for *Allisonella*, respectively. These four genera were exactly the same genera identified using stability selection by Lin et al. (2014). They belong to two bacterial phyla, *Bacteroidetes* and *Firmicutes*. The results indicate that Alistipes in the Bacteroidetes phylum is negatively associated with BMI, which is consistent with previous findings that the gut microbiota in obese mice and humans tend to have a lower proportion of *Bacteroidetes* (Ley et al., 2005, 2006; Turnbaugh et al., 2006). However, for the *Firmucutes* phylum, both the positively associated (*Acidaminococcus* and *Allisonella*) and negatively associated (*Clostridium*) genera were observed to be associated with BMI, suggesting that obesity may be associated with changes in gut microbiome composition at a lower taxonomic level than previously thought.

3.5.2. Subcomposition analysis

The proposed method was then applied to subcomposition analysis, where the number of sequencing reads were converted into compositions of genera within each phylum. This creates four subcompositions of the genera within four phyla. This analysis aims to answer the question whether the composition of genera within a given phylum is associated with BMI, where the log-transformed genera subcompositions are treated as predictors, together with total fat and caloric intakes as

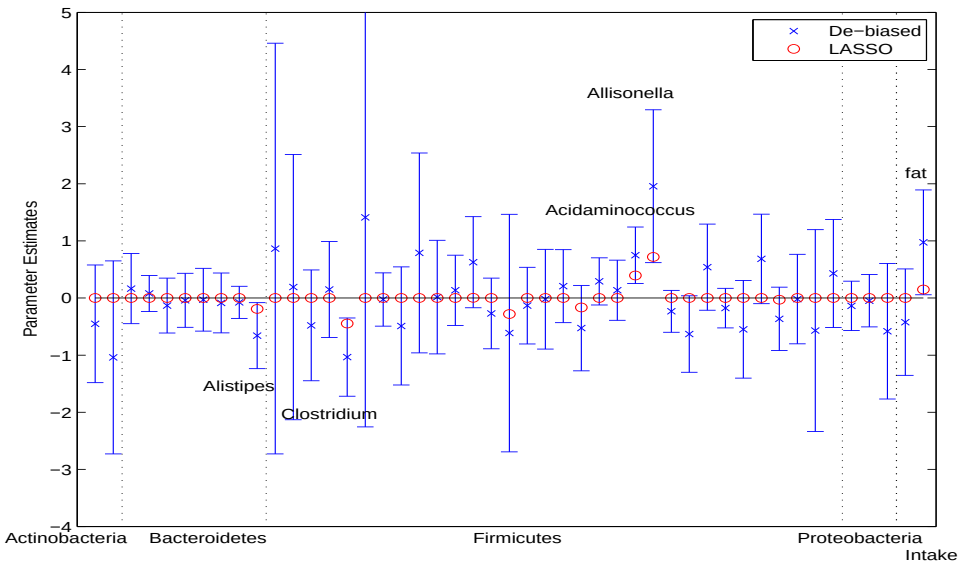


Figure 3.1: Analysis of gut microbiome data. Lasso estimates, de-biased estimates and 95% confidence intervals of the regression coefficients in the model treating the composition of 45 genera as covariates together with total fat and caloric intakes. Dashed vertical lines separate bacterial genus into different phyla.

covariates in the following model,

$$E(\text{BMI}) = \sum_{g=1}^{4} \sum_{s=1}^{m_g} \beta_{gs} \log(X_{gs}) + \gamma_1 \text{FAT} + \gamma_2 \text{CALORIE},$$

where $\sum_{s=1}^{m_g} \beta_{gs} = 0$ for $g = 1, \cdots, 4$, and $\log(X_{gs})$ is the logarithm of the relative abundance of the sth genus of the gth phylum.

Figure 3.2 shows the LASSO estimates, de-biased estimates, and 95% confidence interval of the coefficients of the 45 genera. Four genera were statistically significant with p-value of 0.0036 for *Clostridium*, 0.0056 for *Acidaminococcus*, 0.0116 for *Allisonella*, and 0.0111 for *Oscillibactor*. All four genera belong to phylum *Firmicutes*, indicating that the subcomposition of the bacterial genera within *Firmicutes* is associated with BMI. The genus *Alistipes* has a p-value of 0.0523 in this analysis, which is marginally significant. It is interesting that the bacterial genus *Oscillibactor* was identified as one of the two bacterial genera that are negatively associated with BMI. *Oscillibacter* was observed to be increased on the resistent starch and reduced carbohydrate weight loss diets (Walker et al., 2011) in a strictly diet-controlled experiments in obese men, which may explain its negative association with BMI. A recent study also identified *Oscillibacter*-like organisms as a

potentially important gut microbe that mediates high fat diet-induced gut dysfunction (Lam et al., 2012). It is possible that *Oscillibacter* directly regulates components involved in the maintenance of gut barrier integrity.

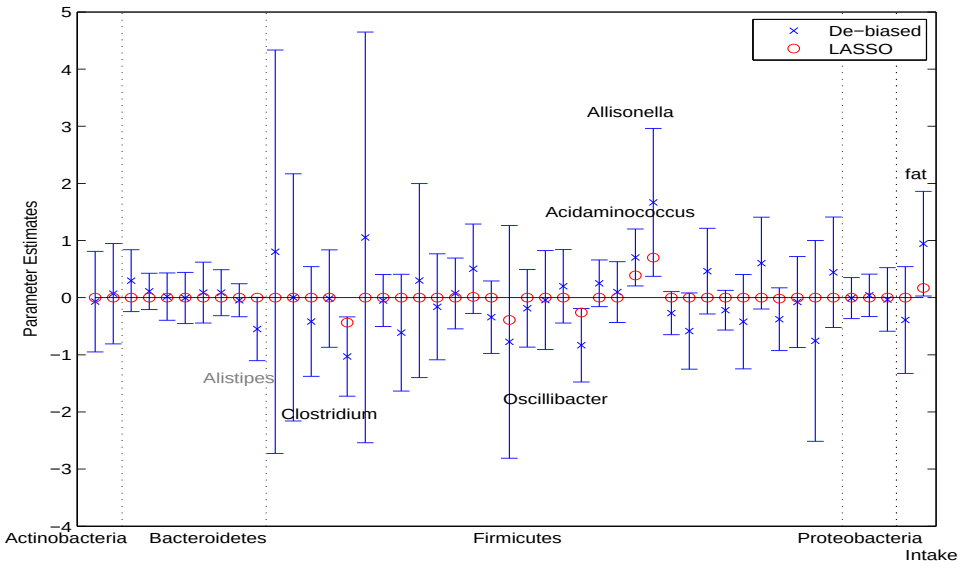


Figure 3.2: Analysis of gut microbiome data. Lasso estimates, de-biased estimates and 95% confidence intervals of the regression coefficients in the model treating the subcompositions of the genera in each phylum as covariates together with total fat and caloric intakes. Dashed vertical lines separate bacterial genus into different phyla.

Figure 3.3 shows the predicted BMI using leave-one-out cross-validation (LOOCV). In each round of LOOCV, the variables were selected based on the estimated 95% confidence intervals and the prediction was performed using refitted coefficients of the selected bacterial genera, together with calorie and fat intakes. An $R^2 = 0.1576$ was obtained between the observed and predicted values. As a comparison, fitting the model with one linear constraint at the genus-level resulted a $R^2 = 0.1361$ based on LOOCV, indicating some gain in prediction by the subcompositional analysis.

3.6. Simulation Evaluation and Comparisons

In order to simulate the compositional covariates, a $n \times p$ matrix W of taxon counts is first generated with each row of W being generated from a log-normal distribution $lnN(\nu, \Sigma)$, where $\Sigma_{ij} = \zeta^{|i-j|}$ with ζ=0.2 or 0.5 is the covariance matrix to reflect different levels of correlation between the taxa counts. Parameters $\nu_j = p/2$ for $j = 1, \ldots, 5$ and $\nu_j = 1$ for $j = 6, \ldots, p$ are set to allow some taxa to be much more abundant than others, as often observed in real microbiome compositional data.

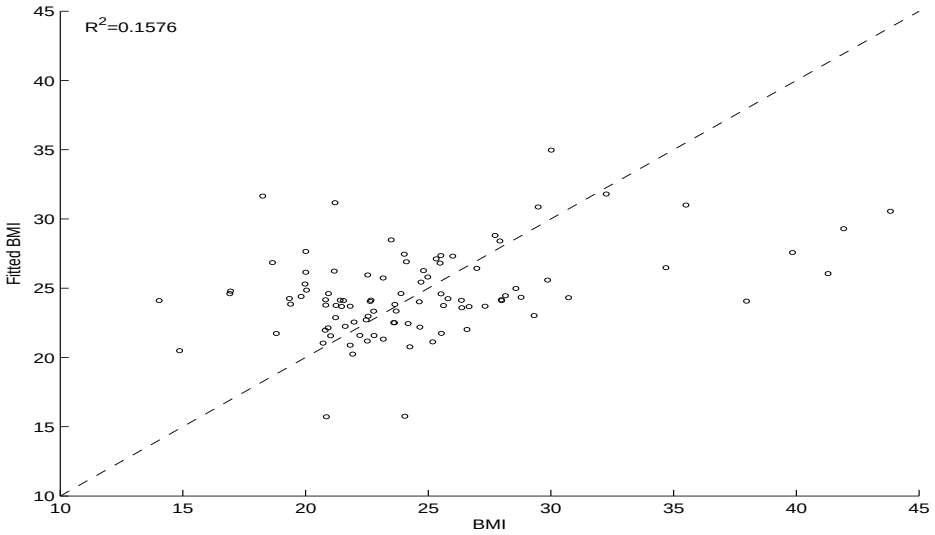


Figure 3.3: Analysis of gut microbiome data. Observed and predicted BMI using LOOCV and variables selected based on 95% confidence intervals, together with total fat and caloric intakes.

The compositional covariate matrix $\mathbf{Z}$ is obtained by normalizing the simulated taxa counts as

$$z_{ij} = \log\left(\frac{w_{ij}}{\sum_{k=1}^{p} w_{ik}}\right), i = 1, \cdots, n, j = 1, \ldots, p$$

. Based on these compositional covariates, the response Y is generated through Model (3.2) with

$$\beta = (1, -0.8, 0.4, 0, 0, -0.6, 0, 0, 0, 0, -1.5, 0, 1.2, 0, 0, 0.3, 0, \ldots, 0)$$

and σ = 0.5. Different dimension/sample size combinations (p,n)=(50,100), (50,200), (50,500), (100,100), (100,200), (100,500) are considered and the simulations are repeated 100 times for each setting. The tuning parameters are chosen using the method described in Section 3.4.3. The regression coefficient β used in the simulation satisfies the following 8 linear constraints

$$\begin{aligned} &\sum_{j=1}^{10} \beta_j = 0, \sum_{j=11}^{16} \beta_j = 0, \sum_{j=17}^{20} \beta_j = 0, \sum_{j=21}^{23} \beta_j = 0, \\ &\sum_{j=24}^{30} \beta_j = 0, \sum_{j=31}^{32} \beta_j = 0, \sum_{j=33}^{40} \beta_j = 0, \sum_{i=41}^{p} \beta_j = 0. \end{aligned} \tag{3.13}$$

3.6.1. *Estimation of confidence intervals*

The model is first fitted under the correct constraints specified in (3.13) and the corresponding confidence intervals are obtained based on our asymptotic results. Figure 3.4 shows the coverage probability for various models and samples sizes, indicating that the coverage probabilities of the confidence intervals are close to the nominal level of 0.95 when the sample size is large. For small sample sizes, the empirical coverage probability is slightly greater than the nominal level of 0.95, indicating some conservativeness. Figure 3.5 shows the lengths of confidence intervals. As expected, larger sample sizes result in shorter lengths and larger correlations among the variables lead to increased length of the confidence intervals.

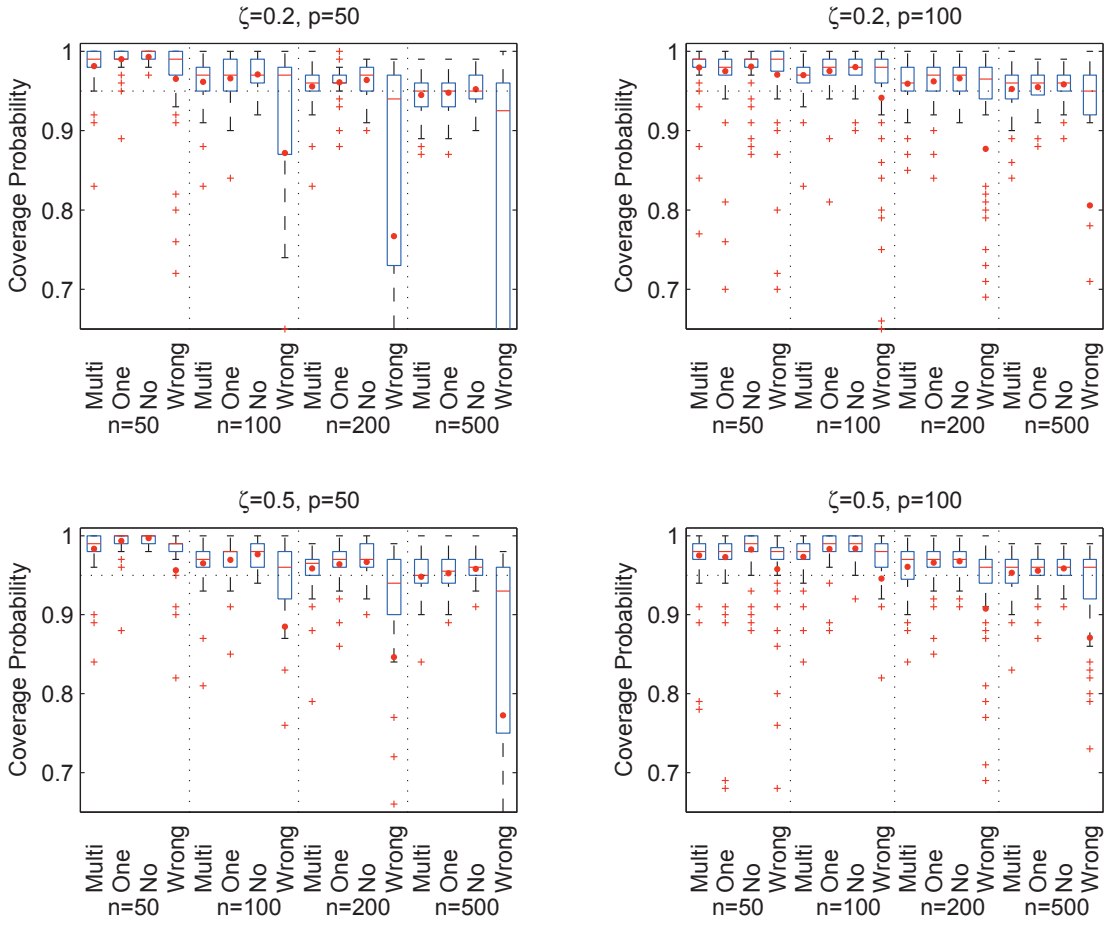


Figure 3.4: Coverage probabilities of confidence intervals based on 100 replications. For each model, minimum, median (in red line), mean (in red dot) and maximum of the coverage probabilities over compositional covariates are shown. The confidence intervals are constructed using multiple, one, no and wrong linear constraints, labeled by 'Multi', 'One', 'No' and 'Wrong' respectively.

