DISTANCE BASED LINEAR REGRESSION MODEL AND ITS APPLICATION TO MICROBIOME ASSOCIATION STUDIES

by

Masoumeh Sheikhi Kiasari

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2021

Approved by:

_____
Dr. Shaoyu Li

_____
Dr. Yanqing Sun

_____
Dr. Eliana Christou

_____
Dr. Donna Kazemi

ABSTRACT

MASOUMEH SHEIKHI KIASARI. Distance Based Linear Regression Model And Its Application To Microbiome Association Studies. (Under the direction of DR. SHAOYU LI)

In the past few decades, pairwise distance based statistical methods have been developed to identify spatial and/or temporal clusters of disease, study the association between the dissimilarity of ecological communities and distance in geographical locations. With emergence of high-throughput technologies, pairwise distance base methods are widely used in the analysis of genetics and genomics data, especially when the data structure fails the fundamental assumptions of classical multivariate analysis, including independency and normality. However, much of existing knowledge has been around non-parametric or semi-parametric estimations usually employing permutation techniques to assess statistical significance, which are known to be computationally expensive and sensitive to the choice of permutation.

Majority of this thesis focuses on linear regression of pairwise distance matrices. We consider the pairwise correlation structure between the distances and investigate the large sample properties of the ordinary least squares estimator of the model coefficients. Extensive simulations are conducted to evaluate the performance of our method with finite sample size.

Another major component of the thesis is the human microbiome data analysis. We analyze the integrative Human Microbiome Project (iHMP) data set of composition of microbial communities in the digestive tracts of humans by using multiple statistical methods, including our proposed method. The results are presented and interpreted. Existing challenges and future works are also discussed.

ACKNOWLEDGEMENTS

First and foremost I am extremely grateful to my advisor, Dr. Shaoyu Li, whose expertise was invaluable in formulating the research questions and methodology. This thesis would have not been possible without your insightful advice, continuous support, and patience during my PhD. You went above and beyond to help me continue my research and prepare for job interviews at the same time. I cannot express in words how thankful I am.

I would also like to extend my gratitude to the Mathematics Department's Graduate Coordinator Dr. Shaozhong Deng, Committee co-Chair Dr. Yanqing Sun, Committee Member Dr. Eliana Christou, and Graduate Faculty Representative Dr. Donna Kazemi for their invaluable feedback and accommodation.

I would also like to thank the Mathematics Department's System Administrator Mark Hamrick for his advice as well as providing computational resources. Also, I would like to acknowledge University Research Computing for proving high-performance computing, hpc.

I would also like to extend my appreciation to Graduate School's Assistant Teaching Professor of Writing Dr. Lisa Russell-Pinson for her notable Dissertation Writing course and Dissertation Writers support group.

Last but not least, I gratefully acknowledge the financial support received towards my PhD from the Graduate School and Mathematics Department.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: DISTANCE–BASED MULTIVARIATE ANALYSIS

1.1 Background

Non-standard structured, high dimensional multivariate data are now emerging in many modern research fields, including neuroimaging, ecology, genomics, and human microbiome studies. It is of great interest to study the functional relationship between these multivariate variables and some status of the cohort. It could be the association between two groups of multivariate variables, such as the relationship between the ecological system and spatial location; or the association between multivariate dependent variable and a univariate independent variable and vice versa. For example the association between the expression level of multiple genes in a molecular functional pathway and a disease outcome. Classical multivariate analysis tools become infeasible because either the massively structured data fail the basic assumptions or how they cluster together between groups of interest in some research field. Powerful statistical tools like multivariate analysis of variance (MANOVA) are based on assumptions of independence, multivariate normal distribution and homogeneity of covariance [1, 2]. But many data sets do not conform with these assumptions. Take for example ecological data where number of each species is considered a variable. Abundance of individual species are usually highly aggregated and skewed and non-normally distributed. Also it is common that the number of spices is larger than the

sample size (i.e. small n and large p problem) [3, 4]. These major challenges lead to two avenue of statistical methodologies for the kind of multivariate data analysis.

1.2 Distance based linear correlation

1.2.1 Mantel test

Mantel test was primarily introduced to identify time–space clustering of disease. The methods was motivated by a biomedical research problem to identify clustering of leukemia patients in location and time [5]. Assume in a study a sample of n observations are recorded and can be represented as two subset of variables $X_{n\times p}$ and $Y_{n\times q}$. The interpersonal differences in X and Y, $D_X = [d_{ij}^X]_{1\le i<j\le n}$ and $D_Y = [d_{ij}^Y]_{1\le i<j\le n}$ are constructed by using suitable dissimilarity measurements s_X and s_Y as shown in Figure 1.1. Let $X_i = [x_{i1}, x_{i2}, \cdots, x_{in}]$ and $X_j = [x_{j1}, x_{j2}, \cdots, x_{jn}]$ be the i-th and j-th rows of X, then $d_{ij}^X = s_X(X_i, X_j)$ for some distance measure $s_X(\cdot,\cdot)$. Similarly, the pairwise distance matrix for Y is constructed. Because the distance

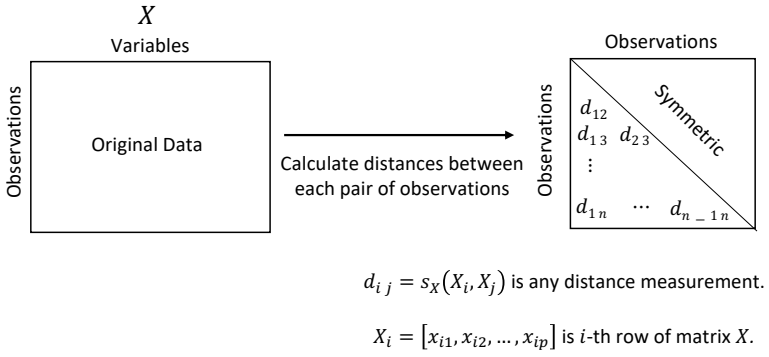


Figure 1.1: Constructing the lower triangular matrix of pairwise distances

matrices are symmetric, only element in the lower triangular matrix are used and the test statistic is

$$Z(D_X, D_Y) = \sum \sum_{(1\le i<j\le n)} d_{ij}^X d_{ij}^Y \tag{1.1}$$

A permutation procedure was proposed to obtain the empirical distribution of

Z values. However, the scale of Z will varies from problem to problem; hence, a normalized version of Mantel test statistic is

$$r(D_X, D_Y) = \sum \sum_{1 \le i < j \le n} \frac{(d_{ij}^X - \bar{d}^X)(d_{ij}^Y - \bar{d}^Y)}{var(d^X)^{1/2} var(d^Y)^{1/2}} \tag{1.2}$$

where $\bar{d}^X = \binom{n}{2}^{-1} \sum \sum_{1 \le i < j \le n} d_{ij}^X$ is the sample mean of lower triangular distance matrix ($\bar{d}^Y$ is similarly defined). The resulting statistic r is analogous to the Pearson correlation coefficient [6, 7]. A permutation based procedure is employed to assess the statistical significance by following steps: 1. compute the observed value of the statistic $r = r(D_X, D_Y)$ using Equation 1.2. 2. Permute rows and corresponding columns of the distance matrix simultaneously as seen in Figure 1.2 to preserve the symmetry of the structure and construct D_X^*. 3. Compute the Mantel statistic $r^* = r(D_X^*, D_Y)$ for all $n!$ possible permutations or a large random set of permutations for large data sets, say $B = 999$ for a precision level 0.001.

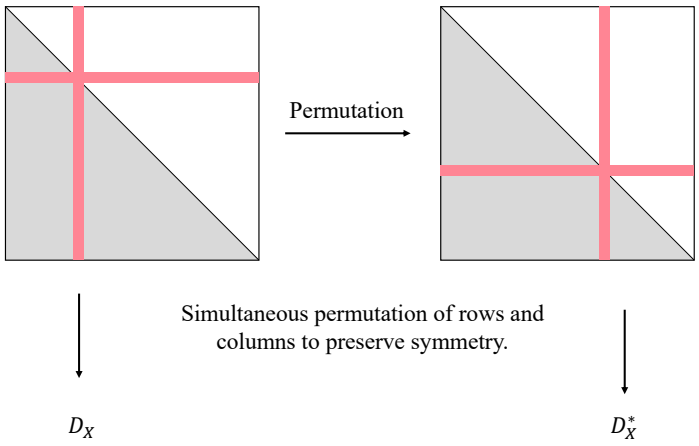


Figure 1.2: Permuting rows and columns jointly to preserve symmetric structure of distance matrix.

4. Calculate empirical p-value by comparing the observed r value with the r^*'s under permutations. For a test involving upper-tail, under the null hypothesis of $r \le 0$, the empirical p-value of the test is the proportion of the r^*'s greater than the observed Mantel statistic r, p-value$=\frac{\#(r^* > r)+1}{\#(r^*)+1}$. For a test involving lower-tail, under the null

hypothesis of $r \geq 0$, the empirical p-value of the test is the proportion of the r^*'s less than the observed Mantel statistic r, p-value$=\frac{\#(r^*<r)+1}{\#(r^*)+1}$. For a two-tailed test that is under the null hypothesis of $r = 0$, the empirical p-value of the test is the proportion of the $|r^*|$'s greater than $|r|$, p-value$=\frac{\#(|r^*|>|r|)+1}{\#(r^*)+1}$ where $|.|$ is the absolute value function.

1.2.2 Partial Mantel test

Partial Mantel test is an extension of Mantel test to three distance matrices that computes partial correlations between the two distance matrices while controlling for the effect of a third distance matrix [7,8]. Given the observed data X, Y and Z, the partial Mantel r statistic is

$$r(D_X, D_Y; D_Z) = \frac{r(D_X, D_Y) - r(D_X, D_Z)r(D_Y, D_Z)}{\sqrt{1 - r(D_X, D_Z)^2}\sqrt{1 - r(D_Y, D_Z)^2}} \tag{1.3}$$

where $r(D_X, D_Y)$ is the simple Mantel statistic calculated by Equation 1.2. Once $r(D_X, D_Y)$, $r(D_X, D_Z)$ and $r(D_Y, D_Z)$ are computed then $r = r(D_X, D_Y; D_Z)$ can be calculated using Equation 1.3. To test hypotheses concerning r, one can perform following steps. Construct D_X^* by random matrix permutation and evaluate $r(D_X^*, D_Y)$ and $r(D_X^*, D_Z)$. Then the partial correlation statistic under permutation, r^*, is the value of $r(D_X^*, D_Y; D_Z)$ computed by Equation 1.3. Repeat the procedure for a large number of times (or possibly for all $n!$ permutations for small data). The empirical p-value of the test is evaluated same way as simple Mantel test.