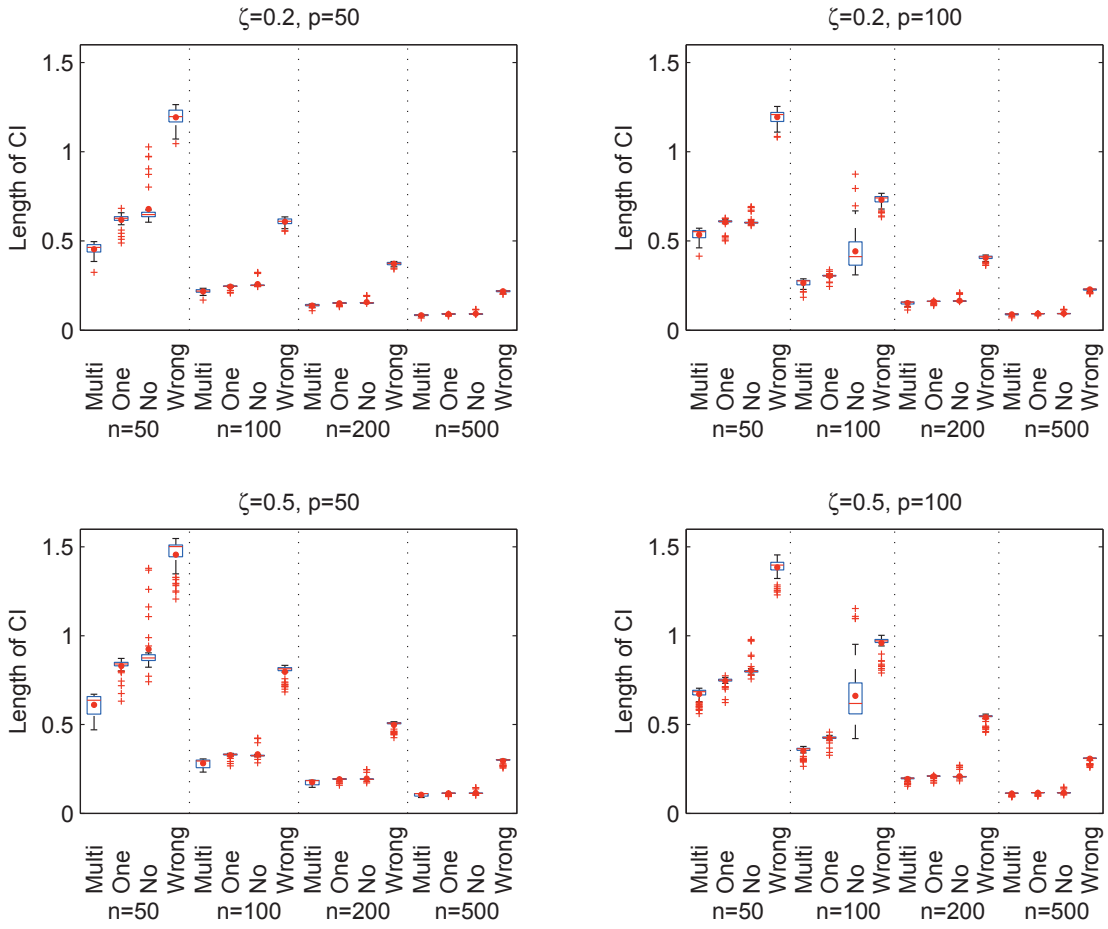


Figure 3.5: Average lengths of confidence intervals based on 100 replications. For each model, minimum, median (in red line), mean (in red dot) and maximum of the lengths of the intervals overall all compositional covariates are shown. The confidence intervals are constructed using multiple, one, no and wrong linear constraints, labeled by 'Multi', 'One', 'No' and 'Wrong' respectively.

As comparisons, the model is also fitted under no constraint, one single constraint, $\sum_{j=1}^{p}\beta_j = 0$, and misspecified constraints,

$$\sum_{j=1}^{5}\beta_j = 0, \sum_{j=6}^{12}\beta_j = 0, \sum_{j=13}^{23}\beta_j = 0, \sum_{j=24}^{30}\beta_j = 0, \sum_{j=31}^{p}\beta_j = 0.$$

The coverage probabilities and the lengths of the confidence intervals are given in Figure 3.4 and Figure 3.5, respectively. While the coverage probabilities are relatively less sensitive to such misspecification, the intervals estimated under the correct linear constraints are much shorter than those obtained with one or none of the linear constraints, especially when sample size is small. Using the wrong constraints results in much longer intervals with less accurate coverage.

3.6.2. Variable selection based on the confidence intervals

The confidence intervals of the regression coefficients can also be applied to choose the variables of interest. For example, a variable can be selected if the nominal 95% confidence interval of the corresponding regression coefficient includes zero. Table 3.1 shows the true positive rate and false positive rate of the variables identified based on 95% confidence intervals under multiple constraints, one single constraint and no constraint. When the sample size is small, imposing the correct linear constraints can lead to more true discoveries while the false positive rates are still controlled under 5%. In contrast, the models with only one or no constraint lead to much lower true positive rates and the standard LASSO without any constraint gives the worst variable selection results.

Table 3.1: True/False positive rates of the significant variables selected based on 95% confidence intervals constructed using multiple, one and no linear constraints, labeled by 'Multi', 'One' and 'No' respectively. Variable correlations ζ, numbers of variables p and sample sizes (n) are considered.

Configuration			True Positive Rate			False Positive Rate		
			Constraints			Constraints		
ζ	p	n	Multi	One	No	Multi	One	No
0.2	50	50	0.9329	0.8514	0.7586	0.0121	0.0056	0.0051
		100	1.0000	1.0000	0.9957	0.0330	0.0286	0.0267
		200	1.0000	1.0000	1.0000	0.0386	0.0333	0.0328
		500	1.0000	1.0000	1.0000	0.0498	0.0477	0.0470
0.2	100	50	0.8571	0.8071	0.7700	0.0131	0.0166	0.0139
		100	1.0000	0.9857	0.9400	0.0265	0.0218	0.0173
		200	1.0000	1.0000	1.0000	0.0374	0.0353	0.0333
		500	1.0000	1.0000	1.0000	0.0441	0.0428	0.0406
0.5	50	50	0.8500	0.7486	0.6543	0.0095	0.0030	0.0019
		100	0.9971	0.9900	0.9871	0.0281	0.0240	0.0223
		200	1.0000	1.0000	1.0000	0.0351	0.0309	0.0305
		500	1.0000	1.0000	1.0000	0.0474	0.0437	0.0412
0.5	100	50	0.7643	0.7157	0.6443	0.0168	0.0173	0.0118
		100	0.9814	0.9300	0.8500	0.0227	0.0137	0.0145
		200	1.0000	1.0000	1.0000	0.0359	0.0320	0.0319
		500	1.0000	1.0000	1.0000	0.0444	0.0417	0.0409

3.6.3. Prediction evaluation

Prediction performances are also evaluated and compared for models with or without linear constraints. The prediction error $||Y - \mathbf{Z}\hat{\beta}||_2^2/n$ is computed from an independent test sample of size n. Table 3.2 shows the prediction errors of the LASSO estimator, refitted estimator with variables selected by LASSO, and refitted estimator with variables selected by the 95% confidence intervals. For each of these three estimators, model fitting and coefficient refitting and prediction are performed

with multiple, one and no linear constraints. Overall, fitting the models with correct multiple constraints substantially decreases the prediction error. The LASSO estimator has the worst prediction performance, while the two refitted estimators have comparable prediction errors.

Table 3.2: Testing set prediction error of the LASSO estimator, refitted estimator with variables selected by by LASSO, and refitted estimator with variables selected based on 95% confidence intervals. For each estimator, model was fit using multiple, one and no linear constraints. Variable correlations ζ, numbers of variables p and sample sizes (n) are considered.

Configuration			LASSO Estimator			Refitted with Selection by LASSO			Refitted with Selection by 95% CI		
			Constraints			Constraints			Constraints		
ζ	p	n	Multi	One	No	Multi	One	No	Multi	One	No
0.2	50	50	0.687	0.926	0.983	0.360	0.502	1.336	0.370	0.487	1.375
		100	0.360	0.391	0.412	0.300	0.309	1.153	0.284	0.296	1.155
		200	0.293	0.302	0.307	0.271	0.273	1.039	0.264	0.269	1.054
		500	0.265	0.269	0.270	0.259	0.261	1.025	0.255	0.258	1.034
0.2	100	50	1.027	1.429	1.438	0.484	0.776	1.531	0.496	0.602	1.483
		100	0.408	0.467	0.491	0.305	0.315	1.164	0.286	0.322	1.300
		200	0.303	0.318	0.322	0.273	0.276	1.066	0.268	0.277	1.076
		500	0.269	0.274	0.274	0.263	0.264	1.041	0.260	0.264	1.049
0.5	50	50	0.806	1.095	1.210	0.520	0.687	1.179	0.441	0.557	1.278
		100	0.400	0.476	0.454	0.300	0.319	0.959	0.283	0.301	0.963
		200	0.305	0.325	0.320	0.270	0.272	0.861	0.263	0.267	0.877
		500	0.269	0.276	0.274	0.258	0.260	0.847	0.255	0.257	0.862
0.5	100	50	1.069	1.494	1.731	0.668	0.993	1.416	0.606	0.690	1.361
		100	0.476	0.604	0.560	0.322	0.366	0.963	0.293	0.342	1.134
		200	0.323	0.358	0.342	0.271	0.273	0.884	0.265	0.270	0.896
		500	0.274	0.284	0.279	0.262	0.262	0.863	0.258	0.261	0.876

3.6.4. Simulation based on real microbiome compositional data

Another set of simulations are conducted where the gut microbiome composition data analyzed in Section 3.5 are used to generate the covariates with $p = 45$ through resampling. The many zeros in the compositional data matrix are replaced with pseudo-count of 0.05 and are renormalized to have unit sum. For each simulation, we resample with replacement from the rows of compositional data matrix to achieve the required sample size. The coefficients β and noise level σ are the same as in prevision section. The sample size is chosen to be $n = 50, 100, 200$ and 500. Each setting is repeated 500 times. The coverage probability and length of confidence intervals are shown in Figure 3.6 for model with multiple, one and no constraints on the coefficients. Similar conclusions are observed. The coverage probabilities are relatively less sensitive to misspecification of linear constraints, however, the intervals estimated under the correct linear constraints are shorter than

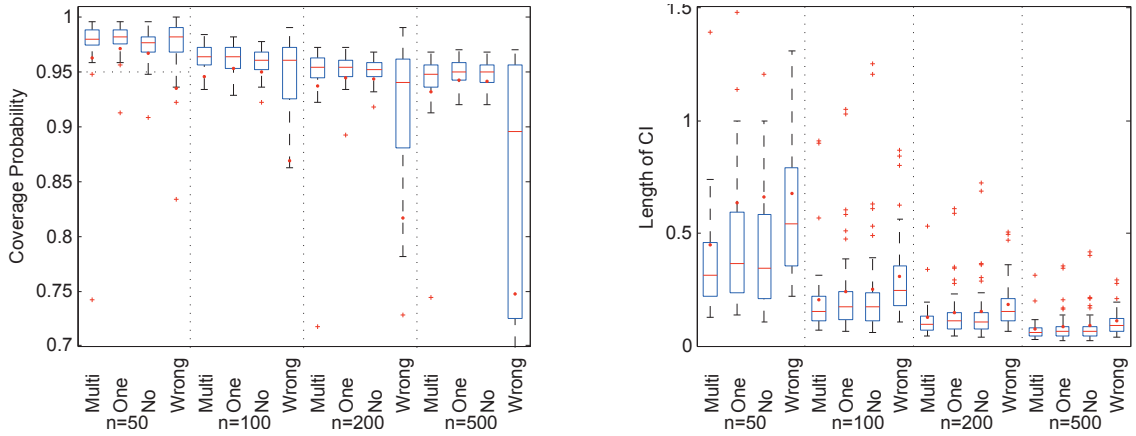


Figure 3.6: Coverage probabilities and length of confidence intervals based on 500 replications. Data are simulated by resampling the gut microbiome composition data in Section 3.5.

those obtained with one or none of the linear constraints, especially when sample size is small. Using the wrong constraints results in much longer intervals with a less accurate coverage.

3.7. Discussion

This chapter has considered the problem of regression analysis for microbiome compositional data obtained through 16S sequencing or metagenomic sequencing. The models and methods in this chapter can be applied to identify the microbial subcompositions that are associated with a continuous response. The idea of imposing the constraints on regression coefficients was motivated by using the log-ratios as covariates. However, the method proposed does not use the log-ratios as covariates, it treats the logarithm of the relative abundances as covariates and allows the response to depend on the relative abundances of certain bacteria instead of the ratios. Imposing linear constraints on coefficients enhances the interpretability and also guarantees the subcompositional coherence. Our method allows selecting taxa in different higher rank taxa. By applying our subcompositional analysis, *Oscillibacter* genus was found to be associated with BMI, even after total fat and caloric intakes were adjusted, indicating that gut microbiome may serve as independent predictor for complex phenotypes such as BMI. Our simulation studies have demonstrated a clear gain in prediction performance when true linear constraints are imposed. However, the small sample size of our data did not allow us to extensively evaluate gain in BMI prediction by incorporating the gut microbiome data.

An estimation procedure through regularization under linear constraints has been developed. In

order to obtain the confidence interval of the regression coefficients, de-biased estimates of the regression coefficients are obtained, which are shown to be approximately normally distributed. The p optimization problems in the de-biased algorithm can be solved efficiently using convex programs. For one simulated data set in Section 3.6, Algorithm 2 took about 36 seconds for $p = 100$ and 300 seconds for $p = 200$ on a PC with a core of Intel i7-3770 CPU3.40GHz. For large p, convex optimization problems can be carried out in parallel. In typical microbiome studies, p is less than 1,000.

The general results presented in this chapter can also be used for statistical inference for the log-contrast model considered in Lin et al. (2014). This type of de-biased estimates were also proposed in Zhang and Zhang (2014) and Geer et al. (2014). Lee et al. (2016) proposed an exact inference procedure for LASSO by characterizing the distribution of a post-selection estimator conditioned on the selection event. It is interesting to extend their approach to the high-dimensional regression problems with constraints. Efron (2014) developed a bootstrap smoothing procedure for computing the standard errors and confidence intervals for predictions, which is different from what was considered in this chapter. Efron's procedure can be applied directly to make inferences on predictions using the methods developed here.

Several extensions are worth considering. Model (3.4) can be extended to include the interaction terms of the form $\lambda_{lk}(\log x_{il} - \log x_{ik})^2$, where x_{il} and x_{ik} are the proportion of the lth and the kth component of subject i, λ_{lk} is the coefficient that corresponds to the interaction between these two components (Aitchison and Bacon-shone, 1984). Similar variable selection and inference procedure can be developed. It is also interesting to develop methods for generalized linear models with high-dimensional compositional data as covariates.

CHAPTER 4

A MODEL FOR PAIRED-MULTINOMIAL DATA AND TESTING ON TAXONOMIC TREE

4.1. Introduction

The human microbiome includes all microorganisms in and on the human body (Gill et al., 2006). These microbes play important roles in human metabolism in order to maintain human health. Dysbiosis of gut microbiome has been shown to be associated with many human diseases such as obesity, diabetes and inflammatory bowel disease (Manichanh et al., 2012; Qin et al., 2012; Turnbaugh et al., 2006). Next generation sequencing technologies make it possible to quantify the relative composition of microbes in high-throughout. Two high-throughput sequencing based approaches have been used in microbiome studies. One approach is based on sequencing the 16S ribosomal RNA (rRNA) amplicons, the resulting reads provide information about the bacterial taxonomic compositions. Another approach is based on shotgun metagenomic sequencing, which sequences all the microbial genomes presented in the sample, rather than just one marker gene. Both 16S rRNA and shotgun sequencing approaches provide bacterial taxonomic composition information and have been widely applied to human microbiome studies, including the Human Microbiome Project (Turnbaugh et al., 2007) and the Metagenomics of the Human Intestinal Tract project (Qin et al., 2010).

Compared to shotgun metagenomics, 16S rRNA sequencing is an amplicon-based approach, which makes the detection of rare taxa easier and requires less starting genomic material than some metagenomic approaches. One important step in analysis of such 16S amplicon sequencing reads data is to assign them to a taxonomy tree. Several computational methods are available for accurate taxonomy assignments, including BLAST (Altschul et al., 1990), the online Greengenes (DeSantis et al., 2006) and RDP (Cole et al., 2007) classifiers, and several tree-based methods. Liu et al. (2008) compared several of these methods and recommended use of Greengenes or RDP classifier. Each taxonomy assignment method produces lineage assignments at the levels of domain, phylum, class, order, family and genus and the final data can be summarized as counts of reads that are assigned to nodes of the existing taxonomic tree.

Given the multivariate nature of the count data measured on the taxonomic tree, methods for anal-

ysis of multivariate count data are greatly needed in the microbiome research. Researchers are interested in testing multivariate hypotheses concerning the effects of treatments or experimental factors on the whole assemblages of bacterial taxa. These types of analyses are useful for studies aiming at assessing the impact of microbiota on human health and on characterizing the microbial diversity in general. Multivariate methods for testing the differences in bacterial taxa composition between groups of metagenomic samples have been developed. The commonly used methods include permutation test such as Mantel test (Mantel, 1967), Analysis of Similarity (ANOSIM) (CLARKE, 1993), and distance-based MANOVA (PERMANOVA) (Anderson, 2001). An alternative test is based on the Dirichlet multinomial (DM) distribution to model the counts of sequence reads from microbiome samples (Chen and Li, 2013; La Rosa et al., 2012). However, this family of DM probability models may not be appropriate for microbiome data because, intrinsically, such models impose a negative correlation among every pair of taxa. The microbiome data, however, display both positive and negative correlations (Mandal et al., 2015). Models that allow for flexible covariance structures are therefore needed.

Many microbiome studies involve collection of 16S amplicon sequencing data over time or over different body sites in order to assess the dynamics of the microbial communities. Such studies generate paired-multinomial count data, where the repeatedly observed microbiomes and therefore the corresponding taxonomic count data are dependent. Modeling such paired-multinomial count data is the focus of this chapter. To the best of our knowledge, there is no flexible model for such paired-multinomial data. In this chapter, a probability distribution for paired multinomial count data, which allows flexible covariance structure, is introduced. The model can be used to model repeatedly measured multivariate counts. Based on this paired-multinomial distribution, a test statistic is developed to test the difference of compositions from paired multivariate count data. An application of the test to the analysis of count data observed on a taxonomic tree is developed in order to test difference in paired microbiome compositions and to identify the subtrees with differential subcompositions.