1.3 Multiple regression on distance matrices

Multiple regression on distance matrices (MRM) is an extension of partial Mantel analysis, for modeling between multiple pair-wise distance matrices [9]. Besides $Y_{n\times q}$ and $X_{n\times p}$, additional variables can be included, for example, $Z_{n\times k}$. Instead of modeling the linear relationship between the original data, MRM aims to model the

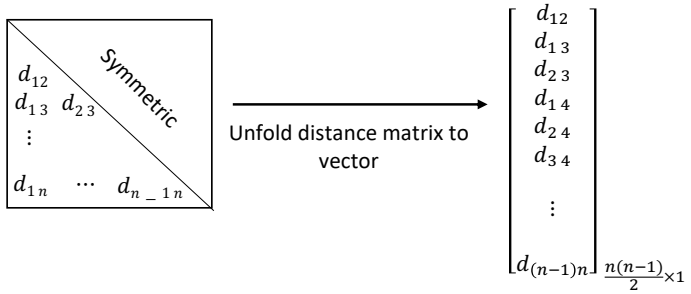


Figure 1.3: Unfolding a distance matrix to a vector

relationship between the pairwise distance matrices, D_Y, D_X and D_Z. Because all are symmetric matrices, the lower triangle of each matrix is vectorized (Figure 1.3), and a linear regression model is fitted.

$$d_{ij}^{Y} = \beta_0 + \beta_1 d_{ij}^{X} + \beta_2 d_{ij}^{Z} + \varepsilon_{ij}, \qquad 1 \le i < j \le n. \tag{1.4}$$

Unknown coefficients of the model are estimated by ordinary least squares. However, the asymptotic distribution of the t-test for the significance of the parameters for linear regression model on independent observations is not feasible because obviously, the pairwise distances are correlated. A permutation based procedure is also suggested to assess the significance of the test: The dependent distance matrix is permuted while holding the explanatory distance matrices unchanged. Let $\hat{\beta}_i^*$ be the estimated regression coefficients in a permutation. Under the null hypothesis $\beta_i = 0$, empirical p-value is the proportion of $\hat{\beta}_i^*$'s larger in absolute value than observed $\hat{\beta}_i$, p-value$=\frac{\#(|\hat{\beta}_i^*|>|\hat{\beta}_i|)+1}{\#(\hat{\beta}_i^*)+1}$, for all possible or a large sample of permutations.

1.4 Distance based multivariate analysis of variance

1.4.1 Background

Powerful multivariate statistical methods, such as the classical multivariate analysis of variance (MANOVA), have existed for decades [10–15]. However, in some

applications, the data structure fails the fundamental assumptions of these methods. For example, abundances of species in a community take discrete values, rare species contribute lots of zeros to the data set and there are more variables than sample size. Especially, when the dimension of the variable p is greater than sample size n, the sample variance covariance matrix becomes singular, so traditional tools, including Hottelling's T^2 [10] and Wilks' Lambda test [11] cannot be used. There are two ways to solve this issue, either using the generalized inverse [16] or distribution approximation [17] of Dempster trace criterion [18, 19] for one and two sample cases. Gower [20,21], Gower and Legendre [22], and Gower and Krzanowski [23] investigated the connection between sample variance and distance, specifically, for Euclidean distance, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i<j}^{n}(x_i - x_j)^2$. They proposed to extend all the necessary features of MANOVA to the dissimilarity matrix, for example, decompose the total variance to be between-group and within-group by using the dissimilarity matrix. The idea has been carried on and motivated various following works on pairwise distance matrix, among them, the PERMANOVA [4] gains the most attention due to its use in recent microbiome studies.

1.4.2 Permutational MANOVA

Permutational Multivariate Analysis of Variance (PERMANOVA) [4] is a non–parametric analogue to traditional MANOVA procedure obtained by partitioning the lower triangle of a distance matrix into between-group distances and within-group distances (Figure 1.4). Let $N = an$ be the total number of observations, where a is the number of groups, and n is the number of observations in each group. Let d_{ij} be the distance between sample units i and j. The total sum of squares is defined as

$$SS_T = \frac{1}{N}\sum \sum_{(1\leq i<j\leq N)} d_{ij}^2$$

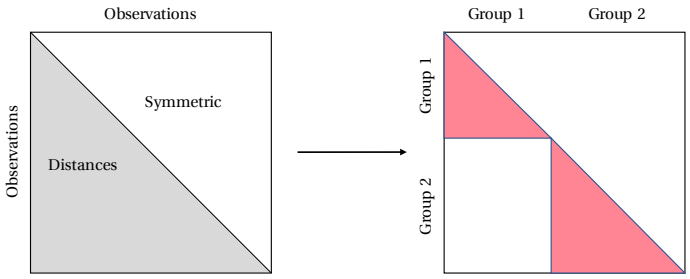


Figure 1.4: Partitioning a distance matrix into groups. In the right panel, red triangles represent within-group and the white rectangle represents the between-group.

and within-group or residual sum of squares is

$$SS_W = \frac{1}{n} \sum \sum_{(1 \leq i < j \leq N)} d_{ij}^2 I(\textit{observations } i \textit{ and } j \textit{ are in same group}),$$

where $I(\cdot)$ is the indicator function. The between-group sum of squares is defined as

$$SS_A = SS_T - SS_W.$$

A pseudo F-ratio test statistic is computed by

$$F = \frac{SS_A/(a-1)}{SS_W/(N-a)}$$

Anderson and Walsh [24] state PERMANOVA assumes exchangeabililty, that is joint conditional distribution $p(X_1, X_2, \cdots, X_n | Y_i = y_i, \textit{ for all } i)$ is invariant under permutation of the sample units among the groups [25]. The null hypothesis tested by PERMANOVA is H_0: centroids of groups as defined in space of chosen resemblance measure are equivalent for all groups. They later explain that, if H_0 were true, centroids of each group is within same distance to the overall centroid (Figure 1.5) [24,26].

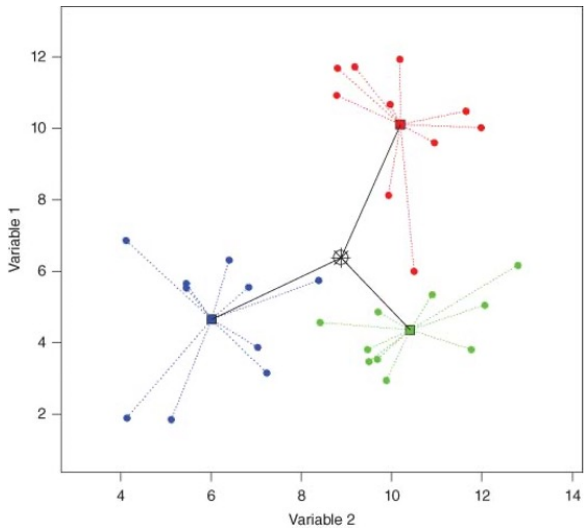


Figure 1.5: Schematic diagram of geometric partitioning for PERMANOVA, shown for g=3 groups of n=10 sampling units per group in two-dimensional (bivariate, p= 2) Euclidean space. First published: `https://doi.org/10.1002/9781118445112.stat07841`

Empirical p-value is obtained via permutation. For each permutation the test statistic F^* is calculated and the empirical p-value is computed using p-value$=\frac{\#(F^* \geq F)+1}{\#(F^*)+1}$.

1.5 Limitations of currently published methods

Variations of Mantel test are usually used in spatial correlation analysis, for example the study of geographical genetic divergence in botanical science. However, Mantel test suffers lack of power for spatially autocorrelated data [27]. Spatial autocorrelation is the term used to describe the presence of systematic spatial variation in a variable. Positive spatial autocorrelation is the tendency for areas or sites that are close together to have similar values [28]. Inflated Type-I error and relatively low power appears to be a general feature of the Mantel test [29–31].

PERMANOVA is routinely used in numerical ecology to test for the location differences in microbial communities. It has been discussed that PERMANOVA fails to detect a significant between-group difference unless it is present in taxa (units of any rank i.e. kingdom, phylum, class, order, family, genus and species, designating an

organism or a group of organisms) with high variance [32]. Moreover, PERMANOVA is prone to inflated Type-I error in presence of heteroscedasticity [33, 34].

Franckowiak *et al.* in [35] examined the ability of model selection criteria based on Akaike's information criterion (AIC) or its small-sample correction (AICc) and the Bayesian information criterion (BIC), to reliably rank candidate models when applied with MRM while varying the sample size; and, strongly discouraged the continued application of AIC, AICc and BIC for model selection with MRM.

A possible explanation of challenges summarized above could be misusing naive permutation testing. The concept of permutation relies on exchangeability and simple permutation could lead to inflated Type-I error or low power [36, 37], particularly, in presence of nuisance effect, unequal variances, correlations, skewness, or unequal sample sizes in two groups [38, 39].

CHAPTER 2: HUMAN MICROBIOME AND INFLAMMATORY BOWEL DISEASES

2.1 Introduction

"What is the human microbiome?" has troubled researchers [40] since Lederberg's coinage of "microbiome" in 2001 [41]. Researchers have confused how to exactly define the human microbiome and interchangeably used terminologies "microbiota" and "microbiome". The term microbiota is referred to the microbial taxa associated with humans to signify the communities of microorganisms within a specific environment [42]. The term microbiome is defined as the collection of the microbial taxa or microbes and their genes [40], [43], that is the entire microbial communities. Depending on the collection of microbes, i.e. the body site in which those microorganisms inhabit in, researchers use specific terms such as gut microbiome, skin microbiome or oral microbiome.

Most sources cite the number of human cells as 10^{13} or 10^{14}, and a recent study reported 3.7×10^{13} human cells in a reference human [44]. Estimates for the number of microbial cells in the body are usually $10^{14} - 10^{15}$ [45], [46] which suggests a ratio of $10 : 1$ microbial cells to human cells. However, more recent studies suggest the ratio to be much closer to $1 : 1$ [47]. Growing evidence suggests that gut microbiota may be an important factor in the pathogenesis of a variety of diseases including

inflammatory [48].

Inflammatory Bowel Disease (IBD) is a broad term that describes conditions characterized by chronic inflammation of the gastrointestinal tract. The two most common inflammatory bowel diseases are ulcerative colitis (UC) and Crohn's disease (CD). Inflammation affects the entire digestive tract in Crohn's disease and only the large intestine (also called the colon) in ulcerative colitis. In 2015, more than 3.5 million people worldwide were diagnosed with IBD (either Crohn's disease or ulcerative colitis) [49]. Dysbiosis of the gut microbiota, an alteration of the microbial community structure associated with disease, has been consistently observed in patients with IBD. Although the dysbiosis may simply be a result of the inflammatory process [33], it may play a role in the pathogenesis of disease where there is an increase in potentially harmful bacterial and a reduction in more protective bacterial species [50].

2.2 The human microbiome project

The National Institutes of Health (NIH) Human Microbiome Project (HMP, `https://hmpdacc.org`) was carried out over ten years and two phases to provide resources, methods, and discoveries that link interactions between humans and their microbiomes to health–related outcomes. The second phase of the HMP, the Integrative HMP (iHMP or HMP2) [51], was designed to explore host-microbiome interplay, including immunity, metabolism, and dynamic molecular activity, to gain a more holistic view of host-microbe interactions over time [52]. The iHMP projects included three studies that followed the dynamics of human health and disease during conditions with known microbiome interactions: (1) Pregnancy and preterm birth (PTB) [53], (2) Inflammatory bowel diseases (IBD) [54] and (3) Stressors that affect individuals with prediabetes [55]. A collection of commentary and research publications from across Nature journals and related publications from HMP2 can be found at `https://www.nature.com/collections/fiabfcjbfj`, and a rich multi-omic data resource at `http://www.ihmpdcc.org`.

2.3 HMP2 Study: gut microbiome and inflammatory bowel diseases

The Inflammatory Bowel Disease Multi'omics Database (IBDMDB) project followed 130 individuals from five clinical centres over the course of one year each as part of HMP2 (see Table 2.1). Integrated longitudinal molecular profiles of microbial and host activity were generated by analyzing 1,785 stool samples (self-collected and sent by mail every two weeks), 651 intestinal biopsies (collected colonoscopically at baseline), and 529 quarterly blood samples [52]. Multiple molecular profiles were generated from the same sets of samples, including stool metagenomes, metatranscriptomes, metaproteomes, viromes, metabolomes, host exomes, epigenomes, transcriptomes, and serological profiles, among others, allowing concurrent changes to be observed in multiple types of host and microbial molecular and clinical activity over time. Protocols and results from the study, further information about its infrastructure, and both raw and processed data products are available through the IBDMDB data portal `http://ibdmdb.org`, from the HMP2 Data Coordination Center (DCC; `http://ihmpdcc.org`). [54, 56–60]

Table 2.1: Number of Subjects by Category in HMP2 study of inflammatory bowel diseases (IBD). nonIBD: not diagnosed with IBD or control group; CD: diagnosed with Crohn's disease; UC: diagnosed with ulcerative colitis.

Disease	Gender	Child	School-age	Adult	Senior	Total
nonIBD	female	3	5	4	0	12
	male	3	2	9	1	15
CD	female	3	8	20	1	32
	male	5	15	13	0	33
UC	female	2	3	14	1	21
	male	1	9	7	1	17
Total		17	42	67	4	130

2.4 Metagenomes: profiling and statistical analyses

National Human Genome Research Institute (NHGRI, `https://www.genome.gov/genetics-glossary/Metagenomics`) defines metagenomics as the study of a collection of genetic material (genomes) from a mixed community of organisms. Metagenomics usually refers to the study of microbial communities. The genome is the entire set of genetic instructions found in a cell. Metagenomics is usually used in the study of microbial communities where one can't separate one microbe from another.

As a part of HMP2, the composition of microbial communities of stool samples were profiled from metagenomic shotgun sequencing data and MetaPhlAn (Metagenomic Phylogenetic Analysis) [61]. Processing microbiome data generates a matrix that relates feature abundance (taxa or genes) to samples. The microbiome data are highly dimensional, often representing thousands of different taxa, and sparse and zeros inflated matrix [62]. Stool samples were collected over the course of one year and metagenomic profiles were generated and classified at seven taxonomic ranks—that is the relative level of a group of organisms (a taxon) in a taxonomic hierarchy, species, genus, family, order, class, phylum and kingdom. At kingdom level, Bacteria make up for over 99% of detected microorganisms. The bacteria detected in adult participants was classified as 12 phyla which includes 581 species (see Table 2.2).

Table 2.2: Kingdom: Bacteria, Age-group: Adults

Rank	count
Phylum	12
Class	25
Order	41
Family	76
Genus	187
Species	581

Attributes such as species richness, evenness and diversity can be used to compare community compositions. Species richness is the number of different species community. Species evenness is a description of the distribution of abundance across the

species in a community. Species diversity is usually described by an index that includes both richness and evenness of the species. Global taxonomic richness is affected by variation in three components: within-community, or alpha diversity, between-community, or beta diversity, and between-region, or gamma diversity [63–70].

2.5 Alpha diversity and evenness

Shannon's diversity index is commonly used in ecology as a measure of alpha diversity. It's based on the Shannon's entropy formula, $H_{Shannon} = -\sum_{i=1}^{S} p_i \ln p_i$ where $p_i = \frac{n}{N}$ is the proportion of the number of individual species i found (n) divided by the total number of individuals found (N) and S is the number of different species. Pielou's evenness index [71] is defined by $E_{Pielou} = \frac{H_{Shannon}}{\ln S}$.

Figure 2.1 shows the distribution of Shannon indices for each cohort. Kruskal-Wallis [72] rank sum test of differences in mean values of Shannon indices as shown on the plot ($\chi^2_2 = 41.08$, p-value ≈ 0) as well as Pielou indices ($\chi^2_2 = 9.72$, p-value=0.0078).

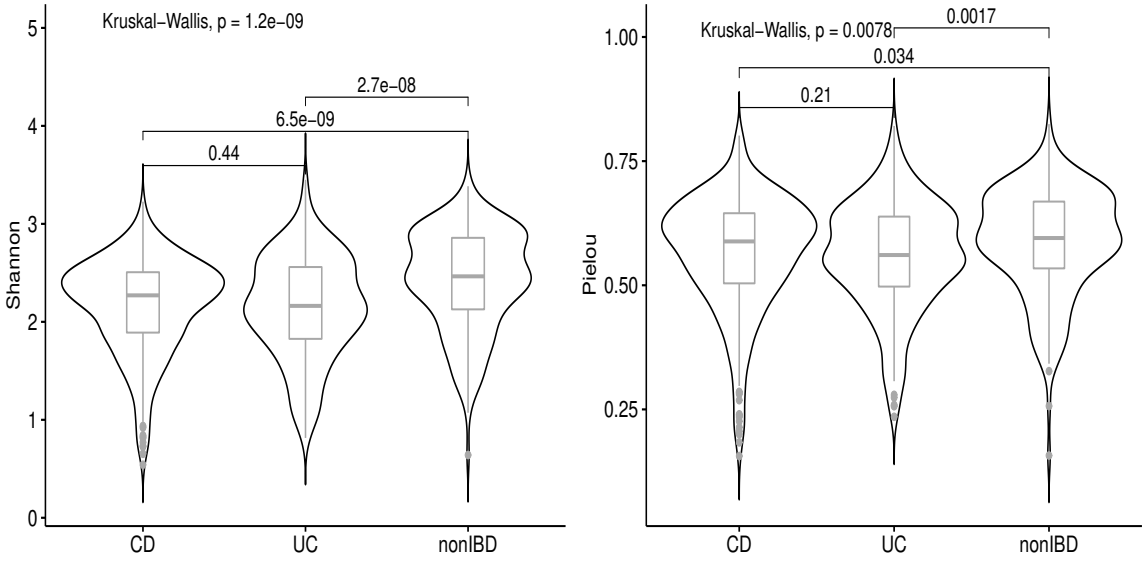


Figure 2.1: Distribution of Shannon's diversity index and Pielou's evenness index

2.6 Beta diversity

Common indices of beta diversity include Bray-Curtis, Unifrac distance. Bray-Curtis distance [73], is a commonly used distance measure (beta diversity) in microbiome data, $BC_{ij} = 1 - \frac{2C_{ij}}{S_i+S_j}$, where S_i is the total number of species counted in sample i, S_j is the total number of species counted in sample j, and, C_{ij} is the sum of only the lesser counts for each species found in both samples.

UniFrac and Weighted UniFrac distances [74–77] are other measures of beta diversity that compute differences between microbial communities based on phylogenetic information. UniFrac measures the phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both. Weighted UniFrac accounts for abundance of observed organisms whereas unweighted UniFrac only considers their presence or absence. A schematic diagram of UniFrac calculations is shown in Figure 2.2

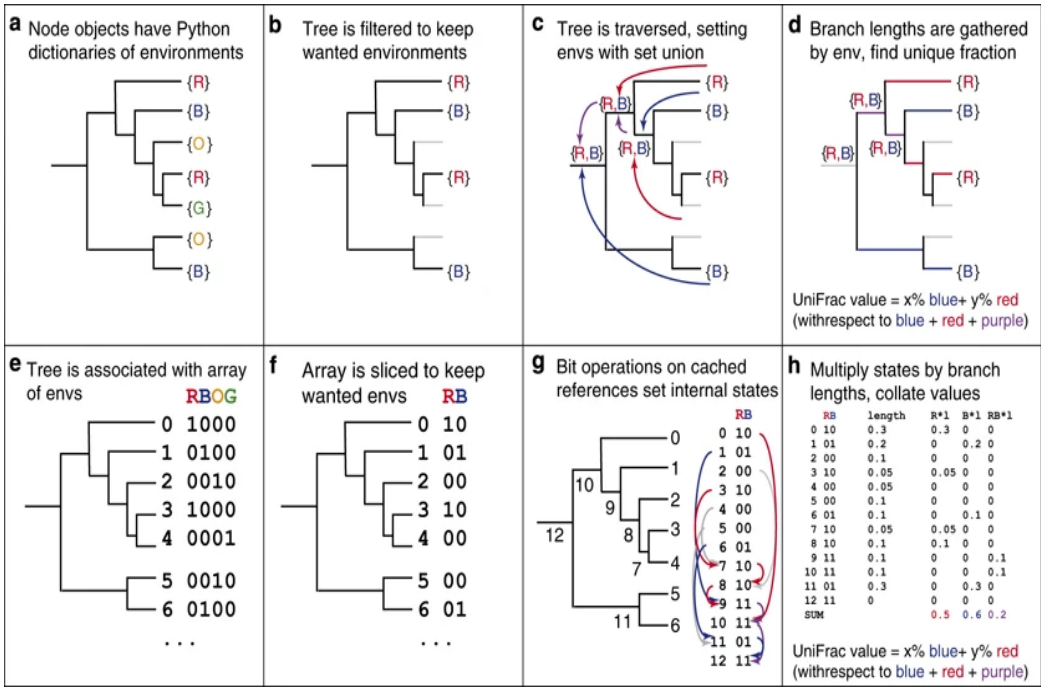
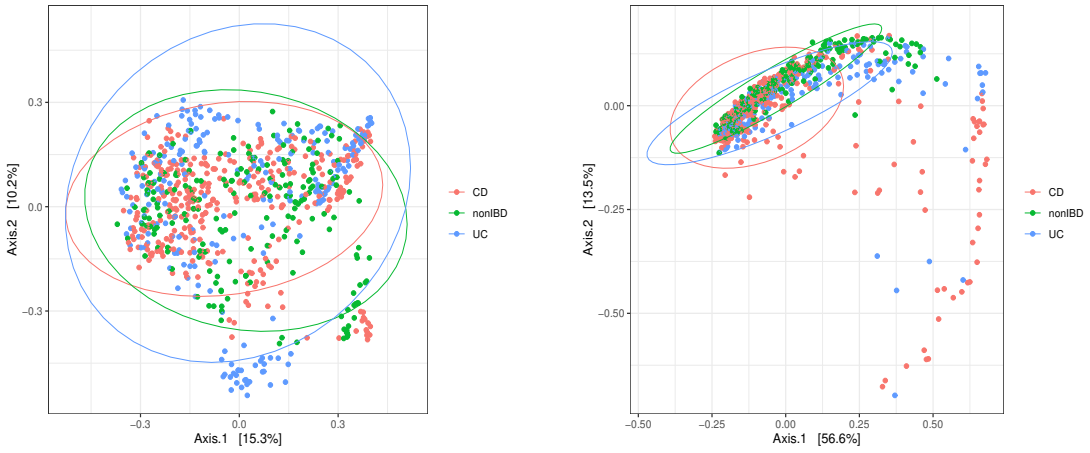


Figure 2.2: Schematic diagram of UniFrac calculations. First published: `https://doi.org/10.1038/ismej.2009.97`

2.6.1 Visualization of beta diversity

A well-known tool to visualize beta diversity is principal coordinates analysis. Principal coordinates analysis (PCoA) is a metric multidimensional scaling method based on projection, which uses spectral decomposition to approximate a matrix of distances/dissimilarities by the distances between a set of points in few dimensions [78]. PCoA is equivalent to principal component analysis (PCA) when euclidean distances are used. Plots of first two axes of PCoA accompanied by 95% student's-t confidence ellipses were generated and displayed in Figure 2.3a using Bray-Curtis distance measure and in Figure 2.3b using Weighted UniFrac distance method at species level.