The chapter is organized as follows. In Section 4.2, the Dirichlet multinomial model and the test of compositional equality based on this model are briefly reviewed. A paired multinomial (PairMN) model for paired count data is defined. In Section 4.3, a statistical test of equal composition based on the paired multinomial model is developed and is applied to counts data observed on the taxo-

nomic tree to test for overall compositional difference and to identify the subtrees that show different subcompositions. Results from simulation studies are reported in Section 4.5 and application to an analysis of gut microbiome data is given in Section 4.6. A brief discussion is given in Section 4.7.

4.2. Paired Multinomial Distribution of Paired Multivariate Count Data

4.2.1. Dirichlet multinomial distribution for multivariate count data and the associated two-sample test

Consider a set of microbiome samples measured on n subjects with d distinct taxa identified across all samples at a given taxonomic level (e.g., phylum, class, etc.). Let $\mathbf{X}_1,\ldots,\mathbf{X}_n \in \mathbb{N}^d$ denote the count data of these n samples, where the jth entry of $\mathbf{X}_i$ is the number of the sequencing reads aligned to the jth taxon from the ith sample. In order to account for overdispersion of the count data in microbiome studies, $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are often assumed to follow a Dirichlet multinomial distribution (Chen and Li, 2013; La Rosa et al., 2012), $DM(N_i,\alpha,\theta), i = 1,\cdots,n$, where N_i is the total number of the reads from the ith sample that are mapped to these d taxa, $\alpha = (\alpha_1,\cdots,\alpha_d)$, $0 \le \alpha_j \le 1$, $\sum_j \alpha_j = 1$ is a vector of the expected taxa composition, and θ is an overdispersion parameter.

Consider the two-group comparison problem, where two groups of microbiome samples, denoted by $\mathbf{X}_{11},\ldots,\mathbf{X}_{n_1 1}$ for the n_1 samples in group 1 and $\mathbf{X}_{12},\ldots,\mathbf{X}_{n_2 2}$ for the n_2 samples in group 2. La Rosa et al. (2012) assumes both independently follow a DM distribution with

$$\begin{aligned} \mathbf{X}_{i1} &\sim DM(N_{i1},\alpha_1,\theta_1), i = 1,\ldots,n_1, \\ \mathbf{X}_{i2} &\sim DM(N_{i2},\alpha_2,\theta_2), i = 1,\ldots,n_2, \end{aligned} \tag{4.1}$$

and propose a test for the following hypothesis:

$$H_0 : \alpha_1 = \alpha_2 \text{ vs } H_a : \alpha_1 \neq \alpha_2. \tag{4.2}$$

Define

$$\hat{\pi}_t = (\sum_{i=1}^{n_t} \mathbf{X}_{it})/(\sum_{i=1}^{n_t} N_{it}), t = 1,2, \tag{4.3}$$

which is a consistent estimator for α_t for $t = 1,2$. Wilson (1989) and La Rosa et al. (2012) proposed

to reject the null hypothesis when

$$\sum_{k=1}^{d} \frac{(\hat{\pi}_{1k} - \hat{\pi}_{2k})^2}{C_1 \hat{\pi}_{1k} + C_2 \hat{\pi}_{2k}} > {\chi^2_{d-1}}^{-1}(1-\alpha), \tag{4.4}$$

where

$$C_t = \frac{1}{N_{\cdot t}^2}\Big(\hat{\theta}_{\alpha_t}\big(\sum_{i=1}^{n_t} N_{it}^2 - N_{\cdot t}\big) + N_{\cdot t}\Big), \quad t = 1,2$$

and $\hat{\theta}_t$ is a consistent estimator of $\theta_t, t = 1,2$.

In many microbiome studies, microbiome data are often observed for the same subjects over two different time points or different body sites. If the microbiome of each subject is measured several times, these repeated measurements are not independent to each other and cannot be handled by the independent DM model. Thus, a new model is developed in the next section to take into account the within-in subject correlations.

4.2.2. Paired Multinomial Distribution for Paired Multinomial Data

Any model for paired multinomial data such as those observed in microbiome studies with repeated measures needs to account for the dependency of the data. For a paired multinomial random variable $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}) \in \mathbb{N}^{d\times 2}, i = 1,\ldots,n$, a paired multinomial (PairMN) distribution can be defined as

$$\mathbf{X}_i \sim \text{PairMN}\big(N_{i1}, N_{i2}, \pi_1, \pi_2, \Sigma_1, \Sigma_2, \Sigma_{12}\big),$$

where

$$\begin{aligned} \mathbf{X}_{it}|\mathbf{P}_{it} &\sim \text{Multinomial}(N_{it}, \mathbf{P}_{it}) \in \mathbb{R}^d, \\ \mathbb{E}\mathbf{P}_{it} &= \pi_t, \\ \text{Var}\mathbf{P}_{it} &= \Sigma_t, \\ \text{Cov}(\mathbf{P}_{i1}, \mathbf{P}_{i2}) &= \Sigma_{12}, \end{aligned} \tag{4.5}$$

for $t = 1,2$. Here, the group-specific composition is represented by π_t. The joint distribution of $(\mathbf{P}_{i1}, \mathbf{P}_{i2})$ is only defined up to its first and second moments so that it includes a wide range of

distributions. Under this probability model, the moments of $\mathbf{X}_{it}$ are given as follow:

$$\begin{aligned} \mathbb{E}\mathbf{X}_{it} &= N_{it}\pi_t, \\ \text{Var}\mathbf{X}_{it} &= N_{it}\left(diag(\pi_t) - \pi_t\pi_t^\top\right) + N_{it}(N_{it}-1)\Sigma_t, \\ \text{Cov}(\mathbf{X}_{i1},\mathbf{X}_{i2}) &= N_{i1}N_{i2}\Sigma_{12}. \end{aligned} \tag{4.6}$$

Compared to the DM model in (4.1), this model has several important features. First, for a given t, the model allows a more flexible covariance structure for the observed counts that is characterized by Σ_t. Second, this model uses Σ_{12} to quantify the correlation between the repeated samples of the same subject. If $\mathbf{P}_{it}$ is assumed to follow a Dirichlet distribution, the proposed model in (4.5) becomes the DM distribution in (4.1). However, a parametric assumption is not needed to achieve the flexible covariance structure.

4.3. Statistical Test Based on Paired Multinomial Samples

4.3.1. A general test for paired multinomial distributions

In order to test if there is any difference in microbiome composition between two correlated samples, consider the following hypotheses:

$$H_0 : \pi_1 = \pi_2 \text{ vs } H_a : \pi_1 \neq \pi_2. \tag{4.7}$$

Define

$$\hat{\pi}_t = \frac{\sum_{i=1}^n \mathbf{X}_{it}}{\sum_{i=1}^n N_{it}},$$

then $\mathbb{E}\hat{\pi}_t = \pi_t$. A Hotelling's T^2 type of statistic based on $\hat{\pi}_1 - \hat{\pi}_2$ can then be developed.

A consistent estimator for $\Sigma_\pi = \text{Var}\,(\hat{\pi}_1 - \hat{\pi}_2)$ is given in the following Lemma.

Lemma 3. *Define*

$$
\begin{aligned}
N_{\cdot t} &= \sum_{i=1}^{n} N_{it} \\
N_{ct} &= \frac{1}{(n-1)N_{\cdot t}} \left(N_{\cdot t}^2 - \sum_{i=1}^{n} N_{it}^2 \right) \\
\mathbf{S}_t &= \frac{1}{n-1} \sum_{i=1}^{n} N_{it} (\hat{\pi}_{it} - \hat{\pi}_t)(\hat{\pi}_{it} - \hat{\pi}_t)^\top \\
\mathbf{G}_t &= \frac{1}{N_{\cdot t} - n} \sum_{i=1}^{n} N_{it} \left(diag(\hat{\pi}_{it}) - \hat{\pi}_{it} \hat{\pi}_{it}^\top \right) \\
\widehat{\Sigma}_{12} &= \frac{1}{(n-1)} \sum_{i=1}^{n} \frac{N_{i1} + N_{i2}}{N_{c1} + N_{c2}} (\hat{\pi}_{i1} - \hat{\pi}_1)(\hat{\pi}_{i2} - \hat{\pi}_2)^T
\end{aligned}
$$

where $\hat{\pi}_{it} = \mathbf{X}_{it}/N_{it}$*, then*

$$
\begin{aligned}
\widehat{\Sigma}_\pi = \sum_{t=1}^{2} \Bigg\{ & \frac{\mathbf{S}_t + (N_{ct} - 1)\mathbf{G}_t}{N_{ct} N_{\cdot t}} + \frac{\sum_{i=1}^{n} N_{it}^2 - N_{\cdot t}}{N_{ct} N_{\cdot t}^2} (\mathbf{S}_t - \mathbf{G}_t) \Bigg\} \\
& - \frac{\sum_{i=1}^{n} N_{i1} N_{i2}}{N_{\cdot 1} N_{\cdot 2}} \left(\widehat{\Sigma}_{12} + \widehat{\Sigma}_{12}^\top \right)
\end{aligned}
\tag{4.8}
$$

is a consistent estimator for $\Sigma_\pi = Var(\hat{\pi}_1 - \hat{\pi}_2)$*. In other words,*

$$
||\widehat{\Sigma}_\pi - \Sigma_\pi||_{\max} \to 0 \text{ in probability as } n \to \infty \tag{4.9}
$$

where $||\cdot||_{\max}$ *is the max norm of matrix.*

A statistic to test H_0 vs H_a specified in (4.7) is defined as

$$
F = \frac{n - d + 1}{(n-1)(d-1)} (\hat{\pi}_1 - \hat{\pi}_2) \widehat{\Sigma}_\pi^\dagger (\hat{\pi}_1 - \hat{\pi}_2)^\top, \tag{4.10}
$$

where $\widehat{\Sigma}_\pi^\dagger$ is the Moore-Penrose pseudoinverse of $\widehat{\Sigma}_\pi$ because $\widehat{\Sigma}_\pi$ is singular due to the unit sum constraint on $\mathbf{P}_{it}$. In the computation, the negative eigenvalues of $\widehat{\Sigma}_\pi$ are truncated to 0 because Σ_π is non-negative definite.

The following Theorem shows that under the null, this test statistic follows an asymptotic F-distribution with degrees of freedom of $d-1$ and $n-d-1$.

Theorem 3. *With test statistic* F *defined in* (4.10)*, an asymptotic level* α *test for testing* (4.7) *is to*

reject H_0 when

$$F > F^{-1}_{d-1,n-d+1}(1-\alpha). \tag{4.11}$$

The p-value for testing (4.7) *is*

$$p = 1 - F_{d-1,n-d+1}(F). \tag{4.12}$$

Remark 1. *Lemma 3 and the proposed test statistic in* (4.10) *can be easily extended to unpaired multivariate count data with unequal sample sizes.*

4.4. Analysis of Microbiome Count Data Measured on the Taxonomic Tree

4.4.1. Identification of subtrees of with differential subcompositions based on the proposed test

The proposed test statistic can be applied to identify the subtrees of the taxonomic tree that are different in their compositions between two measurements (e.g., time or body sites). A rooted taxonomic tree T with nodes $v_1,\ldots,v_{K_0}$ representing for the taxonomic units of T is often available based on 16S sequencing data. For each microbiome sample, 16S RNA reads can be aligned to the nodes of T. Without loss of generality, assume that the first K nodes $v_1,\ldots,v_K$ are all the internal non-leaf nodes and v_1 is the root node. Also, denote $\tau(v_k)$ as the set of all direct child nodes of $v_k, k = 1,\ldots,K$.

For a given internal node k, let the count $\mathbf{Q}(v_k)$ assigned to node v_k be the number of all descending reads of v_k. For convenience, also denote $\mathbf{Q}(S) = \big(\mathbf{Q}(v_{k_1}),\ldots,\mathbf{Q}(v_{k_j})\big)$ for any set of nodes $S = \{v_{k_1},\ldots,v_{k_j}\}$. For each split from a parental node to the child nodes, the reads on the parent node are either assigned to a child node or remain unassigned. For each parent node v_k, the counts of reads assigned to its direct child node are in vector $\mathbf{Q}(\tau(v_k))$, and the count of reads unassigned is $\mathbf{Q}(v_k) - \sum_{j\in\tau(v_k)}\mathbf{Q}(v_j)$. For a subject i with measurement index t, at a given internal node k, $k = 1,\ldots,K$, denote

$$\mathbf{X}^{(k)}_{it} = \Big(\mathbf{Q}_{it}(\tau(v_k)), \mathbf{Q}_{it}(v_k) - \sum_{j:v_j\in\tau(v_k)}\mathbf{Q}_{it}(v_j)\Big)^\top, \tag{4.13}$$

$$N^{(k)}_{it} = \mathbf{Q}_{it}(v_k), \tag{4.14}$$

where $\mathbf{X}^{(k)}_{it}$ is the vector of the read counts that are assigned to each of the child node or unassigned reads and $N^{(k)}_{it}$ is the sum of these read counts. In the subtree shown in Figure 4.2b, $\mathbf{X}^{(k2)}_{it} =$

$(\mathbf{Q}_{it}(v_{k8}), \mathbf{Q}_{it}(v_{k9}), \mathbf{Q}_{it}(v_{k2}) - \mathbf{Q}_{it}(v_{k8}) - \mathbf{Q}_{it}(v_{k9}))$ and $N_{it}^{(k2)} = \mathbf{Q}_{it}(v_{k2})$.

For a study with a pair of repeated microbiome measurements, for each internal node k, $k = 1, \ldots, K$, the paired vectors of read counts are assumed to have PairMN distributions,

$$(\mathbf{X}_{i1}^{(k)}, \mathbf{X}_{i2}^{(k)}) | (N_{i1}^{(k)}, N_{i2}^{(k)}) \sim \text{PairMN}\big(N_{i1}^{(k)}, N_{i2}^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \Sigma_1^{(k)}, \Sigma_2^{(k)}, \Sigma_{12}^{(k)}\big),$$

where $\mathbf{X}_{it}^{(k)}$ are counts for the ith sample in the tth measurement.

In order to identify the subtrees with differential subcompositions between the two measurements, the following hypotheses can be tested using the proposed method in Theorem 3,

$$H_0^{(k)} : \pi_1^{(k)} = \pi_2^{(k)}, \quad k = 1, \ldots, K. \tag{4.15}$$

Define p_k as the p-value from testing $H_0^{(k)}$. Theorem 3 shows that under the null hypotheses, p_k's are asymptotically uniformly distributed. In fact, they are also asymptotically independent under the null. Take Figure 4.2b as an example, under the $H_0^{(k1)}$ and $H_0^{(k2)}$,

$$\begin{aligned}
\mathbb{P}(p_{k1} \le \alpha, p_{k2} \le \beta) &= \int \mathbb{P}(p_{k1} \le \alpha | \mathbf{Q}(v_{k2}), p_{k2} \le \beta) \mathbb{P}(p_{k2} \le \beta | \mathbf{Q}(v_{k2})) dF(\mathbf{Q}(v_{k2})) \\
&= \int \mathbb{P}(p_{k1} \le \alpha | \mathbf{Q}(v_{k2})) \mathbb{P}(p_{k2} \le \beta | \mathbf{Q}(v_{k2})) dF(\mathbf{Q}(v_{k2})) \\
&\overset{a}{=} \alpha \int \mathbb{P}(p_{k2} \le \beta | \mathbf{Q}(v_{k2})) dF(\mathbf{Q}(v_{k2})) \\
&= \alpha \mathbb{P}(p_{k2} \le \beta) \overset{a}{=} \mathbb{P}(p_{k1} \le \alpha) \times \mathbb{P}(p_{k2} \le \beta)
\end{aligned}$$

where $\overset{a}{=}$ are equations that hold asymptotically. Therefore, to control for multiple comparisons, the false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995) can be used to identify the subtrees with different subcompositions between two repeated measurements.

4.4.2. Global test for differential overall compositions on taxonomic tree

The goal for testing the global difference in taxonomic compositions between pairs of measurements can be formulated as the following composite hypothesis,

$$H_0 = \cap_{k=1}^{K} H_0^{(k)} \quad \text{vs} \quad H_a = \cup_{k=1}^{K} (H_0^{(k)})^c, \tag{4.16}$$

where $H_0^{(k)}$ is specified by (4.15).

To test this composite hypothesis, the combined p-value can be obtained using Fisher's method,

$$p_{combined} = 1 - (\chi^2_{2K})^{-1}\left(-2\sum_{k=1}^{K}\log p_k\right) \tag{4.17}$$

or 2^{nd} smallest p-value,

$$p_{combined} = 1 - \big(1+(K-1)p_{(2)}\big)(1-p_{(2)})^{K-1} \tag{4.18}$$

where $p_{(2)}$ is the 2^{nd} smallest p-value of $p_1,\ldots,p_K$. Under the null, the $p_{combined}$ computed using either methods is asymptotically uniformly distributed. Test (4.18) is more powerful if only a small number of subtrees with differential subcompositions between the two measurements, while test (4.17) is more suitable if differences occur in a large number of subtrees.