(a) Bray-Curtis (b) WUniFrac

Figure 2.3: Plots of first two axes of PCoA accompanied by 95% student's-t confidence ellipses; Samples are from adult participants at kingdom level bacteria.

Weighted Unifrac distance method seems to separate CD and UC cases from healthy controls better than Bray-Curtis method. Using Weighted Unifrac 70% of variation is explained by first two axes. In contrast, using Bray-Curtis distance only 25.5% of variation is explained by first to axes and all samples overlapped.

2.6.2 Analysis of beta diversity using PERMANOVA

A PERMANOVA was performed using three groups: CD and UC cases and nonIBD controls using Bray-Curtis distance (see Table 2.3) and Weighted UniFrac distance (see Table 2.4).

Table 2.3: PERMANOVA output using Bray-Curtis based on 9999 permutations; terms added sequentially.

	Df	Sums of Sqs	Mean Sqs	F.Model	Pr(>F)
CD	1	1.981	1.9808	6.5911	0.0001
UC	1	3.822	3.8223	12.7185	0.0001
Residuals	779	234.114	0.3005		
Total	781	239.918			

Table 2.4: PERMANOVA output using WUnifrac based on 9999 permutations; terms added sequentially.

	Df	Sums of Sqs	Mean Sqs	F.Model	Pr(>F)
CD	1	0.742	0.74246	8.3305	0.0001
UC	1	0.580	0.57972	6.5045	0.002
Residuals	779	69.429	0.08913		
Total	781	70.751			

2.6.3 Analysis of beta diversity using Mantel test

We first tried to identify core and rare taxa. Table 2.5 shows the prevalence (number of samples representing the taxon) of the phyla of gut bacteria. Firmicutes and bacteroidetes are present in almost all samples whereas spirochaetes, chlamydiae or cyanobacteria appear in less than 1% of samples. Figure 2.4 shows the relative abundance of bacteria at phylum level for adults in three groups, diagnosed with Crohn's disease (CD), ulcerative colitis (UC) or healthy controls (nonIBD). On average, bacteroidetes and firmicutes make up for 95% of phylum level of gut bacteria.

Table 2.5: Abundances and relative abundances of the phyla of gut bacteria

	Abundance	Relative abundance
Firmicutes	1602	1.00
Bacteroidetes	1572	0.98
Proteobacteria	1499	0.93
Actinobacteria	1440	0.90
Verrucomicrobia	502	0.31
Fusobacteria	76	0.05
Synergistetes	45	0.03
Lentisphaerae	17	0.01
Candidatus Melainabacteria	12	0.01
Spirochaetes	5	0.00
Chlamydiae	2	0.00
Cyanobacteria	1	0.00

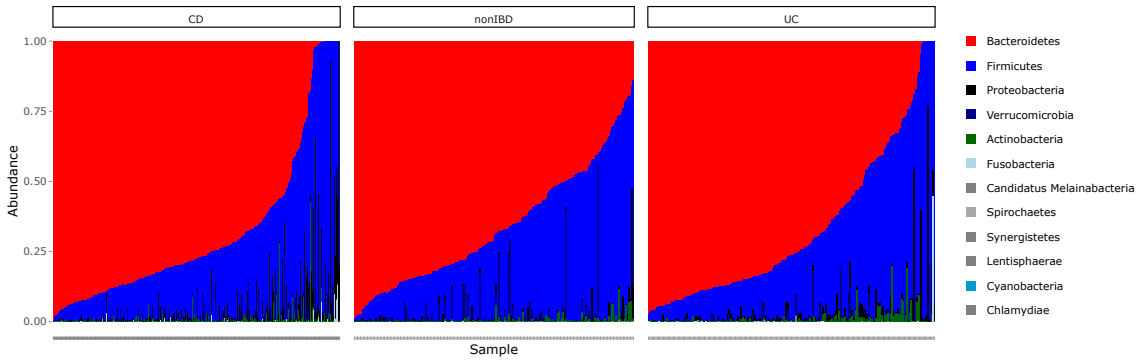


Figure 2.4: Abundance of phyla of gut bacteria for adults with CD ($n = 355$), nonIBD ($n = 196$) and UC ($n = 231$)

Intuitively, box plots of relative abundances and distance measurements are presented in Figure 2.5 and Figure 2.6. Since there were over 500 distinct species with high sparsity, box plots were generated at phylum level for a better resolution. Figure 2.5 displays the relative abundance of phyla grouped by UC and CD cases and nonIBD controls within adults participated in HMP2 study. Bacteroidetes and firmicutes, proteobacteria, actinobacteria and verrucomicrobia have the highest relative abundances (core taxa) and therefore of interest to be investigated. Figure 2.6 displays the distances/dissimilarities of each phylum labeled by between and within cohorts. Bacteroidetes and firmicutes distance measurements present higher interquar-

tile range while proteobacteria, actinobacteria and verrucomicrobia distances are heavily skewed.

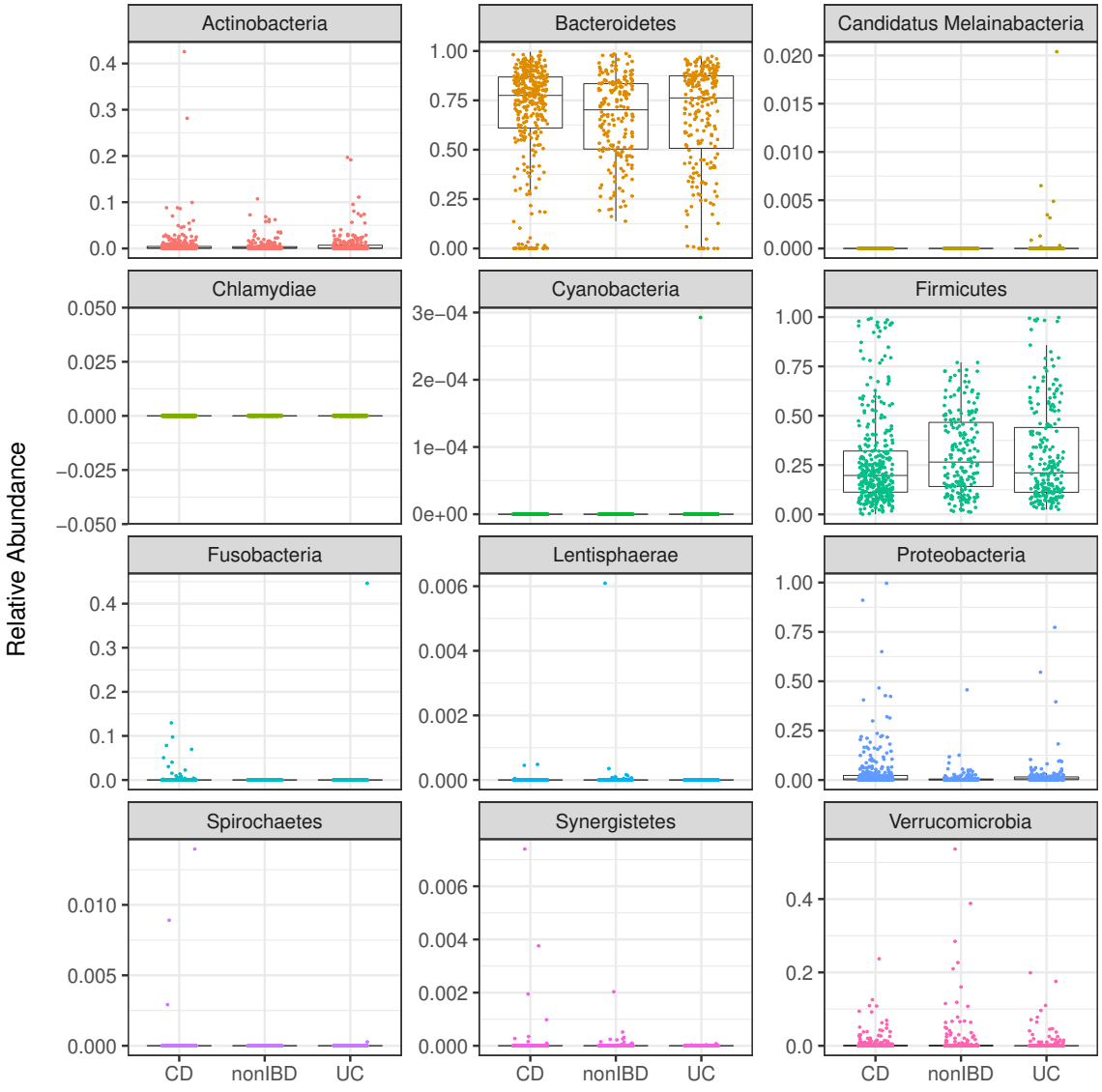


Figure 2.5: Box plots of relative abundance of gut microbiome of adults at phylum level for UC and CD cases and nonIBD controls.