4.5. Simulation Studies

4.5.1. Comparison with test based on the DM model

To compare the performance of our pairMN test statistic in (4.11) with the original unpaired statistic (4.4), two data generating models within the class of PairMN are considered. The first model generates $\mathbf{P}_{it}, i = 1,\ldots,n$ based on a mixture of Dirichlet distributions:

$$\begin{aligned}
&\mathbf{P}_{it} = (1-\rho)\mathbf{P}'_{it} + \rho\mathbf{P}''_i, \quad t = 1,2,\\
&\mathbf{P}'_{it} \sim \mathsf{Dir}(\alpha_t, \theta_{\alpha_t}), \quad t = 1,2,\\
&\mathbf{P}''_i \sim \mathsf{Dir}(\ell, \theta_\ell).
\end{aligned} \tag{4.19}$$

Under this setting,

$$\begin{aligned}
\pi_t =&\ (1-\rho)\alpha_t + \rho\ell, \quad 0<\rho<1, \quad t = 1,2\\
\Sigma_t =&\ (1-\rho)^2\theta_{\alpha_t}\big(diag(\alpha_t) - \alpha_t\alpha_t^\top\big) + \rho^2\theta_\ell\big(diag(\ell) - \ell\ell^\top\big), \quad t = 1,2\\
\Sigma_{12} =&\ \rho^2\theta_\ell\big(diag(\ell) - \ell\ell^\top\big).
\end{aligned} \tag{4.20}$$

In our simulation, the dimension is set as $d = 8$. The parameter ρ is used to control the degree of correlation in Σ_{12}, where ρ ranges from 0 to 0.6. Other parameters are set as $\theta_\ell = 1$, $\theta_{\alpha_1} = 3$, $\theta_{\alpha_2} = 5$,

$\ell = (0.12, 0.06, 0.08, 0.43, 0.02, 0.14, 0.1, 0.05)$, α_1 and α_2 such that $\pi_1 = (0.15, 0.05, 0.22, 0.3, 0.03, 0.1, 0.07, 0.08)$, and under the alternative hypothesis $\pi_2 = (0.1, 0.1, 0.22, 0.3, 0.03, 0.1, 0.07, 0.08)$. The number of total counts N_{it} are simulated from a Poisson distribution with mean 1000. When $\rho = 0$, this model degenerates to the Dirichlet-multinomial distribution.

The second model generates $\mathbf{P}_{it}, i = 1, \ldots, n$ based on a log-normal distribution. Specifically,

$$\mathbf{P}_{it} = \frac{e^{\mathbf{Z}_{it}}}{1^\top e^{\mathbf{Z}_{it}}}, \quad t = 1, 2, \tag{4.21}$$

where

$$(Z_{i1j}, Z_{i2j}) \sim N\left(\begin{bmatrix} \mu_{j1} \\ \mu_{j2} \end{bmatrix}, \begin{bmatrix} \sigma_{j1}^2 & \rho\sigma_{j1}\sigma_{j2} \\ \rho\sigma_{j1}\sigma_{j2} & \sigma_{j2}^2 \end{bmatrix} \right), \quad j = 1, \ldots, d,$$
$$\mathbf{Z}_{it} = (Z_{it1}, \ldots, Z_{itd})^\top, \quad t = 1, 2.$$

Under this setting, no explicit expressions for π_t, Σ_t and Σ_{12} are available, but the correlation can be quantified using ρ, and the difference in π_t can be quantified by the difference in $\mu_t = (\mu_{1t}, \ldots, \mu_{dt})^\top, t = 1, 2$. In our simulation, the dimension of sample is $d = 8$, ρ ranges from 0 to 0.6, $\sigma_t = (\sigma_{1t}, \ldots, \sigma_{dt})^\top = (1, \ldots, 1)$ for $t = 1, 2$, $\mu_1 = (3, 1, 0.5, 1, 0, 1, 1, 0)$, and set $\mu_2 = (3, 1, 1, 0.5, 0, 1, 1, 0)$ under the alternative. The number of total counts N_{it} are also simulated from a Poisson distribution with mean 1000.

For both data generating models, sample sizes of $n = 20, 50$ and 100 are considered. The simulations are repeated 5,000 times for each specific setting and the null hypothesis is rejected at level of $\alpha = 0.05$. The type I error and the empirical power of the various tests are shown in Figure 4.1. It shows that both tests have test size under the nominal level in all of the settings. For data simulated from the paired multinomial-Dirichlet distribution (4.19), the power of the unpaired test is slightly better than the paired test only when ρ is very small, that is, when there is a weak within-subject correlation (Figure 4.1 (a)). This is expected since the unpaired test (4.4) is developed specifically for the Dirichlet-multinomial distribution, i.e. PairMN model with $\rho = 0$. When ρ increases from 0 to 0.6, the paired test has a steadily increasing power with the test size still around the nominal level, while the size and power of the unpaired test gradually decrease. The results suggest that compared with the paired test, the unpaired test tends to be conservative and therefore has reduced

power in detecting the difference in compositions when the within-subject correlation is large.

For data simulated from log-normal-based PairMN model (4.21), the power of our paired test is much larger than the power of the unpaired test for all values of ρ, while the type 1 errors are well controlled (Figure 4.1 (b)). These results show that the proposed paired test performs well in both data generating models, suggesting that our test is very flexible and robust to different distributions of $\mathbf{P}_{it}$.

4.5.2. Simulating count data on a taxonomic tree

The proposed tests in (4.17) and (4.18) are further compared with PERMANOVA test (Anderson, 2001) using L^1 Kantorovich-Rubinstein (K-R) distance (Evans and Matsen, 2012) with unit branch length and with each pair of samples as a stratum. Using the notations in Section 4.4.2, the L^1 K-R distance between two trees $\mathbf{Q}_{i_1t_1}$ and $\mathbf{Q}_{i_2t_2}$ is given by

$$d(\mathbf{Q}_{i_1t_1}, \mathbf{Q}_{i_2t_2}) = \sum_{k=1}^{K_0} |\mathbf{p}_{i_1t_1}(v_k) - \mathbf{p}_{i_2t_2}(v_k)| \tag{4.22}$$

where

$$\mathbf{p}_{it}(v_k) = \big(\mathbf{Q}_{it}(v_k) - \sum_{j:v_j \in \tau(v_k)} \mathbf{Q}_{it}(v_j)\big)/\mathbf{Q}_{it}(v_1) \quad k = 1, \ldots, K_0$$

is the proportion of reads that are assigned to node v_k but cannot be further specified to its child nodes. This is sum of the l_1 distances between two compositional vectors over each branch of the taxonomic tree.

In order to simulate data that mimic real microbiome count data, count data on the taxonomic tree are generated based on sampling from a real 16S microbiome dataset from Flores et al. (2014), where the gut (feces), palm and tongue microbial samples of 85 college-age adults where taken in a range of three months and were characterized using 16S rRNA sequencing. Within the gut microbiome samples, counts of reads are summarized on a taxonomic tree that has 1050 nodes from kingdom to species (see Figure 4.2). Since no large changes are expected in gut microbiome during a three-month period, these samples are assumed to have the same null distribution, which resulted in a total of 638 gut microbial samples.

Using the notation in Section 4.4.1, these samples are denoted as $\mathbf{Q}^o_1, \ldots, \mathbf{Q}^o_{638}$. Before the simula-

tion, the matrix $\mathbf{P}^o \in (0,1)^{638\times 1050}$ with

$$\mathbf{P}^o(i,k) = \left(\mathbf{Q}_i^o(v_k) - \sum_{j:v_j\in\tau(v_k)} \mathbf{Q}_i^o(v_j)\right)/\mathbf{Q}_i^o(v_1), \quad i = 1,\ldots,638, \quad k = 1,\ldots,1050$$

is first calculated, which is the composition of all nodes for each of the 638 gut microbial samples. The total count of reads of all samples are also calculated and recorded using $\mathbf{N}^o \in \mathbb{N}^{638}$.

To simulate a pair of correlated microbiome sample $\mathbf{Q}_{i1}$ and $\mathbf{Q}_{i2}$, three compositions $\mathbf{P}_{i1}^o$, $\mathbf{P}_{i2}^o$ and $\mathbf{P}_{i3}^o$ from $\mathbf{P}^o$ are randomly sampled and two total counts N_{i1}^o and N_{i2}^o are randomly resampled from $\mathbf{N}^o$. Read counts $\mathbf{W}_{i1}$ are sampled from multinomial$(N_{i1}^o, \frac{\mathbf{P}_{i1}^o+\mathbf{P}_{i3}^o}{2})$, and $\mathbf{W}_{i2}$ from multinomial$(N_{i2}^o, \frac{\mathbf{P}_{i2}^o+\mathbf{P}_{i3}^o}{2}) + \mathbf{E}_i$, where $\mathbf{E}_i$ is a perturbation to the genus of Streptococcus, and is drawn from Binomial$(N_{i2}^o, p_\varepsilon)$ at the coordinate corresponding to Streptococcus and zero otherwise. $\mathbf{Q}_{it}$ is then iteratively computed such that $\mathbf{Q}_{it}(v_k) - \sum_{j:v_j\in\tau(v_k)} \mathbf{Q}_{it}(v_j) = \mathbf{W}_{itk}$ for $t = 1,2$ and $i = 1,\ldots,n$, where n is the number of pairs simulated and is set to be 20, 50 and 100 in our simulation. The percent of perturbation p_ε is chosen to range from 0 to 2%. For each scenario, we repeat the simulations for 100 times. For the global test of (4.16), we reject the null hypothesis at the level of 0.05. For the identification of subtrees with differential subcompositions in multiple testing (4.15), we control the FDR at the level of 0.05.

Figure 4.3 compares the rejection rate of PERMANOVA with our method in (4.17) and (4.18) for the global test (4.16). Within method (4.18), which combines p-values using the 2^{nd} smallest p-value, we also compare the our paired test based on PairMN in (4.10) with the unpaired test based on DM in (4.4). When the sample size is small, none of the methods is able to detect the perturbation to Streptococcus. As the sample size increases, the rejection rate of our method using 2^{nd} smallest p-value combination of the paired-tests gradually increases, especially when the percent of perturbation gets closer to 2%. Our method using (4.17) with Fisher's method of p-value combination does not perform as well because the perturbation only occurs to a very small number of subtrees. The method using 2^{nd} smallest p-value combination of the unpaired tests also performs worse than the paired tests.

Figure 4.4 shows the percent of discoveries of the differential subtrees with FDR controlled at 0.05. We observe that the observed FDR is close to the nominal level of 0.05. Since the count of genus Streptococcus is set to be different, the counts on all the parent nodes of Streptococcus are also

Table 4.1: $p-$values of different comparisons between two body sites and between smokers and non-smokers based on the proposed tests and PERMANOVA.

	PairMN (Fisher || 2nd)	PERMANOVA
Nasopharynx and Oropharynx (Left Side)	0 || 0	$<$0.001
Nasopharynx and Oropharynx (Right Side)	0 || 0	$<$0.001
Smoker vs Nonsmoker (nasopharynx)	2.1e-07 || 8.6e-05	0.003
Smoker vs Nonsmoker (oropharynx)	1.2e-07 || 6.2e-04	0.005
Left vs Right (nasopharynx)	0.16 || 0.65	0.053
Left vs Right (oropharynx)	0.37 || 0.79	0.99

changed. Therefore, the differential subtrees denoted by their root nodes are: (a) Kingdom Bacteria, (b) Phylum Firmicutes, (c) Class Bacilli, (d) Order Lactobacillales, (e) Family Streptococcaceae and (f) Genus Streptococcus (see Figure 4.2). Among these, (c) and (e) are not identified in any scenario because these subtrees have counts mostly mapped in one child node and thus make any changes nearly impossible to detect. The test does not have power to identify (a) because the perturbation is too small to detect given the large counts on the child nodes of (a). All the other three subtrees are identified by our method when the percent of perturbation and sample size get larger.

4.6. Analysis of Microbiome Data in the Upper Respiratory Tract

The human nasopharynx and oropharynx are two body sites located very close to each other in the upper respiratory tract. The nasopharynx is the ecological niche for many commensal bacteria. It is interesting to understand whether these nearby sites have similar microbiome composition and how smoking perturbs their compositions. Charlson et al. (2010) collected the left and right nasopharynx and oropharynx microbiome samples from 32 current smokers and 36 nonsmokers. The samples were sequenced using 16S rRNA sequencing, and the count of reads are aligned onto a taxonomic tree with 213 nodes from kingdom to species.

Several comparisons of the overall microbiome compositions were compared and the results are summarized in Table 4.1. As expected, no significant differences were observed between left and right nasopharynx or oropharynx. However very significant differences were observed between nasopharynx and oropharynx both in the left and right sides, further confirming the niche-specific colonization at discrete anatomical sites. In addition, smoking had strong effects on microbiome composition in both nasopharynx and oropharynx

4.6.1. Comparison of nasopharynx and oropharynx microbiome for nonsmokers

Since a large overall microbiome composition differences was observed, it is interesting to identify which subtrees and their corresponding subcompositions led to such a difference. The proposed subtree identification procedure in Section 4.4.2 using the pairNM test in (4.11) was applied to identify the subtrees with differential subcompositions between the two body sites at an FDR=0.05. The identified parental nodes, their child nodes and the corresponding subcompositions are shown in Figure 4.5. One advantage of the proposed method is to identify these subtrees at various taxonomic levels. For example, at the phylum level, nasopharynx clearly had more *Firmicutes*, however, oropharynx had more *Bacteroidetes*. At the genus level, *Streptococcus* appeared more frequently in oropharynx, but *Lactococcus* occurred more in nasopharynx.

4.6.2. Comparison of microbiome between smokers and non-smokers

The proposed procedure was also applied to identify the differential subtrees with differential subcompositions between smokers and nonsmokers in nasopharynx and oropharynx and the results are shown in Figure 4.6 for a FDR=0.05. For nasopharynx, the subcompositions of classes under *Firmicutes*, classes under *Bacteroidetes* and families under *Clostridiale* were different, with fewer *Bacilli* in *Firmicutes*, more *Bacteroidia* in *Bacteroidetes*, and fewer *Veillonellaceae* in *Clostridiales* being observed in smokers (Figure 4.6a).

For oropharynx, differences in the subcomposition of phyla and species under *Prevotalla* were observed, with more *Firmicuates* and more Melaninogenica in Prevotella observed in smokers (Figure 4.6b).

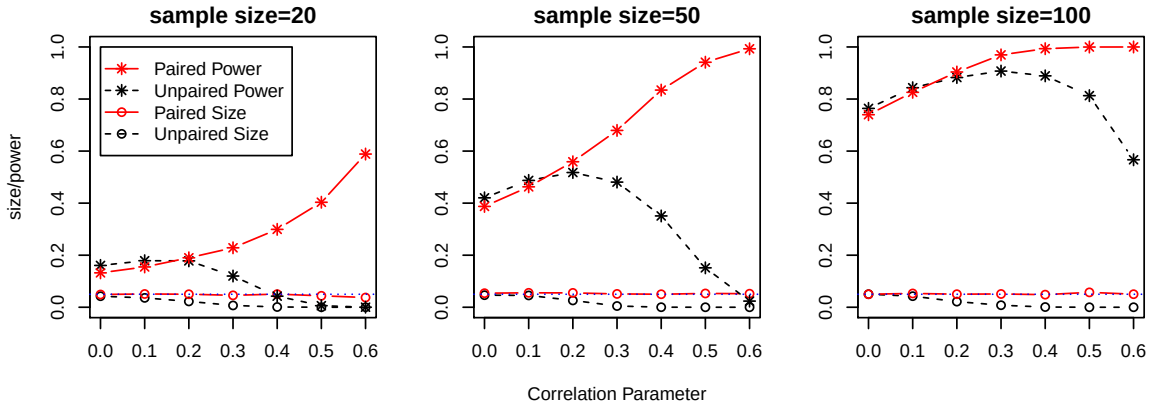
4.7. Discussion

This chapter has introduced a flexible model for paired multinomial data in modeling the dependency of of the data. Based on this model, a T^2 type of test statistic has been developed for testing equality of the overall composition between two repeatedly measured multinomial data. The test can be used for analysis of count data observed on a taxonomic tree to identify the subtrees that show differential subcompositions in repeated measures. Our simulations have shown that the proposed test has correct type 1 errors and much increased power than the commonly used tests based on DM model or the PERMANOVA test. The test proposed in this chapter can be applied to

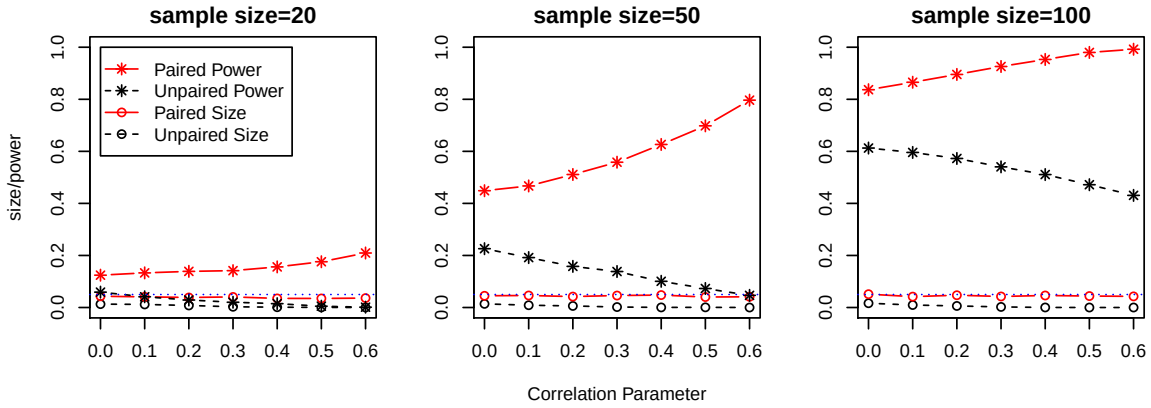
both independent and repeated measurement data. For independent data, the proposed test allows more flexible dependent structure among the taxa than the Dirichlet multinomial model, which only allows negative correlations among the taxa. The proposed test statistics are also computationally more efficient than the commonly used permutation-based procedures such as PERMANOVA, which enables their applications in large-scale microbiome studies.