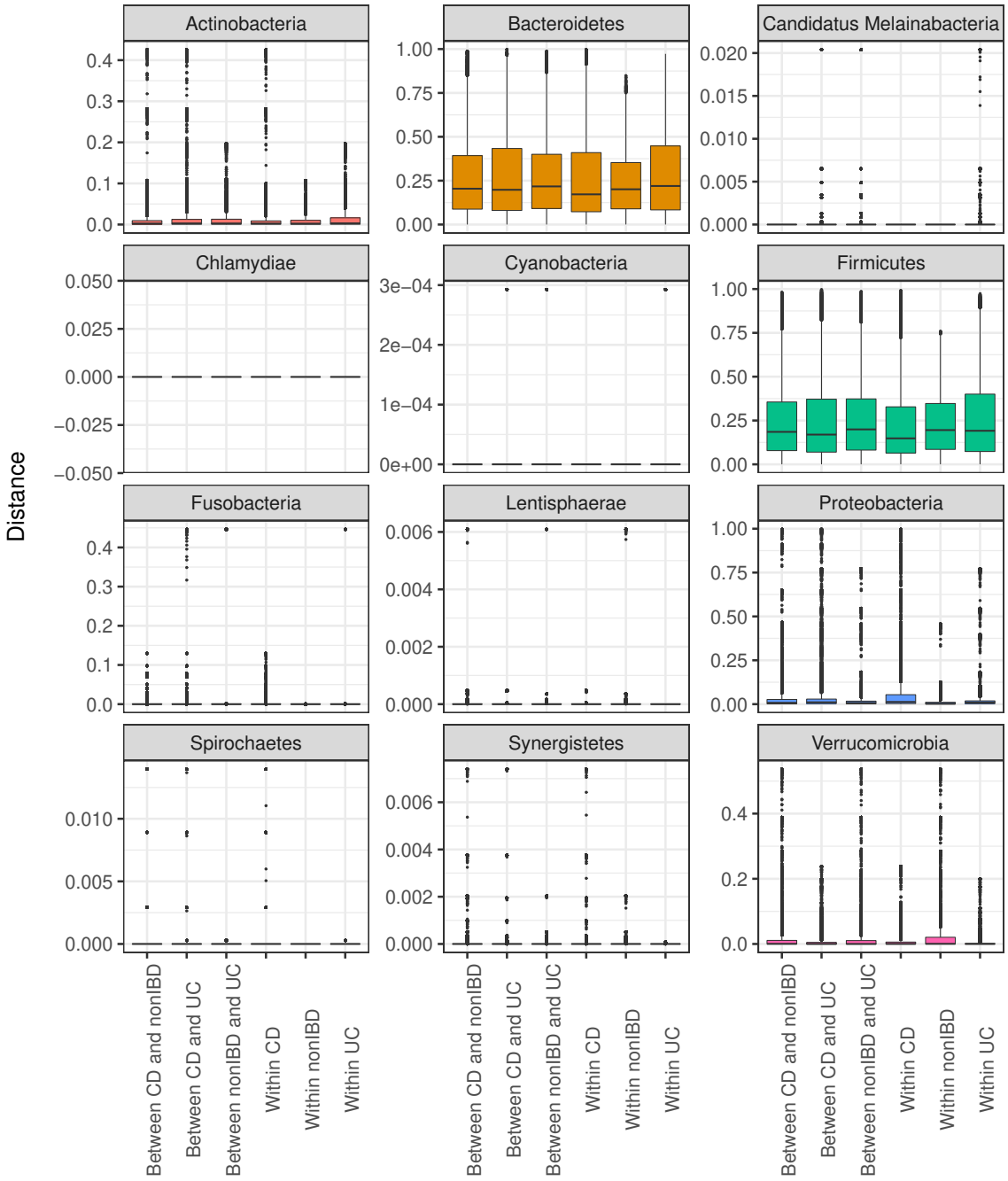


Figure 2.6: Box plots lay out the distribution of pairwise distances of gut microbiome of adults at phylum level, labeled by between and within cohorts.

Mantel test (Table 2.6) was used to investigate whether lower dissimilarities of species with phylum level bacteroidetes and firmicutes correspond to lower (or higher) dissimilarities of species with phylum level proteobacteria, actinobacteria and verru-

comicrobia for each group. The null hypothesis for "pval1" is that the Mantel r statistic will be equal to or smaller than zero, i.e. negative correlation. Conversely, the null hypothesis for "pval2" is that the Mantel r statistic is equal to or greater than zero, i.e. positive correlation. The null hypothesis for "pval3" is that the Mantel r statistic is equal to zero, i.e. no correlation.

Table 2.6: Mantel r-statistic and p-values: "pval1" under $H_0 : r \leq 0$; "pval2" under $H_a : r \geq 0$; "pval3" when $H_0 : r = 0$; accompanied by lower and upper limits of 95% CI

	Mantel r	pval1	pval2	pval3	llim.2.5%	ulim.97.5%
within CD	0.0423	0.0086	0.9915	0.0106	0.0282	0.0572
within UC	0.0688	0.0008	0.9993	0.0008	0.0479	0.0863
within nonIBD	0.0991	0.0001	1.0000	0.0001	0.0851	0.1128

We learn that dissimilarities of species with phylum level bacteroidetes and firmicutes are positively correlated to dissimilarities of species with phylum level proteobacteria, actinobacteria and verrucomicrobia for each cohort. It is of interest to study the biological relationships and interactions of species grouped as above. Smaller Mantel r for UC and CD cases might be a biological signal that needs further investigations.

2.6.4 Analysis of beta diversity using MRM

For application of MRM on this data set, distances at species level were computed and labeled as withing and between UC, CD and nonIBD groups. Considering the

within nonIBD distances as reference group, the MRM model is

$$\begin{aligned} d_{ij} = \beta_0 &+ \beta_1 I_{\text{Within CD}} \\ &+ \beta_2 I_{\text{Between CD and UC}} \\ &+ \beta_3 I_{\text{Between CD and nonIBD}} \\ &+ \beta_4 I_{\text{ Within UC}} \\ &+ \beta_5 I_{\text{Between nonIBD and UC}} + \varepsilon_{ij} \end{aligned} \tag{2.1}$$

where $I(\cdot)$ is the indicator function. MRM coefficients and p-values based on 9999 permutations are summarized for choice of distance Bray-Curtis in Table 2.7a and for Weighted UniFrac in Table 2.7b.

Table 2.7: MRM coefficients and p-values using 9999 permutations

	Distance	pval
Intercept	0.7061	1.0000
Within CD	0.0622	0.0001
CD–UC	0.0742	0.0001
CD–nonIBD	0.0552	0.0001
Within UC	0.0699	0.0002
nonIBD–UC	0.0632	0.0001

(a) Bray-Curtis

	Distance	pval
Intercept	0.3174	0.9972
Within CD	0.0525	0.0295
CD–UC	0.0571	0.0118
CD–nonIBD	0.0390	0.0017
Within UC	0.0531	0.0460
nonIBD–UC	0.0367	0.0076

(b) WUnifrac

Differences on the p-values obtained on different distances indicate that MRM may suffer inflated Type-I error rate when Bray-Curtis distance is used.

CHAPTER 3: A DISTANCE-BASED LINEAR REGRESSION MODEL

3.1 Introduction

3.1.1 Statistical model and parameter estimation

Suppose that we have n independent data draws denoted as $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \cdots, n$, where $\mathbf{x}_i \in \mathcal{R}^p$ and $\mathbf{y}_i \in \mathcal{R}^q$. Pairwise distances or dissimilarities between all combinations of n objects of $\{\mathbf{y}_i\}_{i=1}^n$ are constructed and denoted as $D_{\mathbf{Y}} = [y_{ij}]_{1 \le i < j \le n}$. Let $K \le p$ be the number of sub-groups of explanatory variables, that is $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \cdots, \mathbf{x}_i^{(K)})'$, then for each sub-group distance matrices are constructed and denoted as $D_{\mathbf{x}}^{(1)} = [x_{ij}^{(1)}]_{1 \le i < j \le n}$, $\ldots$, $D_{\mathbf{x}}^{(K)} = [x_{ij}^{(K)}]_{1 \le i < j \le n}$, respectively. Depending on the application, the pairwise distance measurements may reflect distance in species abundances, geographical location, and genetic distance using compound or individual distance measure. The pairwise distance transforms the original independent observations $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \cdots, n$, to correlated pairwise distances, which are denoted as $y_{ij} = s_y(\mathbf{y}_i, \mathbf{y}_j)$ and $x_{ij}^{(K)} = s_k(\mathbf{x}_i^{(k)}, \mathbf{x}_j^{(k)})$, $k = 1, 2, \cdots, K$ and $1 \le i < j \le n$. We then model the relationship between y_{ij} and $x_{k,ij}$'s via the following regression model:

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + \varepsilon_{ij}, \quad 1 \le i < j \le n, \tag{3.1}$$

here, $\mathbf{x}_{ij} = [1, x_{ij}^{(1)}, \cdots, x_{ij}^{(K)}]'$ and the vector of coefficients $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_K]'$.

A least squares estimator of the regression coefficients $\boldsymbol{\beta}$ can be derived as in classic linear regression model for independent observations, by minimizing the following sum of squares:

$$U_n(\boldsymbol{\beta}) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} (y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta})^2 = \frac{1}{\binom{n}{2}} (Y - X\boldsymbol{\beta})'(Y - X\boldsymbol{\beta}), \tag{3.2}$$

where

$$Y = [y_{12}, y_{13}, \cdots, y_{1n}, y_{23}, \cdots, y_{(n-1)n}]'$$

and

$$X = [\mathbf{x}_{12}, \mathbf{x}_{13}, \cdots, \mathbf{x}_{1n}, \mathbf{x}_{23}, \cdots, \mathbf{x}_{(n-1)n}]'.$$

The ordinary least squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$. This is the multivariate regression model (MRM) of pairwise distances [9] that we reviewed in Chapter 1. The model was motivated by ecological studies and is more flexible in terms of incorporating multiple distance matrices.

Although the parameter estimation is straightforward, the large sample properties of the estimators are different from the classical linear regression model because y_{ij}'s are clearly correlated. Lichstein [9] applied a permutation based procedure to test the significance of every individual model coefficient. However, a well known pitfall of permutation method is that it is computationally expensive because the model needs to be re-fitted for a large number of times. Especially in large scale studies, such as microbiome genome-wide association studies, the association between microbiome community and genetic markers may require tens of thousands of tests. Moreover, multiple testing corrections are always required, which means the result of a test can be called significant if empirical p-value is smaller than 10^{-6}. In that case, at least 10^6 permutations are needed for each test . Therefore, we investigated the large sample properties of model coefficients including asymptotic consistency and normality of

the least squares estimator of MRM in the following section. Based on the derived theoretical results, a computationally much efficient inference procedure is developed.

3.1.2 Large sample theory

$U_n(\boldsymbol{\beta}) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} (y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta})^2$ forms a second-order U-statistic. Large sample properties of U-statistics were used to drive the theoretical distribution of $\hat{\boldsymbol{\beta}}$. [79–82]

Theorem 1 (Asymptotic consistency)**.** *Assuming Θ is compact, if $E(\varepsilon_{ij}^2) < \infty$, then the least squares estimator defined by minimizing $U_n(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}$, is consistent for $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} E(U_n(\boldsymbol{\beta}))$.*

Proof. Since $U_n(\boldsymbol{\beta})$ is a second order U-statistic, by the strong law of large numbers for U-statistic, $U_n(\boldsymbol{\beta}) \to E(U_n(\boldsymbol{\beta}))$ almost surely. Moreover, $\boldsymbol{\beta}_0$ is the unique minimizer of $U(\boldsymbol{\beta}) = E(U_n(\boldsymbol{\beta}))$ [79, 80, 83]. Then the consistency can be derived by following the argument for consistency of M-estimators. □

Theorem 2 (Asymptotic Normality)**.** *If $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$ and $E(h^2) < \infty$, then $\hat{\boldsymbol{\beta}}$ is asymptotically normal,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, \Sigma), \tag{3.3}$$

where

$$\Sigma = 4H_0^{-1}V_0H_0^{-1}$$

with

$$H_0^{-1} = E(\mathbf{x}_{ij}\mathbf{x}'_{ij})$$

and

$$V_0 = Var(E[\mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}_0)|\mathbf{x}_i, \mathbf{y}_i]).$$

Proof. Define a normalized score function:

$$Q_n(\boldsymbol{\beta}) = \sqrt{n}\frac{\partial U_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\sqrt{n}}{\binom{n}{2}} \sum_{1\le i<j\le n} \mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}) = -\frac{\sqrt{n}}{\binom{n}{2}} \sum_{1\le i<j\le n} h(\mathbf{x}_i, y_i, \mathbf{x}_j, y_j; \boldsymbol{\beta}) \tag{3.4}$$