As demonstrated in our simulations, the proposed overall test of composition is more powerful than PERMANOVA type of tests when the overall composition difference is due to a few subcompositions since our test considers each subtree and subcomposition separately and then combines the p-values. Since the tests for differential subcomposition condition on the total counts of the parental nodes, all the p-values are independent that facilitates simple combination of p-values and identification of subtrees based on FDR controlling.

Although the chapter has focused on using existing taxonomic tree and 16S sequencing data, the tests proposed in this chapter can also be applied to shotgun metagenomic sequencing data. One possible approach is to build phylogenetic trees based on a small set of universal marker genes (Sunagawa et al., 2013) and to align the sequencing reads to these phylogenetic trees. The proposed methods can be applied to each of these trees and the results can be combined. This deserves further investigation.

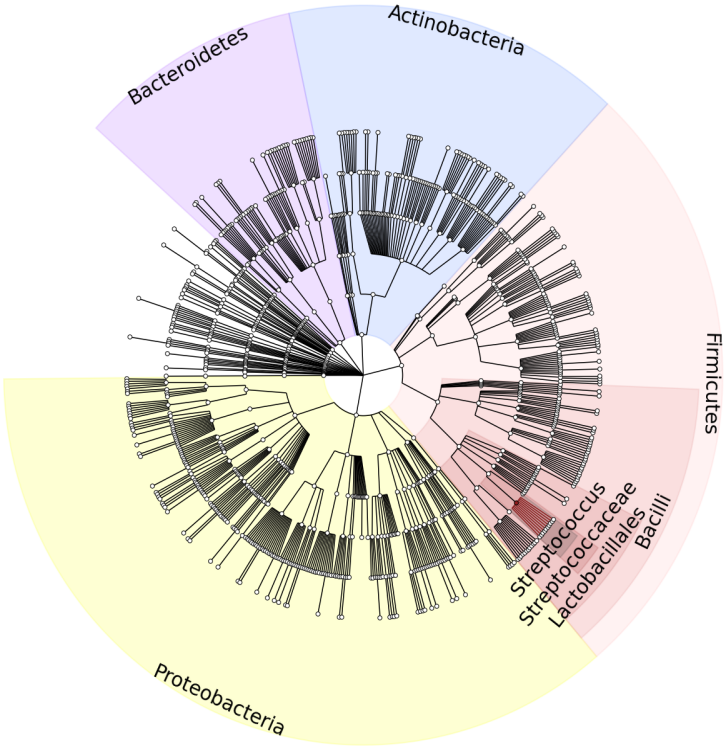


(a) Mixed Dirichlet P_{it}.

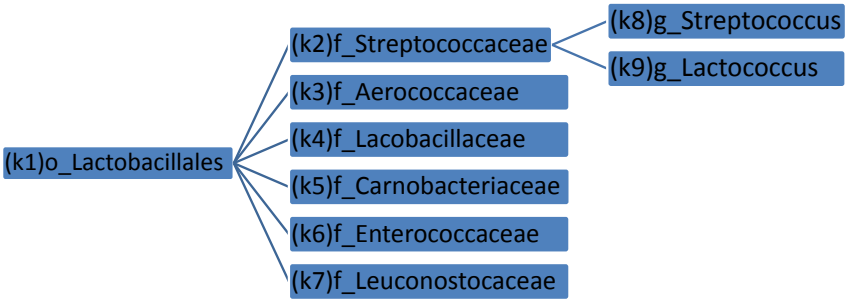


(b) Log-normal based P_{it}.

Figure 4.1: Simulation results: size and power of the paired and unpaired tests for data simulated under the PairMN model (a) and the correlated log-normal model (b) for sample size $n = 20, 50$ and 100. x-axis is the correlation parameter ρ.



(a) Entire taxonomic tree. This figure is generated using GraPhlAn (Asnicar et al., 2015).



(b) Detail tree structure above genus of Streptococcus.

Figure 4.2: Taxonomic tree of the gut microbiome samples from which the simulated data are generated. In our simulations, the count of genus Streptococcus is perturbed to generate samples from the alternative distribution.

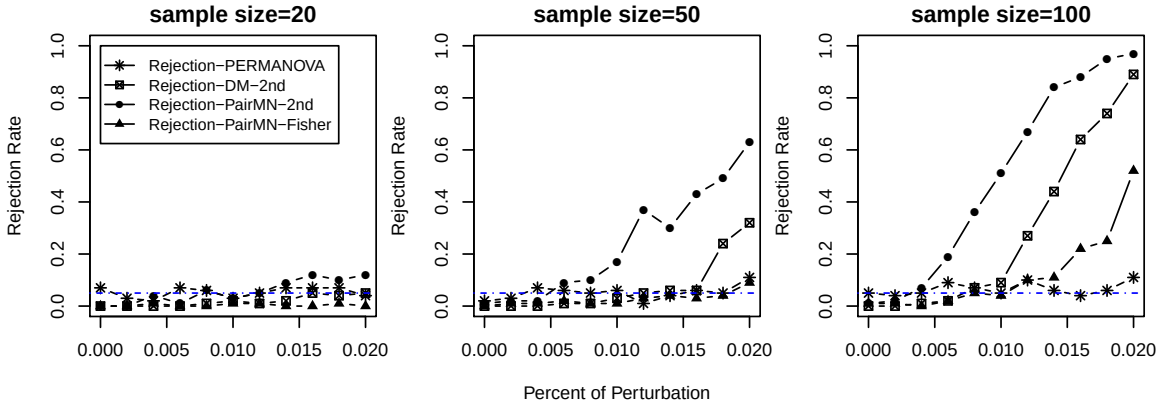


Figure 4.3: Comparison of rejection rate of the proposed method with PERMANOVA with the level of test at $\alpha = 0.05$. X-axis is the perturbation percentage p_ε, where $p_\varepsilon = 0$ corresponds to the null hypothesis.

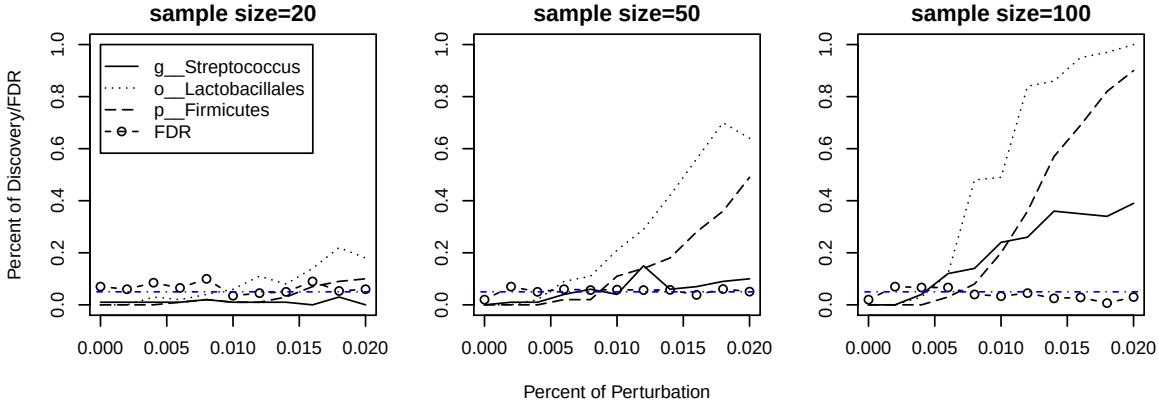


Figure 4.4: Identification of subtrees with differential subcompositions with FDR set to 0.05. Y-axis shows the percent of discovery of the corresponding subtree in 100 simulations with FDR controlled at 0.05. The empirical FDR is close to 0.05.

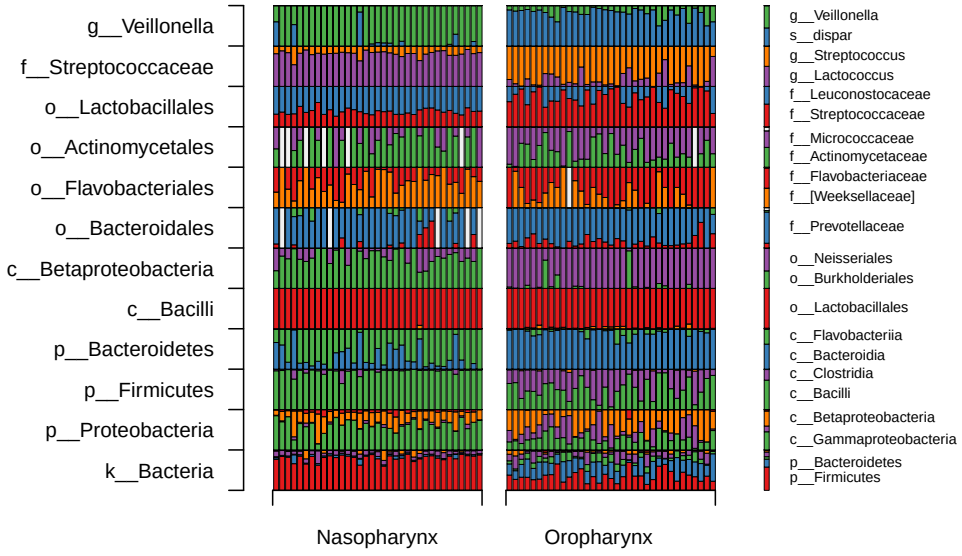
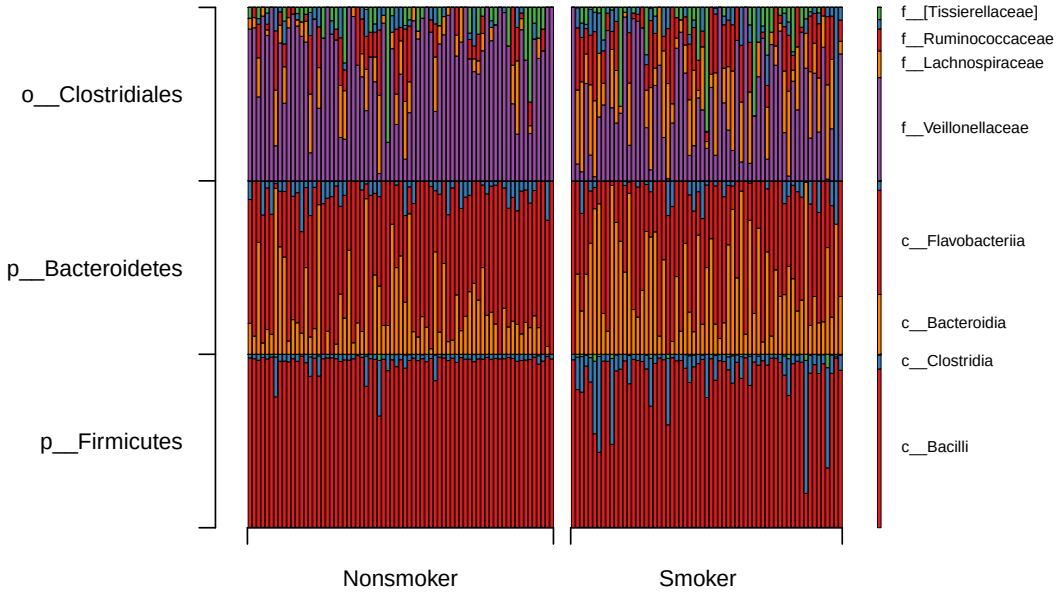
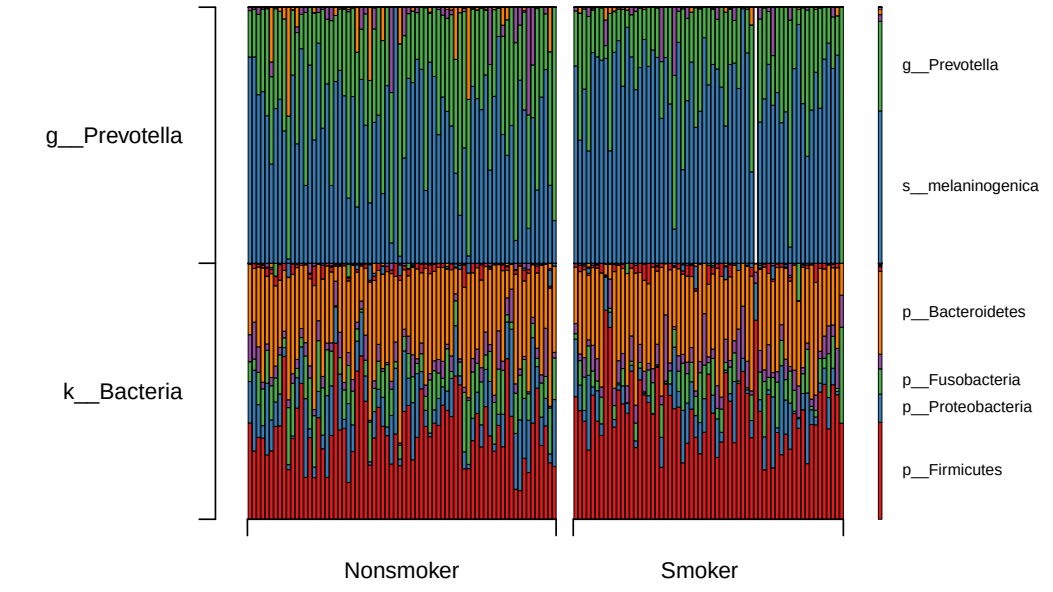


Figure 4.5: Parental nodes and the child nodes that showed differential subcomposition between nasopharynx and oropharynx.



(a) Nasopharynx.



(b) Oropharynx.

Figure 4.6: Parental nodes and the child nodes that showed differential subcomposition between smokers and nonsmokers in nasopharynx (a) and oropharynx (b).

CHAPTER 5

FUTURE TOPICS

Part of my future research will be related to my current work. I am hoping to develop more statistical methods for problems in microbiome research and metagenomic studies. The following are projects I am interested in working on in the near future.

5.1. Log-Contrast Generalized Linear Models

In human microbiome research, besides continuous responses such as BMI, discrete responses such as presence/absence of Crohn's disease are also common. Since ordinary linear model is not suitable for discrete response, it is important to extend our work in Chapter 3 to log-contrast generalized linear model (GLM). Suppose the response y_i follows an exponential family distribution with the log composition Z_i as covariates in the following way:

$$f(y_i|\beta, Z_i) = h(y_i)\exp\{\eta_i y_i - A(\eta_i)\}, \quad \eta_i = Z_i^\top \beta,$$

where β is the regression coefficients. This is a special case of the exponential family, which includes binomial distribution with $A(\eta) = \log(1+e^\eta)$. Then, similar to Model (3.4), the estimation of β can be formulated as a convex optimization problem below

$$\begin{aligned} \widehat{\beta}^n &= \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^n \log f(y_i|\beta, Z_i) + \lambda ||\beta||_1 \right\} \text{ subject to } C^\top \beta = 0 \\ &= \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n}[Y^\top \mathbf{Z}\beta - \sum_{i=1}^n A(Z_i^\top \beta)] + \lambda ||\beta||_1 \right\} \text{ subject to } C^\top \beta = 0 \end{aligned} \tag{5.1}$$

Define $\mu(\eta) = \nabla_{\eta_i} A(\eta_i)$ and $v(\eta) = \nabla^2_{\eta_i} A(\eta_i)$. To obtain the confidence intervals for β, we can also construct a de-biased estimator $\widehat{\beta}^u$ with some modification to (3.12) by

$$\widehat{\beta}^u = \widehat{\beta}^n + \frac{1}{n}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top (Y - \mu(\widehat{\beta}^n, \widetilde{\mathbf{Z}})). \tag{5.2}$$

where $\mu = (\mu(\widetilde{Z}_1^\top \widehat{\beta}^n), \ldots, \mu(\widetilde{Z}_n^\top \widehat{\beta}^n))$, and $\widetilde{\mathbf{M}}$ can be computed by (3.10) and (3.11) with $\widehat{\Sigma} = (\widetilde{\mathbf{Z}}^\top \mathbf{V}\widetilde{\mathbf{Z}})/n$ and $\mathbf{V} = diag(v(\widetilde{Z}_1^\top \widehat{\beta}^n), \ldots, v(\widetilde{Z}_n^\top \widehat{\beta}^n))$. However, detailed theoretical analyses of the estimator and

confidence intervals require more assumptions on the transformed design matrix $\widetilde{\mathbf{Z}}$ and careful handling of function $A(\eta)$.

5.2. Statistical Inference for Signal-Noise-Ratio

Another important question in microbiome research is to estimate the fraction of the variance in a response such as BMI can be explained by the observed microbiome composition, as opposed to other unknown or unmeasured factors. This quantity corresponds to the "R-square" and signal-noise-ratio in the linear regression model. To be specific, we consider the high-dimensional regression model with log-transformed microbiome composition as covariates and linear constraints on the regression coefficients:

$$y_i = Z_i^\top \beta + \varepsilon_i, \quad \mathbf{C}^\top \beta = 0, \quad i = 1, \ldots, n. \tag{5.3}$$

Here $Z_1, \cdots, Z_n$ are independent log-transformed microbiome compositions, $\varepsilon_1, \cdots, \varepsilon_n$ are errors which are uncorrelated with $Z_1, \cdots, Z_n$. The problem of quantifying variance explained by covariates is to make statistical inference for the signal-noise-ratio (SNR):

$$SNR = 1 - \frac{\sigma^2}{\mathrm{Var}(y_i)} = 1 - \frac{\sigma^2}{\mathrm{Var}(Z_i^\top \beta) + \sigma^2},$$

where $\sigma^2 = \mathrm{Var}(\varepsilon_i)$ is the noise level, $\mathrm{Var}(Z_i^\top \beta)$ is the variance which can be explained by covariates. Since $\mathrm{Var}(y_i)$ can be naturally estimated by $\sum_{i=1}^n y_i^2/n$, we can instead focus on statistical inference for σ^2. The problem of statistical inference for σ^2 has attracted a lot of recent interest. For example, Fan, Guo, and Hao (2012) and Sun and Zhang (2012) both provide estimates of σ^2 with sparsity assumption on β. Dicker (2014) and Janson, Barber, and Candès (2015) develop methods for the estimation and inference of σ^2 without assuming β to be sparse but require the covariance structure of Z_i to be known. However, in microbiome research, the covariance structure of Z_i is often complicated and hardly known and β is sometimes dense, it is more desirable to develop a method of inference on SNR when β is not necessarily sparse and covariance structure of Z_i is limited.

APPENDIX

PROOFS

A.1. Proofs for Chapter 2

Proof of Theorem 1. Let $\hat{S} = \{j : \hat{\beta}_j \neq 0\}$ denote the support of $\hat{\beta}$. If $p \in \hat{S}$, the optimality conditions for problem (2.3) can be written as

$$-n^{-1}(Z^p_{\hat{S}_{\setminus p}})^\top (y - Z^p \hat{\beta}_{\setminus p}) + \lambda\{\operatorname{sgn}(\hat{\beta}_{\hat{S}_{\setminus p}}) - \operatorname{sgn}(\hat{\beta}_p) 1_{s-1}\} = 0, \tag{A.1}$$

$$\|n^{-1}(Z^p_{\hat{S}^c})^\top (y - Z^p \hat{\beta}_{\setminus p}) + \lambda \operatorname{sgn}(\hat{\beta}_p) 1_{p-s}\|_\infty \leq \lambda, \tag{A.2}$$

where $\hat{\beta}_p = -1^\top_{p-1} \hat{\beta}_{\setminus p}$. The idea of the proof is to define an event that occurs with high probability and, conditioning on that event, find some $\hat{\beta}$ with the desired properties such that (A.1) and (A.2) hold.

For $J \subset \{1, \ldots, p\}$, let Z_J denote the submatrix formed by the jth columns of Z with $j \in J$. By the union bound and the classical Gaussian tail bound, we have

$$Pr\{\|n^{-1}(Z_S)^\top \varepsilon\|_\infty \geq \lambda/2\} \leq \sum_{j \in S} Pr\{|n^{-1} z_j^\top \varepsilon| \geq \lambda/2\} \leq s \exp\{-n\lambda^2/(8\sigma^2)\}$$

and, since $\Pi \equiv I - n^{-1} Z^p_{S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} (Z^p_{S_{\setminus p}})^\top$ is a projection matrix and has spectral norm at most 1,

$$\begin{aligned} &Pr\{\|n^{-1}(Z^p_{S^c})^\top \varepsilon - C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} n^{-1} (Z^p_{S_{\setminus p}})^\top \varepsilon\|_\infty \geq \lambda\xi\} = Pr\{\|n^{-1}(Z^p_{S^c})^\top \Pi\varepsilon\|_\infty \geq \lambda\xi\} \\ &\leq \sum_{j \in S^c} Pr\{|n^{-1}(z_j - z_p)^\top \Pi\varepsilon| \geq \lambda\xi\} \leq (p-s) \exp\{-n\lambda^2\xi^2/(8\sigma^2)\}, \end{aligned}$$

where we have used the normalization assumption $\max_j \|z_j\|_2 \leq \sqrt{n}$. Thus, with probability at least $1 - p\exp\{-n\lambda^2\xi^2/(8\sigma^2)\}$, the following inequalities hold:

$$\|n^{-1}(Z_S)^\top \varepsilon\|_\infty \leq \lambda/2, \quad \|n^{-1}(Z^p_{S^c})^\top \varepsilon - C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} n^{-1} (Z^p_{S_{\setminus p}})^\top \varepsilon\|_\infty \leq \lambda\xi. \tag{A.3}$$

In what follows, we condition on the event that (A.3) holds and analyze the optimality conditions (A.1) and (A.2) using deterministic arguments.

First, we take $\hat{\beta}_{S^c} = 0$. Substituting $y = Z^p_{S_{\setminus p}} \beta^*_{S_{\setminus p}} + \varepsilon$ and replacing $\hat{S}$ by S, we write (A.1) as

$$\hat{\beta}_{S_{\setminus p}} - \beta^*_{S_{\setminus p}} = (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} [n^{-1} (Z^p_{S_{\setminus p}})^\top \varepsilon - \lambda \{ \operatorname{sgn}(\hat{\beta}_{S_{\setminus p}}) - \operatorname{sgn}(\hat{\beta}_p) 1_{s-1} \}]. \tag{A.4}$$

Now define $\hat{\beta}_{S_{\setminus p}}$ by (A.4) with $\operatorname{sgn}(\hat{\beta}_{S_{\setminus p}})$ and $\operatorname{sgn}(\hat{\beta}_p)$ replaced by $\operatorname{sgn}(\beta^*_{S_{\setminus p}})$ and $\operatorname{sgn}(\beta^*_p)$, respectively. By (A.3), (A.4), and the triangle inequality, we have

$$\begin{aligned}
\|\hat{\beta}_S - \beta^*_S\|_\infty &= \|D_{SS_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} [n^{-1} (Z^p_{S_{\setminus p}})^\top \varepsilon - \lambda \{ \operatorname{sgn}(\hat{\beta}_{S_{\setminus p}}) - \operatorname{sgn}(\hat{\beta}_p) 1_{s-1} \}] \|_\infty \\
&\leq \|D_{SS_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} (D_{SS_{\setminus p}})^\top \|_\infty \|n^{-1} (Z_S)^\top \varepsilon\|_\infty \\
&\quad + \lambda \|D_{SS_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} (D_{SS_{\setminus p}})^\top \|_\infty \\
&\leq \varphi \lambda / 2 + \varphi \lambda = 3 \varphi \lambda / 2 < \beta_{\min}
\end{aligned}$$

by assumption. This implies that $\operatorname{sgn}(\hat{\beta}_S) = \operatorname{sgn}(\beta^*_S)$, and hence we have found a $\hat{\beta}$ such that the desired properties and (A.1) hold.

It remains to verify that $\hat{\beta}$ also satisfies (A.2). Substituting $y = Z^p_{S_{\setminus p}} \beta^*_{S_{\setminus p}} + \varepsilon$ and (A.4), we write

$$\begin{aligned}
& n^{-1} (Z^p_{S^c})^\top (y - Z^p \hat{\beta}_{\setminus p}) + \lambda \operatorname{sgn}(\hat{\beta}_p) 1_{p-s} \\
&= n^{-1} (Z^p_{S^c})^\top \varepsilon - C^p_{S^c S_{\setminus p}} (\hat{\beta}_{S_{\setminus p}} - \beta^*_{S_{\setminus p}}) + \lambda \operatorname{sgn}(\beta^*_p) 1_{p-s} \\
&= n^{-1} (Z^p_{S^c})^\top \varepsilon - C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} n^{-1} (Z^p_{S_{\setminus p}})^\top \varepsilon \\
&\quad + C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} \lambda \{ \operatorname{sgn}(\beta^*_{S_{\setminus p}}) - \operatorname{sgn}(\beta^*_p) 1_{s-1} \} + \lambda \operatorname{sgn}(\beta^*_p) 1_{p-s}.
\end{aligned}$$

By (A.3), Condition 1, and the triangle inequality, we have

$$\begin{aligned}
& \|n^{-1} (Z^p_{S^c})^\top (y - Z^p \hat{\beta}_{\setminus p}) + \lambda \operatorname{sgn}(\hat{\beta}_p) 1_{p-s}\|_\infty \\
&\leq \|n^{-1} (Z^p_{S^c})^\top \varepsilon - C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} n^{-1} (Z^p_{S_{\setminus p}})^\top \varepsilon\|_\infty \\
&\quad + \lambda \|C^p_{S^c S_{\setminus p}} (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} \{ \operatorname{sgn}(\beta^*_{S_{\setminus p}}) - \operatorname{sgn}(\beta^*_p) 1_{s-1} \} + \operatorname{sgn}(\beta^*_p) 1_{p-s}\|_\infty \\
&\leq \lambda \xi + \lambda (1 - \xi) = \lambda,
\end{aligned}$$

which verifies (A.2) and completes the proof. □

of Proposition 1. To apply Theorem 4 of Rockafellar (1976) for the convergence of the method of multipliers, we need only verify for problem (2.3) that (a) Slater's condition is satisfied, and (b) there exists a constant c such that the c-sublevel set of feasible points $B_c = \{\beta : Q(\beta) \le c \text{ and } \sum_{j=1}^{p} \beta_j = 0\}$ is nonempty and bounded, where $Q(\cdot)$ is the objective function in problem (2.3). Claim (a) holds since in this case Slater's condition reduces to feasibility and 0 is a feasible point. To show (b), take any $c \ge Q(0)$; then B_c is nonempty since $0 \in B_c$, and is bounded since $\|\beta\|_1 \le c/\lambda$ for $\beta \in B_c$. Proposition 1 follows from the aforementioned result. □

Proof of Proposition 2. For $J \subset \{1, \ldots, p-1\}$, let Z^r_J denote the submatrix formed by the jth columns of Z^r with $j \in J$. Define

$$P^r = \begin{pmatrix} I_{r-1} & -1 & 0 \\ 0 & \vdots & I_{p-1-r} \\ 0 & -1 & 0 \end{pmatrix} \in \mathbb{R}^{(p-1)\times(p-1)},$$

and let $E^r \in \mathbb{R}^{(p-1)\times(p-1)}$ denote the matrix with 1s in the rth column and 0s elsewhere. Then we have $\operatorname{sgn}(\beta^*_{S_{\setminus r}}) - \operatorname{sgn}(\beta^*_r)1_{s-1} = P^r_{S_{\setminus r}S_{\setminus p}}\{\operatorname{sgn}(\beta^*_{S_{\setminus p}}) - \operatorname{sgn}(\beta^*_p)1_{s-1}\}$, $Z^r_{S_{\setminus r}} = Z^p_{S_{\setminus p}}(P^r_{S_{\setminus r}S_{\setminus p}})^\top$, and $Z^r_{S^c} = Z^p_{S^c} - Z^p_{S_{\setminus p}}(E^r_{S^c S_{\setminus p}})^\top$. Furthermore,

$$\begin{aligned}
C^r_{S_{\setminus r}S_{\setminus r}} &= n^{-1}(Z^r_{S_{\setminus r}})^\top Z^r_{S_{\setminus r}} = n^{-1}P^r_{S_{\setminus r}S_{\setminus p}}(Z^p_{S_{\setminus p}})^\top Z^p_{S_{\setminus p}}(P^r_{S_{\setminus r}S_{\setminus p}})^\top \\
&= P^r_{S_{\setminus r}S_{\setminus p}} C^p_{S_{\setminus p}S_{\setminus p}}(P^r_{S_{\setminus r}S_{\setminus p}})^\top, \\
C^r_{S^c S_{\setminus r}} &= n^{-1}(Z^r_{S^c})^\top Z^r_{S_{\setminus r}} = n^{-1}\{Z^p_{S^c} - Z^p_{S_{\setminus p}}(E^r_{S^c S_{\setminus p}})^\top\}^\top Z^p_{S_{\setminus p}}(P^r_{S_{\setminus r}S_{\setminus p}})^\top \\
&= n^{-1}\{(Z^p_{S^c})^\top Z^p_{S_{\setminus p}} - E^r_{S^c S_{\setminus p}}(Z^p_{S_{\setminus p}})^\top Z^p_{S_{\setminus p}}\}(P^r_{S_{\setminus r}S_{\setminus p}})^\top \\
&= (C^p_{S^c S_{\setminus p}} - E^r_{S^c S_{\setminus p}} C^p_{S_{\setminus p}S_{\setminus p}})(P^r_{S_{\setminus r}S_{\setminus p}})^\top.
\end{aligned}$$

Substituting these identities into the left-hand side of (2.9) yields

$$
\begin{aligned}
&C^r_{S^c S_{\setminus r}}(C^r_{S_{\setminus r}S_{\setminus r}})^{-1}\{\mathrm{sgn}(\beta^*_{S_{\setminus r}}) - \mathrm{sgn}(\beta^*_r)1_{s-1}\} + \mathrm{sgn}(\beta^*_r)1_{p-s} \\
&\quad = (C^p_{S^c S_{\setminus p}} - E^r_{S^c S_{\setminus p}} C^p_{S_{\setminus p}S_{\setminus p}})(C^p_{S_{\setminus p}S_{\setminus p}})^{-1}\{\mathrm{sgn}(\beta^*_{S_{\setminus p}}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\} + \mathrm{sgn}(\beta^*_r)1_{p-s} \\
&\quad = C^p_{S^c S_{\setminus p}}(C^p_{S_{\setminus p}S_{\setminus p}})^{-1}\{\mathrm{sgn}(\beta^*_{S_{\setminus p}}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\} \\
&\qquad - E^r_{S^c S_{\setminus p}}\{\mathrm{sgn}(\beta^*_{S_{\setminus p}}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\} + \mathrm{sgn}(\beta^*_r)1_{p-s} \\
&\quad = C^p_{S^c S_{\setminus p}}(C^p_{S_{\setminus p}S_{\setminus p}})^{-1}\{\mathrm{sgn}(\beta^*_{S_{\setminus p}}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\} \\
&\qquad - \{\mathrm{sgn}(\beta^*_r) - \mathrm{sgn}(\beta^*_p)\}1_{p-s} + \mathrm{sgn}(\beta^*_r)1_{p-s} \\
&\quad = C^p_{S^c S_{\setminus p}}(C^p_{S_{\setminus p}S_{\setminus p}})^{-1}\{\mathrm{sgn}(\beta^*_{S_{\setminus p}}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\} + \mathrm{sgn}(\beta^*_p)1_{p-s},
\end{aligned}
$$

and (2.10) follows similarly. □

A.2. Proofs for Chapter 3

Definition 1. *For any matrix M, define the restricted orthogonal constant (ROC) of order k_1 and k_2 as below*

$$
\theta_{k_1,k_2}(M) = \sup\Big\{ \frac{|\langle M\alpha_1, M\alpha_2\rangle|}{||\alpha_1||_2||\alpha_2||_2} : \alpha_1 \textit{ is } k_1\textit{-sparse vector}, \alpha_2 \textit{ is } k_2\textit{-sparse vector}, \\
\alpha_1 \textit{ and } \alpha_2 \textit{ have non-overlapping support}\Big\}
$$

Before proving Theorem 2, we need to present Lemma 4 and Theorem 4.