Since $\hat{\boldsymbol{\beta}}$ is a consistent estimator, Taylor series expansion of $Q_n(\hat{\boldsymbol{\beta}})$ at $\boldsymbol{\beta}_0$:

$$Q_n(\hat{\boldsymbol{\beta}}) = Q_n(\boldsymbol{\beta}_0) + \frac{\partial Q_n(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1) \tag{3.5}$$

therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -H_n^{-1} Q_n(\boldsymbol{\beta}_0) + o_p(1) \tag{3.6}$$

where $H_n = \frac{1}{\binom{n}{2}} X'X$.
According to the theories of second order U-statistics [83],

$$Q_n(\boldsymbol{\beta}_0) = \hat{Q}_n(\boldsymbol{\beta}_0) + o_p(1) \tag{3.7}$$

where

$$\begin{aligned}
\hat{Q}_n(\boldsymbol{\beta}_0) &= \sum_{i=1}^{n} E(Q_n(\boldsymbol{\beta}_0)|y_i, \mathbf{x}_i) \\
&= -\sum_{i=1}^{n} \sqrt{n}\binom{n}{2}^{-1} \sum_{1\le i<j\le n} E[\mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}_0)|y_i, \mathbf{x}_i] \\
&= -\sqrt{n}\binom{n}{2}^{-1} \sum_{i=1}^{n} \left\{ \binom{n-1}{1} E[\mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}_0)|y_i, \mathbf{x}_i] \right\} \\
&= -\frac{2}{\sqrt{n}} \sum_{i=1}^{n} r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0).
\end{aligned} \tag{3.8}$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -H_n^{-1} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) + o_p(1). \tag{3.9}$$

Since $r(y_i, \mathbf{x}_i)$ are i.i.d, by the central limit theorem,

$$\frac{2}{\sqrt{n}} \sum_{i=1}^{n} r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) \to N(\mathbf{0}, 4V_0) \tag{3.10}$$

where $V_0 = var(r(y_i, \mathbf{x}_i))$.

By the law of large number, $H_n \to E(\mathbf{x}_{ij}\mathbf{x}'_{ij}) \equiv H_0$. The Slutsky theorem implies that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(\mathbf{0}, 4H_0^{-1}V_0H_0^{-1}). \tag{3.11}$$

□

3.1.3 Estimating the covariance matrix and hypothesis testing

We estimate the conditional expectation $r(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0)$ by

$$\hat{r}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) \equiv \binom{n-1}{1}^{-1} \sum_{j \neq i} \mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}) \tag{3.12}$$

then,

$$\frac{1}{n} \sum_{i=1}^{n} \hat{r}(y_i, \mathbf{x}_i, \boldsymbol{\beta}_0) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbf{x}_{ij}(y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}) = -Q_n(\hat{\boldsymbol{\beta}})/\sqrt{n} = 0. \tag{3.13}$$

Therefor

$$\hat{V}_n \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{r}(y_i, x_i, \hat{\boldsymbol{\beta}})\hat{r}(y_i, x_i, \hat{\boldsymbol{\beta}})' \tag{3.14}$$

is an unbiased consistent estimator of V_0 [83].

3.1.4 Hypothesis testing

Hypothesis testing is an essential component of statistical inference because it is often of practical interest to test if a certain covariate is significantly associated with the response variable. This testing problem can be accommodated by considering the form: $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$, if the coefficient of the k-th covariate in the regression model is zero. Since we have derived the asymptotic normality of the least squares estimator of $\boldsymbol{\beta}$, we may use a Wald type test statistic $T_k = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$. The p-value of the test can be calculated by using the normal approximation based on the asymptotic normality of each $\hat{\beta}_k$.

3.2 Simulation studies

In this section, we aim to assess the accuracy of estimation as well as investigate the finite sample performance of our proposed testing procedure by conducting extensive simulation studies. We considered six different scenarios to evaluate the performance of our method from multiple perspectives. Simulations' set-up and results of each scenario are detailed in the following sub-sections.

3.2.1 Simulation for assessing the accuracy of estimation

Scenario I: In this simulation, we mimic genetic association studies and generate genotype data of 20 single nucleotide polymorphisms (SNPs) as detailed in Figure 3.1. Each genotype $\mathbf{g}_i$ is a sequence of 20 elements $g_{ik} \in \{0, 1, 2\}$ where $i = 1, \cdots n$, and $k = 1, \cdots, 20$, with the multinomial distribution,

$$\mathbf{g}_i \sim multinomial(p = (maf_i^2, 2maf_i(1 - maf_i), (1 - maf_i)^2)).$$

Here, maf_i is short for minor allele frequency of the i-th, which is the frequency at which the rare allele occurs in diploid human genome. Three different values of mafs are considered ($maf = 0.1, 0.3, 0.5$) in this simulation. We calculate the genomic distance between individual i and j by $x_{ij} = \sum_{k=1}^{20} |g_{ik} - g_{jk}|$, and simulate observations of the response variable via the following model: $y_{ij} = 0.2x_{ij} + \varepsilon_{ij}$, where $\varepsilon_{ij} = \varepsilon_i - \varepsilon_j$ for each $\varepsilon_i \sim N(0, \sigma^2)$.

	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
(0) AA Pr=maf^2										
(1) Aa Pr=$2 \cdot maf(1 - mfa)$										
(2) aa Pr=$(1 - maf)^2$										
g	2	2	2	2	2	1	2	2	2	0

Figure 3.1: An intuitive illustration of a gene with 10 SNPs constructed using alleles AA, Aa and aa.

Based on 1000 replications, we summarize the Mean Estimates that is the means of coefficients estimates, the Mean Standard Error (MSE), which is the mean of asymptotic standard errors of parameter estimations, the Empirical Standard Deviation (ESD), which is the sample standard deviation of the estimates and the confidence interval coverage probability (CI CP), which is the proportion of confidence intervals covering true parameter. These summary statistics of the simulation are represented in Table 3.1. When sample size increases, the coefficients estimates approach the true parameters. Means of asymptotic standard errors are almost always equivalent to sample standard deviations of estimates and 95% Confidence interval coverage proportions are about 0.95.

Table 3.1: A summary of simulations based on Scenario I under the assumption that pairwise distance matrices of response and independent variables have a linear relationship.

σ	maf	n	Mean Estimates		MSE		ESD		95% CI CP	
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
0.05	0.1	50	-0.00	0.20	0.02	0.00	0.02	0.00	0.95	0.95
		100	0.00	0.20	0.02	0.00	0.02	0.00	0.95	0.94
		200	0.00	0.20	0.01	0.00	0.01	0.00	0.94	0.93
	0.3	50	-0.00	0.20	0.02	0.00	0.02	0.00	0.96	0.95
		100	-0.00	0.20	0.02	0.00	0.02	0.00	0.95	0.95
		200	-0.00	0.20	0.01	0.00	0.01	0.00	0.95	0.95
	0.5	50	-0.00	0.20	0.03	0.00	0.03	0.00	0.96	0.95
		100	-0.00	0.20	0.02	0.00	0.02	0.00	0.94	0.95
		200	-0.00	0.20	0.01	0.00	0.01	0.00	0.96	0.96
0.5	0.1	50	-0.00	0.20	0.21	0.03	0.23	0.03	0.92	0.93
		100	-0.01	0.20	0.15	0.02	0.15	0.02	0.94	0.94
		200	-0.00	0.20	0.11	0.01	0.11	0.01	0.96	0.95
	0.3	50	0.01	0.20	0.24	0.02	0.24	0.02	0.94	0.94
		100	-0.00	0.20	0.17	0.01	0.16	0.01	0.95	0.96
		200	0.00	0.20	0.12	0.01	0.12	0.01	0.94	0.94
	0.5	50	0.01	0.20	0.26	0.02	0.25	0.02	0.95	0.95
		100	-0.00	0.20	0.18	0.01	0.18	0.01	0.94	0.95
		200	-0.00	0.20	0.13	0.01	0.12	0.01	0.95	0.95
1	0.1	50	-0.00	0.20	0.42	0.05	0.44	0.06	0.94	0.94
		100	0.03	0.20	0.31	0.04	0.30	0.04	0.95	0.95
		200	0.00	0.20	0.22	0.03	0.21	0.03	0.95	0.94
	0.3	50	-0.01	0.20	0.48	0.03	0.48	0.03	0.94	0.94
		100	0.00	0.20	0.33	0.02	0.33	0.02	0.95	0.95
		200	0.00	0.20	0.23	0.02	0.24	0.02	0.95	0.94
	0.5	50	0.02	0.20	0.52	0.03	0.52	0.03	0.94	0.95
		100	0.00	0.20	0.36	0.02	0.36	0.02	0.95	0.95
		200	-0.00	0.20	0.25	0.02	0.26	0.02	0.94	0.95

Scenario II: Similar to Scenario I, this simulation was set up to imitate the genome structure of a human gene. In this case each genotype $\mathbf{g}_i$ is a sequence of 20 elements $g_{ik} \in \{0, 1, 2\}$ where $i = 1, \cdots n$, and $k = 1, \cdots, 20$, with the frequency distribution of *MTR* gene [84]. A covariate z_i was introduced to the model that is uniformly distributed on $(0, 1)$ and an error term ε_i was drawn from normal distribution $N(0, .5^2)$. We use the true model: $y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + \varepsilon_{ij}$, where, $\boldsymbol{\beta} = (0, 1, 1)'$ and $x_{ij} = \sum_{k=1}^{p} |g_{ik} - g_{jk}|/(2p)$ (normalized Manhattan distance), $z_{ij} = (z_i^2 + z_j^2)^{1/2}$ and $\varepsilon_{ij} = \varepsilon_i - \varepsilon_j$. Based on 1000 replications, we summarized the Empirical Bias, MSE, ESD, and 95% CI CP in Table 3.2.

We can see that with all the considered sample sizes, the empirical mean biases are small, and estimated standard errors are close to the mean standard errors. Besides, as sample size increases, mean bias decreases and mean of standard errors approaches sample standard deviation of estimates, and the coverage probability of 95% confidence intervals are about the nominal level. The results suggest that when the underlying model between the pairwise distance matrices is linear, our proposed estimators are consistent.

Table 3.2: A summary of simulations based on Scenario II under the assumption that pairwise distances are linearly related through a multiple linear regression model.

n	Empirical Bias		MSE		ESD		95% CI CP	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
50	0.0015	-0.001	0.365	0.095	0.363	0.088	0.95	0.95
200	-0.007	-0.001	0.176	0.041	0.174	0.041	0.95	0.95
400	-0.005	-0.0004	0.124	0.028	0.125	0.028	0.95	0.96

3.2.2 Simulations for hypothesis testing

In our first two simulation scenarios, we make assumptions on having a true linear relationship between the distance matrices. Note that the distance-based methods are also widely used in association studies about the relationship between the orig-

inal data, that is, between X and Y. A considerable amount of literature has been published on the large sample theory of U-statistics and equivalency of distance covariances and kernel functions embedded in kernel machine regression. A kernel machine regression is a test of association between the original X and Y through a kernel function [79, 81, 84–99]. A question to ask is "Can the pairwise distance based linear regression model capture the underlying association of response and explanatory variables?" That is to say if there is a true association between the original variables, will the association retain between pairwise distance matrices?