Lemma 4. *Suppose $||\mathbf{I}_p - P_C||_\infty \le k_0$, then for any matrix A, we have*

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

Proof of Lemma 4. By definition of $||\cdot||_\infty$ for matrices, for any vector $a \in \mathbb{R}^p$,

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

□

Theorem 4. *Let $\widehat{\beta}^n$ be the estimator obtained by solving optimization problem* (3.6) *for Model* (3.5)*,*

where β is s-sparse. If $(3\tau-1)\delta_{2s}^{-}(\widetilde{\mathbf{Z}}/\sqrt{n})-(\tau+1)\delta_{2s}^{+}(\widetilde{\mathbf{Z}}/\sqrt{n})\geq 4\tau\phi_0$ for some constant $\phi_0>0$, and $||\widetilde{\mathbf{Z}}^\top\varepsilon||_\infty\leq n\lambda/\tau$, then,

$$||\widehat{\beta}^n-\beta||_1\leq s\lambda(k_0+1/\tau)/\phi_0.$$

Proof. By the definition of $\widehat{\beta}^n$, we have

$$\frac{1}{2n}||y-\widetilde{\mathbf{Z}}\widehat{\beta}^n||_2^2+\lambda||\widehat{\beta}^n||_1\leq\frac{1}{2n}||y-\widetilde{\mathbf{Z}}\beta||_2^2+\lambda||\beta||_1.$$

Denote $h=\widehat{\beta}^n-\beta$, and S_h be the set of index of the s largest absolute values of h.
Then by $Y=\widetilde{\mathbf{Z}}\beta+\varepsilon$, we have

$$\frac{1}{2n}(||\varepsilon-\widetilde{\mathbf{Z}}h||_2^2-||\varepsilon||_2^2)\leq\lambda(||\beta||_1-||\widehat{\beta}^n||_1). \tag{A.5}$$

Notice that

$$\begin{aligned}||\beta||_1-||\widehat{\beta}^n||_1&=||\beta_{supp(\beta)}||_1-||\widehat{\beta}^n_{supp(\beta)}||_1-||\widehat{\beta}^n_{supp(\beta)^c}||_1\\&\leq||\beta_{supp(\beta)}-\widehat{\beta}^n_{supp(\beta)}||_1-||h_{supp(\beta)^c}||_1\\&\leq||h_{supp(\beta)}||_1-||h_{supp(\beta)^c}||_1\\&\leq||h_{S_h}||_1-||h_{S_h^c}||_1.\end{aligned}$$

Also,

$$\begin{aligned}\frac{1}{2n}(||\varepsilon-\widetilde{\mathbf{Z}}h||_2^2-||\varepsilon||_2^2)&=-\frac{1}{2n}(\widetilde{\mathbf{Z}}h)^\top(2\varepsilon-\widetilde{\mathbf{Z}}h)\geq-\frac{1}{n}h^\top\widetilde{\mathbf{Z}}^\top\varepsilon\geq-\frac{1}{n}||\widetilde{\mathbf{Z}}^\top\varepsilon||_\infty||h||_1\\&=-\frac{1}{n}||\widetilde{\mathbf{Z}}^\top\varepsilon||_\infty(||h_{S_h}||_1+||h_{S_h^c}||_1).\end{aligned}$$

Then, by $||\widetilde{\mathbf{Z}}^\top\varepsilon||_\infty\leq n\lambda/\tau$ and (A.5), we have

$$-(||h_{S_h}||_1+||h_{S_h^c}||_1)\leq\tau(||h_{S_h}||_1-||h_{S_h^c}||_1).$$

Therefore,

$$||h_{S_h^c}||_1\leq\frac{\tau+1}{\tau-1}||h_{S_h}||_1. \tag{A.6}$$

By the KKT condition of optimization problem (3.6), we have

$$||\widetilde{\mathbf{Z}}^\top(y-\widetilde{\mathbf{Z}}\widehat{\beta}^n)+C\mu||_\infty \leq n\lambda$$

for some $\mu \in \mathbb{R}^r$. Then by Lemma 4,

$$\begin{aligned}||\widetilde{\mathbf{Z}}^\top(y-\widetilde{\mathbf{Z}}\widehat{\beta}^n)||_\infty &= ||(\mathbf{I}_p - P_C)(\widetilde{\mathbf{Z}}^\top(y-\widetilde{\mathbf{Z}}\widehat{\beta}^n)+C\mu)||_\infty \\ &\leq k_0||\widetilde{\mathbf{Z}}^\top(y-\widetilde{\mathbf{Z}}\widehat{\beta}^n)+C\mu||_\infty \leq k_0 n\lambda.\end{aligned}$$

Then

$$||\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{Z}}h||_\infty \leq ||\widetilde{\mathbf{Z}}^\top(y-\widetilde{\mathbf{Z}}\widehat{\beta}^n)||_\infty + ||\widetilde{\mathbf{Z}}^\top(y-\widetilde{\mathbf{Z}}\beta)||_\infty \leq k_0 n\lambda + ||\widetilde{\mathbf{Z}}^\top\varepsilon||_\infty.$$

Using Lemma 5.1 in Cai and Zhang (2013), we can get

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

Then,

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

The last inequality comes from $\theta_{k_1,k_2}(A) \leq \frac{1}{2}(\delta_{k_1+k_2}^+(A) - \delta_{k_1+k_2}^-(A))$ for any matrix A from Lemma 1 of Kang et al. (2015). The inequality above gives us

$$||h_{S_h}||_1 \leq s\frac{k_0 n\lambda + ||\widetilde{\mathbf{Z}}^\top\varepsilon||_\infty}{\frac{n}{2(\tau-1)}\left((3\tau-1)\delta_{2s}^-(\widetilde{\mathbf{Z}}/\sqrt{n}) - (\tau+1)\delta_{2s}^+(\widetilde{\mathbf{Z}}/\sqrt{n})\right)} = s\frac{k_0 n\lambda + ||\widetilde{\mathbf{Z}}^\top\varepsilon||_\infty}{2n\tau\phi_0/(\tau-1)}.$$

Therefore, by $||\widetilde{\mathbf{Z}}^\top \varepsilon||_\infty \leq n\lambda/\tau$ and (A.6), we have

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

□

Proof of Theorem 2.

$$\begin{aligned}
\widehat{\beta}^u - \beta &= \widehat{\beta}^n - \beta + \frac{1}{n}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top \varepsilon + \frac{1}{n}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{Z}}(\beta - \widehat{\beta}^n) \\
&= \frac{1}{n}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top \varepsilon + (\widetilde{\mathbf{M}}\widehat{\Sigma} - \mathbf{I}_p)(\beta - \widehat{\beta}^n) \\
&= \frac{1}{n}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top \varepsilon + (\widetilde{\mathbf{M}}\widehat{\Sigma} - \mathbf{I}_p + P_C)(\beta - \widehat{\beta}^n), \quad \text{since } C^\top \beta = C^\top \widehat{\beta}^n = 0.
\end{aligned}$$

Thus, $\sqrt{n}(\widehat{\beta}^u - \beta) = B + \Delta$ where $B = \frac{1}{\sqrt{n}}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top \varepsilon$.
Notice that $(\mathbf{I}_p - P_C)\widehat{\Sigma}(\mathbf{I}_p - P_C) = \widehat{\Sigma}$, and $B = \frac{1}{\sqrt{n}}\widetilde{\mathbf{M}}\widetilde{\mathbf{Z}}^\top \varepsilon = \frac{1}{\sqrt{n}}\widetilde{\mathbf{M}}(\mathbf{I}_p - P_C)\mathbf{Z}^\top \varepsilon$. Thus,

$$B|\mathbf{Z} \sim N\left(0, \sigma^2 \widetilde{\mathbf{M}}(\mathbf{I}_p - P_C)\widehat{\Sigma}(\mathbf{I}_p - P_C)\widetilde{\mathbf{M}}^\top\right) = N(0, \sigma^2 \widetilde{\mathbf{M}}\widehat{\Sigma}\widetilde{\mathbf{M}}^\top).$$

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

The last inequality is by Lemma 4.

By Lemma 1, when choosing $\gamma = c\sqrt{(\log p)/n}$, Ω is a feasible solution of the optimization problem (3.10) with probability at least $1 - 2p^{-c''}$. Therefore, $\left|M\widehat{\Sigma} - (\mathbf{I}_p - P_C)\right|_\infty \leq \gamma = c\sqrt{(\log p)/n}$ with probability at least $1 - 2p^{c''}$.

By Theorem 4, take $\lambda = \tau\tilde{c}\sigma\sqrt{(\log p)/n}$,

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

Altogether, we have

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

□

Proof of Lemma 1. Note that $\Sigma^{1/2}\Omega^{1/2}\widetilde{Z}_l = (\mathbf{I}_p - P_C)\widetilde{Z}_l = \widetilde{Z}_l$. Therefore,

$$\begin{aligned}
\Omega\widehat{\Sigma} - (\mathbf{I}_p - P_C) &= \frac{1}{n}\sum_{l=1}^{n}\left\{\Omega\widetilde{Z}_l\widetilde{Z}_l^\top - (\mathbf{I}_p - P_C)\right\} \\
&= \frac{1}{n}\sum_{l=1}^{n}\left\{\Omega^{1/2}\Omega^{1/2}\widetilde{Z}_l\widetilde{Z}_l^\top\Omega^{1/2}\Sigma^{1/2} - (\mathbf{I}_p - P_C)\right\}.
\end{aligned}$$

Define $v_l^{(ij)} = \Omega_{i,\cdot}^{1/2}\Omega^{1/2}\widetilde{Z}_l\widetilde{Z}_l^\top\Omega^{1/2}\Sigma_{\cdot,j}^{1/2} - (\mathbf{I}_p - P_C)_{i,j}$. Since $\mathbb{E}\Omega\widetilde{Z}_l\widetilde{Z}_l^\top = \Omega\Sigma = (\mathbf{I}_p - P_C)$, we have $\mathbb{E}v_l^{(ij)} = 0$. Then, by the proof of Lemma 6.2 in Javanmard and Montanari (2014),

$$\begin{aligned}
||v_l^{(ij)}||_{\psi_1} &\leq 2||\Omega_{i,\cdot}^{1/2}\Omega^{1/2}\widetilde{Z}_l\widetilde{Z}_l^\top\Omega^{1/2}\Sigma_{\cdot,j}^{1/2}||_{\psi_1} \\
&\leq 2||\Omega_{i,\cdot}^{1/2}\Omega^{1/2}\widetilde{Z}_l||_{\psi_2}||\Sigma_{j,\cdot}^{1/2}\Omega^{1/2}\widetilde{Z}_l||_{\psi_2} \\
&\leq 2||\Omega_{i,\cdot}^{1/2}||_2||\Sigma_{j,\cdot}^{1/2}||_2||\Omega^{1/2}\widetilde{Z}_l||_{\psi_2}||\Omega^{1/2}\widetilde{Z}_l||_{\psi_2} \\
&\leq 2\sqrt{\sigma_{\max}(\Sigma)\sigma_{\max}(\Omega)}\kappa^2 \\
&\leq 2\sqrt{C_{\max}/C_{\min}}\kappa^2 \equiv \kappa',
\end{aligned}$$

where $||X||_{\psi_1}$ is the sub-exponential norm of a random variable X and is defined as

$$||X||_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|X|^p)^{1/p}.$$

Applying Bernstein-type inequality for centered sub-exponential random variables (Bühlmann and Van De Geer, 2011), we get

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

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

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

Therefore, by union bounding over all pairs of i and j,

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

□

Proof of Lemma 2. Suppose $\widetilde{\mathbf{M}} = (\widetilde{m}_1, \ldots, \widetilde{m}_p)^\top$. Since

$$\left| \widetilde{\mathbf{M}}\widehat{\Sigma} - (\mathbf{I}_p - P_C) \right|_\infty \leq k_0 \left| M\widehat{\Sigma} - (\mathbf{I}_p - P_C) \right|_{\max} \leq k_0\gamma,$$

we have $1 - P_{Ci,i} - e_i^\top \widehat{\Sigma}\widetilde{m}_i \leq k_0\gamma$. Therefore, for all $L \geq 0$,

$$\begin{aligned}
\widetilde{m}_i^\top \widehat{\Sigma}\widetilde{m}_i &\geq \widetilde{m}_i^\top \widehat{\Sigma}\widetilde{m}_i + L(1 - P_{Ci,i} - k_0\gamma) - Le_i^\top \widehat{\Sigma}\widetilde{m}_i \\
&\geq \min_m \left\{ m^\top \widehat{\Sigma} m + L(1 - P_{Ci,i} - k_0\gamma) - Le_i^\top \widehat{\Sigma} m \right\} \\
&= L(1 - P_{Ci,i} - k_0\gamma) - \frac{L^2}{4}\widehat{\Sigma}_{i,i} \quad \text{(The minimizer } m = Le_i/2\text{)} \\
&\geq \min_{L \geq 0} \left\{ L(1 - P_{Ci,i} - k_0\gamma) - \frac{L^2}{4}\widehat{\Sigma}_{i,i} \right\} \\
&\geq \frac{(1 - P_{Ci,i} - k_0\gamma)^2}{\widehat{\Sigma}_{i,i}} \quad \text{(take } L = 2(1 - P_{Ci,i} - k_0\gamma)/\widehat{\Sigma}_{i,i}\text{)}.
\end{aligned}$$

□

A.3. Proofs for Chapter 4

Proof of Lemma 3. By (4.6), we have

$$\begin{aligned} \mathrm{Var}\hat{\pi}_t &= \frac{\sum_{i=1}^n N_{it}^2 - N_{\cdot t}}{N_{\cdot t}^2}\Sigma_t + \frac{1}{N_{\cdot t}}\big(diag(\pi_t) - \pi_t\pi_t^\top\big) \\ \mathrm{Cov}(\hat{\pi}_1, \hat{\pi}_2) &= \frac{\sum_{i=1}^n N_{i1}N_{i2}}{N_{\cdot 1}N_{\cdot 2}}\Sigma_{12} \end{aligned}$$

It can also be shown that

$$\begin{aligned} \mathbb{E}(\mathbf{S}_t - \mathbf{G}_t) &= N_{ct}\Sigma_t \\ \mathbb{E}\big(\mathbf{S}_t + (N_{ct}-1)\mathbf{G}_t\big) &= N_{ct}\big(diag(\pi_t) - \pi_t\pi_t^\top\big) \\ \mathbb{E}\widehat{\Sigma}_{12} &= \Sigma_{12} \end{aligned}$$

Thus,

$$\begin{aligned} \Sigma_\pi &= \mathrm{Var}(\hat{\pi}_1 - \hat{\pi}_2) \\ &= \frac{\sum_{i=1}^n N_{it}^2 - N_{\cdot t}}{N_{ct}N_{\cdot t}^2}\mathbb{E}(\mathbf{S}_t - \mathbf{G}_t) + \frac{1}{N_{ct}N_{\cdot t}}\mathbb{E}\big(\mathbf{S}_t + (N_{ct}-1)\mathbf{G}_t\big) \\ &\quad - \frac{\sum_{i=1}^n N_{i1}N_{i2}}{N_{\cdot 1}N_{\cdot 2}}\mathbb{E}\big(\widehat{\Sigma}_{12} + \widehat{\Sigma}_{12}^\top\big) \end{aligned} \tag{A.7}$$

By central limit theorem, we have

$$\begin{aligned} ||(\mathbf{S}_t - \mathbf{G}_t) - \mathbb{E}(\mathbf{S}_t - \mathbf{G}_t)||_{\max} &\to 0 \\ ||\big(\mathbf{S}_t + (N_{ct}-1)\mathbf{G}_t\big) - \mathbb{E}\big(\mathbf{S}_t + (N_{ct}-1)\mathbf{G}_t\big)||_{\max} &\to 0 \\ ||\Sigma_{12} - \mathbb{E}\widehat{\Sigma}_{12}||_{\max} &\to 0 \end{aligned} \tag{A.8}$$

Combining (4.8), (A.7) and (A.8), we have

$$||\widehat{\Sigma}_\pi - \Sigma_\pi||_{\max} \to 0 \quad \text{in probability as } n \to \infty$$

□

Proof of Theorem 3. Define $S^{d-1} = \{x \in \mathbb{R}^p : 1^\top x = 0\}$. Then $\pi_1 - \pi_2, \hat{\pi}_1 - \hat{\pi}_2 \in S^{d-1}$. Therefore

$$(\hat{\pi}_1 - \hat{\pi}_2)^\top \Sigma_\pi^\dagger (\hat{\pi}_1 - \hat{\pi}_2) \to \chi^2_{d-1}$$

Now we are going to show that $\widehat{\Sigma}_\pi^\dagger \to \Sigma_\pi^\dagger$ in probability.

Let Γ be a projection matrix in the form of $[\mathbf{V}, 1_d^\top/\sqrt{d}]$. Then, because $1_d^\top \widehat{\Sigma}_\pi = 1_d^\top \Sigma_\pi = 0$, by Lemma 3, we have

$$||\mathbf{V}^\top(\widehat{\Sigma}_\pi - \Sigma_\pi)\mathbf{V}||_2/d \le ||\mathbf{V}^\top(\widehat{\Sigma}_\pi - \Sigma_\pi)\mathbf{V}||_{\max} = ||\Gamma^\top(\widehat{\Sigma}_\pi - \Sigma_\pi)\Gamma||_{\max} \to 0 \quad \text{in probability}$$

where $||\cdot||_2$ is the spectral norm of matrix.

Define $\Delta = \mathbf{V}^\top(\widehat{\Sigma}_\pi - \Sigma_\pi)\mathbf{V}$. Using Neumann series expansion,

$$\begin{aligned}
&(\mathbf{V}^\top \widehat{\Sigma}_\pi \mathbf{V})^{-1} - (\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1} = \sum_{i=1}^{\infty} \left((\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1}\Delta\right)^n (\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1} \\
\Rightarrow &||(\mathbf{V}^\top \widehat{\Sigma}_\pi \mathbf{V})^{-1} - (\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1}||_2 \le \sum_{i=1}^{\infty} ||\mathbf{V}^\top \Sigma_\pi \mathbf{V}||_2^{i+1} ||\Delta||_2^i \to 0 \quad \text{in probability}
\end{aligned}$$

Therefore,

$$\begin{aligned}
& ||(\mathbf{V}^\top \widehat{\Sigma}_\pi \mathbf{V})^{-1} - (\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1}||_{\max} \le ||(\mathbf{V}^\top \widehat{\Sigma}_\pi \mathbf{V})^{-1} - (\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1}||_2 \to 0 \quad \text{in probability} \\
\Rightarrow \quad & ||\mathbf{V}(\mathbf{V}^\top \widehat{\Sigma}_\pi \mathbf{V})^{-1}\mathbf{V}^\top - \mathbf{V}(\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1}\mathbf{V}^\top||_{\max} \to 0 \quad \text{in probability} \qquad \text{(A.9)}
\end{aligned}$$

Suppose we have the eigenvalue decomposition of Σ_π as $\Sigma_\pi = \mathbf{U}\Lambda\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{d\times(d-1)}$ and $\Lambda \in \mathbb{R}^{(d-1)\times(d-1)}$. Then $1_d^\top \mathbf{U} = 0$. Also, $\mathbf{UV}$ is orthogonal because

$$\mathbf{UVV}^\top\mathbf{U}^\top = \mathbf{U}(\mathbf{I}_d - 1_d 1_d^\top/d)\mathbf{U}^\top = \mathbf{UU}^\top = \mathbf{I}_{d-1}$$

Therefore,

$$\begin{aligned}
&\mathbf{V}(\mathbf{V}^\top \Sigma_\pi \mathbf{V})^{-1}\mathbf{V}^\top = \mathbf{V}(\mathbf{V}^\top \mathbf{U}\Lambda\mathbf{U}^\top \mathbf{V})^{-1}\mathbf{V}^\top \\
&= \mathbf{V}(\mathbf{U}^\top \mathbf{V})^{-1}\Lambda^{-1}(\mathbf{V}^\top \mathbf{U})^{-1}\mathbf{V}^\top = \mathbf{V}(\mathbf{U}^\top \mathbf{V})^\top \Lambda^{-1}(\mathbf{V}^\top \mathbf{U})^\top \mathbf{V}^\top \\
&= (\mathbf{I}_d - 1_d 1_d^\top/d)\mathbf{U}\Lambda^{-1}\mathbf{U}^\top(\mathbf{I}_d - 1_d 1_d^\top/d) = \mathbf{U}\Lambda^{-1}\mathbf{U}^\top \\
&= \Sigma_\pi^\dagger
\end{aligned}$$

Similarly, we have

$$\widehat{\Sigma}_\pi^\dagger = \mathbf{V}(\mathbf{V}^\top \widehat{\Sigma}_\pi \mathbf{V})^{-1}\mathbf{V}^\top$$

Combining (A.9), we have

$$||\widehat{\Sigma}_\pi^\dagger - \Sigma_\pi^\dagger||_{\max} \to 0 \text{ in probability}$$

Then by Slutsky Theorem, for fixed d and $n \to \infty$,

$$F = \frac{n-d+1}{(n-1)(d-1)}(\hat{\pi}_1 - \hat{\pi}_2)^\top \widehat{\Sigma}_\pi^\dagger(\hat{\pi}_1 - \hat{\pi}_2) \to \chi^2_{d-1}/(d-1)$$

Since $F_{d-1,n-d+1} \to \chi^2_{d-1}/(d-1)$ for fixed d and $n \to \infty$, $F > F^{-1}_{d-1,n-d+1}(1-\alpha)$ is an asymptotic level α test.

□

BIBLIOGRAPHY

Aitchison, J (2003). *The statistical analysis of compositional data*. Blackburn Press.

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

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

Altschul, SF, Gish, W, Miller, W, Myers, EW, and Lipman, DJ (1990). Basic local alignment search tool. *Journal of molecular biology* 215.3, 403–410.

Anderson, MJ (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology* 26.1, 32–46.

Asnicar, F, Weingart, G, Tickle, TL, Huttenhower, C, and Segata, N (2015). Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* 3, e1029.

Atchison, J and Shen, SM (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* 67.2, 261–272.

Benjamini, Y and Hochberg, Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bertsekas, DP (2014). *Constrained optimization and Lagrange multiplier methods*. Academic press.

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

Bühlmann, P et al. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* 19.4, 1212–1242.

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

Cani, PD and Delzenne, NM (2011). The gut microbiome as therapeutic target. *Pharmacology & therapeutics* 130.2, 202–212.

Caporaso, JG, Kuczynski, J, Stombaugh, J, Bittinger, K, Bushman, FD, Costello, EK, Fierer, N, Pena, AG, Goodrich, JK, Gordon, JI, et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* 7.5, 335–336.

Charlson, ES, Chen, J, Custers-Allen, R, Bittinger, K, Li, H, Sinha, R, Hwang, J, Bushman, FD, and Collman, RG (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PloS one* 5.12, e15216.

Chen, J and Li, H (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The annals of applied statistics* 7.1.

Chen, J, Bittinger, K, Charlson, ES, Hoffmann, C, Lewis, J, Wu, GD, Collman, RG, Bushman, FD, and Li, H (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28.16, 2106–2113.

Chen, J, Bushman, FD, Lewis, JD, Wu, GD, and Li, H (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* 14.2, 244–258.

CLARKE, KR (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology* 18.1, 117–143.

Cole, JR, Chai, B, Farris, RJ, Wang, Q, Kulam-Syed-Mohideen, A, McGarrell, DM, Bandela, A, Cardenas, E, Garrity, GM, and Tiedje, JM (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic acids research* 35.suppl 1, D169–D172.

Consortium, HMP et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486.7402, 207–214.

Cornell, JA (2011). *Experiments with mixtures: designs, models, and the analysis of mixture data*. Vol. 895. John Wiley & Sons.

DeSantis, TZ, Hugenholtz, P, Larsen, N, Rojas, M, Brodie, EL, Keller, K, Huber, T, Dalevi, D, Hu, P, and Andersen, GL (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* 72.7, 5069–5072.

Dicker, LH (2014). Variance estimation in high-dimensional linear models. *Biometrika* 101.2, 269–284.

Efron, B (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109.507, 991–1007.

Ehrlich, SD, Consortium, M, et al. (2011). MetaHIT: The European Union Project on metagenomics of the human intestinal tract. In: *Metagenomics of the Human Body*. Springer, 307–316.

Evans, SN and Matsen, FA (2012). The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, 569–592.

Fan, J, Guo, S, and Hao, N (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.1, 37–65.

Fan, Y and Tang, CY (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.3, 531–552.

Flores, GE, Caporaso, JG, Henley, JB, Rideout, JR, Domogala, D, Chase, J, Leff, JW, Vázquez-Baeza, Y, Gonzalez, A, Knight, R, et al. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biol* 15.12, 531.

Friedman, J, Hastie, T, Höfling, H, Tibshirani, R, et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* 1.2, 302–332.

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

Gevers, D, Kugathasan, S, Denson, LA, Vázquez-Baeza, Y, Van Treuren, W, Ren, B, Schwager, E, Knights, D, Song, SJ, Yassour, M, et al. (2014). The treatment-naive microbiome in new-onset Crohns disease. *Cell host & microbe* 15.3, 382–392.

Gill, SR, Pop, M, DeBoy, RT, Eckburg, PB, Turnbaugh, PJ, Samuel, BS, Gordon, JI, Relman, DA, Fraser-Liggett, CM, and Nelson, KE (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312.5778, 1355–1359.

Gower, JC (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53.3-4, 325–338.

Grant, M and Boyd, S (2013). *CVX: Matlab software for disciplined convex programming, version 2.0 beta*. Tech. rep. http://cvxr.com/cvx, September 2013.

Grice, EA, Kong, HH, Conlan, S, Deming, CB, Davis, J, Young, AC, Bouffard, GG, Blakesley, RW, Murray, PR, Green, ED, et al. (2009). Topographical and temporal diversity of the human skin microbiome. *science* 324.5931, 1190–1192.

Hsiao, EY, McBride, SW, Hsien, S, Sharon, G, Hyde, ER, McCue, T, Codelli, JA, Chow, J, Reisman, SE, Petrosino, JF, et al. (2013). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* 155.7, 1451–1463.

Huson, DH, Auch, AF, Qi, J, and Schuster, SC (2007). MEGAN analysis of metagenomic data. *Genome research* 17.3, 377–386.

James, GM, Paulson, C, and Rusmevichientong, P (2015). *Penalized and constrained regression*.

Janson, L, Barber, RF, and Candès, E (2015). Eigenprism: Inference for high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1505.02097*.

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

Kang, H, Zhang, A, Cai, TT, and Small, DS (2015). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association* just-accepted.

Kim, OS, Cho, YJ, Lee, K, Yoon, SH, Kim, M, Na, H, Park, SC, Jeon, YS, Lee, JH, Yi, H, et al. (2012). Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *International journal of systematic and evolutionary microbiology* 62.3, 716–721.

Kong, HH, Oh, J, Deming, C, Conlan, S, Grice, EA, Beatson, MA, Nomicos, E, Polley, EC, Komarow, HD, Murray, PR, et al. (2012). Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome research* 22.5, 850–859.

Kostic, AD, Gevers, D, Pedamallu, CS, Michaud, M, Duke, F, Earl, AM, Ojesina, AI, Jung, J, Bass, AJ, Tabernero, J, et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome research* 22.2, 292–298.

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

La Rosa, PS, Brooks, JP, Deych, E, Boone, EL, Edwards, DJ, Wang, Q, Sodergren, E, Weinstock, G, and Shannon, WD (2012). Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PloS one* 7.12, e52078.

Lam, YY, Ha, CW, Campbell, CR, Mitchell, AJ, Dinudom, A, Oscarsson, J, Cook, DI, Hunt, NH, Caterson, ID, Holmes, AJ, et al. (2012). Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice. *PloS one* 7.3, e34233.

Lee, JD, Sun, DL, Sun, Y, and Taylor, JE (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, in press.

Ley, RE, Bäckhed, F, Turnbaugh, P, Lozupone, CA, Knight, RD, and Gordon, JI (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* 102.31, 11070–11075.

Ley, RE, Turnbaugh, PJ, Klein, S, and Gordon, JI (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444.7122, 1022–1023.

Lin, W, Shi, P, Feng, R, and Li, H (2014). Variable selection in regression with compositional covariates. *Biometrika*, asu031.

Liu, Z, DeSantis, TZ, Andersen, GL, and Knight, R (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic acids research* 36.18, e120–e120.

Lozupone, C and Knight, R (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* 71.12, 8228–8235.

Mandal, S, Van Treuren, W, White, RA, Eggesbø, M, Knight, R, and Peddada, SD (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease* 26.

Manichanh, C, Borruel, N, Casellas, F, and Guarner, F (2012). The gut microbiota in IBD. *Nature Reviews Gastroenterology and Hepatology* 9.10, 599–608.

Mantel, N (1967). The detection of disease clustering and a generalized regression approach. *Cancer research* 27.2 Part 1, 209–220.

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

Meyer, F, Paarmann, D, D'Souza, M, Olson, R, Glass, EM, Kubal, M, Paczian, T, Rodriguez, A, Stevens, R, Wilke, A, et al. (2008). The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics* 9.1, 386.

Morgan, JL, Darling, AE, and Eisen, JA (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PloS one* 5.4, e10209.

Morris, A, Beck, JM, Schloss, PD, Campbell, TB, Crothers, K, Curtis, JL, Flores, SC, Fontenot, AP, Ghedin, E, Huang, L, et al. (2013). Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *American journal of respiratory and critical care medicine* 187.10, 1067–1075.

Peterson, J, Garges, S, Giovanni, M, McInnes, P, Wang, L, Schloss, JA, Bonazzi, V, McEwen, JE, Wetterstrand, KA, Deal, C, et al. (2009). The NIH human microbiome project. *Genome research* 19.12, 2317–2323.

Qin, J, Li, R, Raes, J, Arumugam, M, Burgdorf, KS, Manichanh, C, Nielsen, T, Pons, N, Levenez, F, Yamada, T, et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *nature* 464.7285, 59–65.

Qin, J, Li, Y, Cai, Z, Li, S, Zhu, J, Zhang, F, Liang, S, Zhang, W, Guan, Y, Shen, D, et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490.7418, 55–60.

Rockafellar, RT (1976). Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research* 1.2, 97–116.

Sanger, F, Nicklen, S, and Coulson, AR (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 74.12, 5463–5467.

Segata, N, Izard, J, Waldron, L, Gevers, D, Miropolsky, L, Garrett, WS, and Huttenhower, C (2011). Metagenomic biomarker discovery and explanation. *Genome Biol* 12.6, R60.

Segata, N, Waldron, L, Ballarini, A, Narasimhan, V, Jousson, O, and Huttenhower, C (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* 9.8, 811–814.

Seshadri, R, Kravitz, SA, Smarr, L, Gilna, P, and Frazier, M (2007). CAMERA: a community resource for metagenomics. *PLoS biology* 5.3.

Shi, P and Li, H (2016). A Model for Paired-Multinomial Data and Its Application to Analysis of Data on a Taxonomic Tree, technical report.

Shi, P, Zhang, A, and Li, H (2016). Regression Analysis for Microbiome Compositional Data. *Annals of Applied Statistics*, in press.

Smits, LP, Bouter, KE, Vos, WM de, Borody, TJ, and Nieuwdorp, M (2013). Therapeutic potential of fecal microbiota transplantation. *Gastroenterology* 145.5, 946–953.

Snee, RD (1973). Techniques for the analysis of mixture data. *Technometrics* 15.3, 517–528.

Spor, A, Koren, O, and Ley, R (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology* 9.4, 279–290.

Sun, T and Zhang, CH (2012). Scaled sparse linear regression. *Biometrika*, ass043.

Sunagawa, S, Mende, DR, Zeller, G, Izquierdo-Carrasco, F, Berger, SA, Kultima, JR, Coelho, LP, Arumugam, M, Tap, J, Nielsen, HB, et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods* 10.12, 1196–1199.

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

Tibshirani, RJ, Taylor, JE, Candes, EJ, and Hastie, T (2011). *The solution path of the generalized lasso*. Stanford University.

Torgerson, WS (1958). Theory and methods of scaling.

Turnbaugh, PJ, Ley, RE, Mahowald, MA, Magrini, V, Mardis, ER, and Gordon, JI (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *nature* 444.7122, 1027–131.

Turnbaugh, PJ, Ley, RE, Hamady, M, Fraser-Liggett, C, Knight, R, and Gordon, JI (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449.7164, 804.

Virgin, HW and Todd, JA (2011). Metagenomics and personalized medicine. *Cell* 147.1, 44–56.

Wainwright, MJ (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *Information Theory, IEEE Transactions on* 55.5, 2183–2202.

Walker, AW, Ince, J, Duncan, SH, Webster, LM, Holtrop, G, Ze, X, Brown, D, Stares, MD, Scott, P, Bergerat, A, et al. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *The ISME journal* 5.2, 220–230.

Wang, H and Xia, Y (2012). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*.

Wilson, JR (1989). Chi-square tests for overdispersion with multiparameter estimates. *Applied Statistics*, 441–453.

Wu, GD, Chen, J, Hoffmann, C, Bittinger, K, Chen, YY, Keilbaugh, SA, Bewtra, M, Knights, D, Walters, WA, Knight, R, et al. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334.6052, 105–108.

Wu, GD, Compher, C, Chen, EZ, Smith, SA, Shah, RD, Bittinger, K, Chehoud, C, Albenberg, LG, Nessel, L, Gilroy, E, et al. (2014). Comparative metabolomics in vegans and omnivores reveal constraints on diet-dependent gut microbiota metabolite production. *Gut*, gutjnl–2014.

Yatsunenko, T, Rey, FE, Manary, MJ, Trehan, I, Dominguez-Bello, MG, Contreras, M, Magris, M, Hidalgo, G, Baldassano, RN, Anokhin, AP, et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486.7402, 222–227.

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

Zhao, P and Yu, B (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* 7, 2541–2563.