Assume $(\mathbf{x}_i, \mathbf{y}_i), i = 1, 2, \cdots, n$, are n independent observations where $\mathbf{x}_i \in \mathcal{R}^p$ and $\mathbf{y}_i \in \mathcal{R}^q$. To test an association like $\mathbf{y}_i = \beta_0 + f(\mathbf{x}_i) + \varepsilon_i$ we set up the null hypothesis $H_0 : f(\cdot) = 0$. A useful type of test is a score test. Let

$$\begin{aligned} U &= (\mathbf{Y} - \bar{\mathbf{Y}})' K (\mathbf{Y} - \bar{\mathbf{Y}}) \\ &= tr[(\mathbf{Y} - \bar{\mathbf{Y}})' K (\mathbf{Y} - \bar{\mathbf{Y}})] \\ &= tr[K(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})] \\ &= tr\left[K\left(I - \frac{1_n 1_n'}{n}\right)\mathbf{Y}\mathbf{Y}'\left(I - \frac{1_n 1_n'}{n}\right)\right] \\ &= tr\left[KH\mathbf{Y}\mathbf{Y}'H\right] \end{aligned}$$

where $\bar{\mathbf{Y}}$ denotes a matrix of same dimensions of $\mathbf{Y}$ induced by column-wise means of $\mathbf{Y}$ embedded in each column, $H = \left(I - \frac{1_n 1_n'}{n}\right)$ and the kernel function $K = [k_{ij}]_{n\times n} = [k(\mathbf{x_i}, \mathbf{x_j})]_{n\times n}$ is a matrix of similarity or dissimilarity measures. The outer product $\mathbf{Y}\mathbf{Y}'$ are replaced with any symmetric distance matrix $D_{\mathbf{Y}}$ [90]. Let y_{ij} and x_{ij} be the (i, j)-th elements of $D_{\mathbf{Y}}$ and $D_{\mathbf{X}}$, respectively. For $y_{ij} = \alpha_0 + \alpha_1 \mathbf{x}_{ij} + \varepsilon_{ij}$, a choice of $D_{\mathbf{X}}$ (possibly any center-normalized distance matrix) and a kernel $K = [k_{ij}]_{n\times n}$ satisfying (a) $0 \le k_{ij} \le 1$ and (b) $\mathbf{x}_{ij} = 1 - k_{ij}$ implies $y_{ij} = \alpha_0^* + \alpha_1^* k_{ij} + \varepsilon_{ij}$. Let

$U_\alpha = \sum_{1\le i<j\le n} k_{ij}(y_{ij} - \bar{d_Y})$, then

$$\begin{aligned} U_\alpha &= \frac{1}{2}\sum_{1\le i,j\le n} k_{ij}(y_{ij} - \bar{d_Y}) \\ &= \frac{1}{2}\left(\sum_{1\le i,j\le n} k_{ij}\cdot y_{ij} - \bar{d_Y}\sum_{1\le i,j\le n} k_{ij}\right) \\ &\equiv \frac{U}{2} - \frac{\bar{d_Y}}{2}\sum_{1\le i,j\le n} k_{ij}. \end{aligned}$$

Consequently, $U_\alpha \sim N(0, 4V_0)$ and the original hypothesis $H_0 : f(\cdot) = 0$ can be reduced to $H_0 : \alpha_1^* = 0$.

Scenario III: This setup pertains to simulation study of kernel machine regression (KMR) by Hua and Ghosh [90]. Assuming $Y = c * h(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta} + \varepsilon$ as true model when $h(\cdot)$ is either $f_1(\cdot) = g(\cdot)$ or $f_2(\cdot) = sign(g)|g(\cdot)|^{1/2}$ and

$$g(X_i) = 1 + \sum_{k=1}^{p} X_{ik}\eta_k + \sum_{k=2}^{p} X_{i1}X_{ik}\gamma_k$$

with $\eta_1 = 0.4$, $\eta_2 = \cdots = \eta_p = 0.7$, $\gamma_2 = \cdots = \gamma_p = 0.2$. Quantifier c specified the departure from H_0 denoted as effect size. $X_{1\times p}$ was generated similar to Scenario II and covariate $\boldsymbol{Z}$ from standard bivariate normal distribution. The regression coefficients $\boldsymbol{\beta}$ =$(1, 1)$ and ε's were drawn from standard normal, student's $t_{df=3}$ or $\chi_1^2 - 1$ distributions—subtract 1 from χ_1^2 to achieve a mean of 0. For each sample, we first adjusted the effect of covariate using ordinary least squares regression and defined $\tilde{Y} = Y - \boldsymbol{Z}'\hat{\boldsymbol{\beta}}$. In this case, the distance based linear regression model in Equation 3.1 is $\tilde{y}_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$, where x_{ij} is the pairwise Manhattan distance. Note that when $c = 0$, it corresponds to the null hypothesis of no effect of X. We set significance level $\alpha = 0.05$ to check the Type-I error. To examine power, different c values were considered for f_1 (c = 0.1, 0.2, 0.3, 0.4, 0.5) and f_2 (c = 0.6, 1.2, 1.8, 2.4,

3) respectively. Quantifiers were assigned different range of values because the magnitudes of f_1 and f_2 are different. Results are shown in Table 3.3. Empirical size is the proportion of null hypothesis rejected when null hypothesis was true and empirical power is the proportion of null hypothesis rejected when null hypothesis was in fact false. For both size and power of the test, the level of significance was set to $\alpha = 0.05$. Empirical Type-I error for all combinations of n, f, c lies within (-0.012,0.009) of 0.05 by which we learn false positive rate is not inflated. Empirical power is small for sample size $n = 50$ even with highest effect size and rapidly increases as sample size increases. Empirical power gradually increases as effect size increases.

Table 3.3: $Y = c * f_j(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta} + \varepsilon$; No sign of inflated false positive rate is observed. Power of test is sensitive to both sample and effect size.

				Empirical size	Empirical power				
h	ε	n	c	0	0.1	0.2	0.3	0.4	0.5
f_1	$N(0,1)$	50		0.055	0.06	0.104	0.19	0.297	0.409
		200		0.058	0.065	0.351	0.76	0.918	0.981
		400		0.054	0.121	0.667	0.972	0.998	1
f_1	$t_{df=3}$	50		0.038	0.055	0.061	0.103	0.183	0.225
		200		0.054	0.049	0.115	0.301	0.563	0.735
		400		0.048	0.056	0.192	0.525	0.78	0.95
f_1	$\chi^2_1 - 1$	50		0.05	0.054	0.081	0.144	0.188	0.287
		200		0.054	0.052	0.152	0.415	0.694	0.841
		400		0.059	0.085	0.275	0.709	0.988	0.988
			c	0	0.6	1.2	1.8	2.4	3
f_2	$N(0,1)$	50		0.047	0.066	0.1	0.134	0.206	0.304
		200		0.057	0.065	0.226	0.568	0.748	0.892
		400		0.057	0.095	0.437	0.887	0.983	0.998

Scenario IV: Following the line of Scenario III, for multivariate outcome, k = 1, 2 and 3, data were generated from the model $Y_k = c * h_k(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta_k} + \varepsilon_k$ where $X_{1\times p}$ was generated using a frequency distribution of SNPs on the *SLC17A1* gene [90] with $p = 9$ SNPs. $\boldsymbol{Z}$ randomly sampled from bivariate normal distribution with means $\boldsymbol{\mu}$=$(0.2, 0.4)'$ and identity variance-covariance matrix, and ε_k has a multivariate nor-

mal distribution MVNormal($\mathbf{0}, \Sigma$). Two choices for the effects of h_k were considered. First, the sparse effect, where $h_1 = c(x_1 + x_2 + x_3 + x_1x_4x_5 - x_6/3 - x_7x_8/2 + (1 - x_9))$, for $c = 0, 0.1, 0.2$ and $h_2 = h_3 = 0$. Second, the common effect, where $h_1^* = h_1 + cx_3$, and $h_2 = h_3 = cx_3$ with $a = 0, 0.1, 0.2$. The variance–covariance matrix Σ was designed to have an independent structure ($\Sigma = \Sigma_1$) and a more dependent structure ($\Sigma = \Sigma_2$) as follows. Similar to scenario III, effect of covariate was adjusted using least squares regression, $\tilde{Y} = Y - \boldsymbol{Z}'\hat{\boldsymbol{\beta}}$. Then, the distance based linear regression model was carried out using theoretical model $\tilde{y}_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$. Results are shown in Table 3.4.

$$\Sigma_1 = \begin{bmatrix} 0.95 & 0 & 0 \\ 0 & 0.86 & 0 \\ 0 & 0 & 0.89 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 0.95 & 0.57 & 0.43 \\ 0.57 & 0.86 & 0.24 \\ 0.43 & 0.24 & 0.89 \end{bmatrix}$$

Table 3.4: Multivariate Outcome $Y_k = c * h_k(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta_k} + \varepsilon_k$; Inflated rate of false negative is observed due to complex structure of true model.

				Empirical size	Empirical power	
h	Σ	n	c	0	0.1	0.2
		50		0.041	0.044	0.088
$(h_1, 0, 0)$	Σ_1	200		0.046	0.065	0.213
		400		0.052	0.087	0.452
		50		0.044	0.038	0.067
$(h_1, 0, 0)$	Σ_2	200		0.051	0.048	0.273
		400		0.048	0.077	0.476
			c	0	0.1	.2
		50		0.046	0.05	0.098
(h_1^*, h_2, h_3)	Σ_1	200		0.042	0.073	0.331
		400		0.054	0.119	0.633
		50		0.05	0.043	0.08
(h_1^*, h_2, h_3)	Σ_2	200		0.056	0.064	0.341
		400		0.044	0.081	0.572

Scenario V: The aim of this simulation study is to investigate the performance of DBLR in the world of compositional data. The structure of a data set X is referred as compositional when sample points comply with unit-sum constraint. Take for example the ecological community data where the relative abundance of microbial organisms per observation sums to one.

Aitchison [100] proposed to relax the unit-sum constraint by performing statistical analysis through log ratios. Among various forms of log-ratio transformations, the centered log-ratio transformation has attractive features and has been widely used. Based on [101], compositional data X was generated given $X_{ij}^{(k)} = W_{ij}^{(k)} / \sum_{l=1}^{p} W_{il}^{(k)}$, $i = 1, \ldots, n_k$, $j = 1, \ldots, p$, $k = 1, 2$ and $W_{n_k \times p}^{(k)} = (W_1^{(k)}, \ldots, W_{n_k}^{(k)})'$ denoting the matrices of unobserved bases. Let $Z_i^{(k)} = (Z_{i1}^{(k)}, \ldots, Z_{ip}^{(k)})$ be the log basis vectors, where $Z_{ij}^{(k)} = \log(W_{ij}^{(k)})$. For the observed compositional data $X^{(k)}$ $(k = 1, 2)$, the centered log-ratio matrices $Y^{(k)}$'s are defined by $Y_{ij}^{(k)} = \log\{X_{ij}^{(k)} / g(X_i^{(k)})\}$, $i = 1, \ldots, n_k$, $j = 1, \ldots, p$, $k = 1, 2$, where $g(x) = (\prod_{j=1}^{p} x_j)^{1/p}$ denotes the geometric mean of a vector x. Defining $G = I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p'$, this relation can be expressed as matrix form $Y_j^{(k)} = G \log(X_j^{(k)})$. According to scale-invariance property of the centered log-ratios, $X_j^{(k)}$ can be replaced by $W_j^{(k)}$ and obtain $Y_j^{(k)} = GZ_j^{(k)}$.

Simulation setup: Log basis vectors were generated from multivariate normal distribution, $Z_i^{(k)} \sim MVN_p(\mu_k, \Omega)$ $(k = 1, 2$ and $i = 1 \ldots n)$. The components of μ_1 were drawn from a uniform distribution $U(0, 10)$. Null and alternative hypothesis are considered as

$$\mathrm{H}_0 : \mu_2 = \mu_1 \text{ vs. } \mathrm{H}_a : \mu_{2j} = \mu_{1j} - \delta_j \omega_{jj}^{1/2} \left(\frac{\log p}{n} \right)^{1/2}, \; j = 1, \cdots, n$$

Then $W^{(k)}$ and $X^{(k)}$ were generated through the transformations $W_{ij}^{(k)} = \exp(Z_{ij}^{(k)})$ and $X_{ij}^{(k)} = W_{ij}^{(k)} / \sum_{l=1}^{p} W_{il}^{(k)}$. Signal vector δ has support of size $s = 0$, $\lfloor 0.05p \rfloor$, $\lfloor 0.1p \rfloor$ or $\lfloor 0.5p \rfloor$, p=50, 100, 200. Non-zero elements of δ were drawn from $Unif[-2\sqrt{2}, 2\sqrt{2}]$ with index chosen uniformly from $\{1, \ldots, p\}$. For covariance matrix, Ω was defined $\Omega = D^{1/2} A D^{1/2}$, where D is a diagonal matrix with entries drawn from $Unif(1, 3)$

and A has non-zero entries $a_{jj} = 1$ and $a_{j-1,j} = a_{j+1,j} = -0.5$.

Analysis: To assess the differential composition of two groups, first centered log-ratio transformation was applied to observations, $Y_j = clr(X_j)$ and then distance matrix D_Y was constructed for response variable. An indicator function was used to label elements of response as either within (w) or between groups (b). Using DBLR model

$$y_{ij} = \beta_0 I(\textit{samples } i \textit{ and } j \textit{ in group 1}) + \beta_1 I(\textit{at least one sample from group 2}) + \varepsilon_{ij},$$

and tested for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. Rejection of test concludes that there is shift in composition of group 2 regarding the composition of group 1; however it does not infer that dispersion of group 2 is different from group 1. See Table 3.5 for empirical size and power of tests. Note that the two groups have same size $n_1 = n_2 = 100$ and significance level $\alpha = 0.05$.

Table 3.5: Empirical Size and Power of tests concerning the differential composition of groups. False positive rate if less than 1% and power increases as n or p increases.

s	$p = 50$	$p = 100$	$p = 200$
0	0.004	0.006	0.004
$\lfloor 0.05p \rfloor$	0.018	0.07	0.299
$\lfloor 0.1p \rfloor$	0.108	0.383	0.84
$\lfloor 0.5p \rfloor$	0.992	1	1

In continuation, two samples were generated in similar manner except $Z_i^{(1)}$ and $Z_i^{(2)}$ correlated and generated from a $2p$-dimensional joint distribution with mean $\mu^* = (\mu_1^{(k)T}, \mu_2^{(k)T})'$ and variance-covariance matrix

$$\Omega^* = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \otimes \Omega.$$

DBLR was fitted similar to the previous part and results are summarized in Table 3.6. Empirical power follows the same trend as before. Moreover, power of the tests show

less sensitivity to magnitude of p compared to uncorrelated design.

Table 3.6: Empirical Power for Samples with correlated mean structure.

s	$p = 50$	$p = 100$	$p = 200$
$\lfloor 0.05p \rfloor$	0.031	0.168	0.407
$\lfloor 0.1p \rfloor$	0.29	0.712	0.985
$\lfloor 0.5p \rfloor$	1	1	1

Scenario VI: In scope of analysis of differential composition, data was simulated to imitate community microbiome. Two groups of samples were generated with identical or differential mean relative abundance, incorporating a negative binomial distribution with fixed parameters [102], that is

$$x_i \sim NB\left(x_i \middle| \mu_i, \theta_i\right) = \frac{\Gamma\left(x_i + \theta_i\right)}{\Gamma\left(\theta_i\right) x_i!} \cdot \left(\frac{\theta_i}{\mu_i + \theta}\right)^{\theta_i} \cdot \left(\frac{\mu_i}{\mu_i + \theta_i}\right)^{x_i}, \tag{3.15}$$

where μ_i and $\phi_i = \frac{1}{\theta}$ are the mean and the dispersion parameter, respectively, and $\Gamma(\cdot)$ is the gamma function. In the microbiome setting, μ_{ij} is considered as a product of the mean relative abundances $\rho_j = \frac{\sum_{i=1}^{n} x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}}$ of taxon j and the library size $s_i = \sum_{j=1}^{p} x_{ij}$ of sample i, that is $\mu_{ij} = \rho_j s_i$. Library sizes were estimated using random subsets of taxonomic profiles of stool samples in HMP1 with replacement. Parameters ρ and θ were set to vectors of fixed values. Two groups of taxa ($p = 250$) tables with equal number of observations ($n_1 = n_2 = 50, 100, 200$) were simulated under the assumptions: (I) no differential abundance, i.e. fold changes were set to 1, (II) differential abundance by multiplying a fraction (true positive rate TPR = 0.25, 05, 0.75) of means and a fold change (FC = 1.5, 3, 5). Bray-Curtis distance [73], $BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$, is a commonly used distance measure in ecological data, where S_i is the total number of specimens counted on site i, S_j is the total number of specimens counted on site j, and, C_{ij} is the sum of only the lesser counts for each specimen found on both sites. Bray-Curtis distances were computed and labeled as

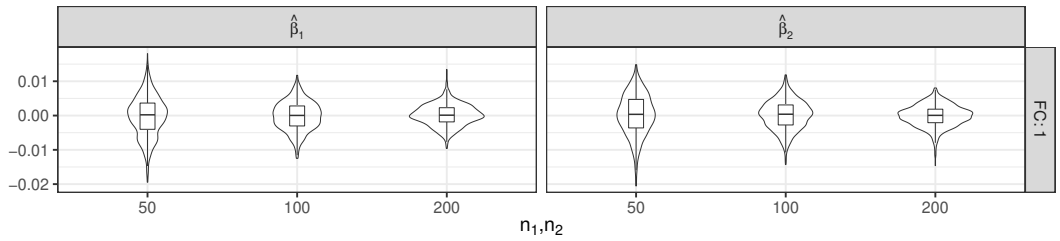
within group 1, within group 2 and between groups. Considering the between-group distances as reference group, the DBLR model is

$$BC_{ij} = \beta_0 + \beta_1 I_{\text{within group 1}} + \beta_2 I_{\text{within group 2}} + \varepsilon_{ij}, \tag{3.16}$$

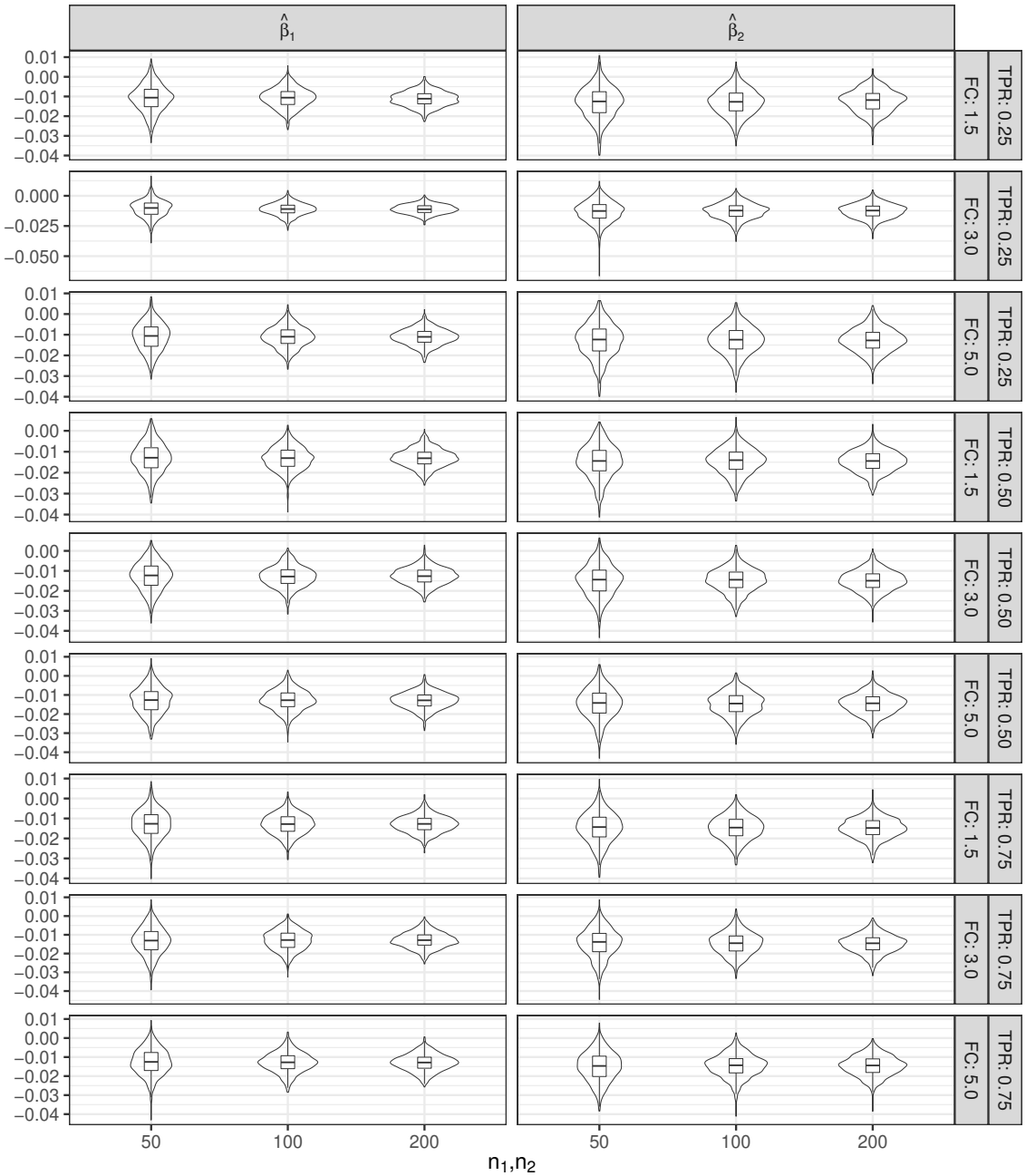
where $I(\cdot)$ is the indicator function. When $\beta_i > 0$ it can be interpreted as signal of additional variation of samples within that group to the mean distances of samples between two groups. When $\beta_i < 0$ it implies mean distances of samples within that group have less dissimilarities compared to their dissimilarities to other sample. However, an individual test cannot differentiate the cause whether it's due to a shift of centroid and/or scales of dispersion of groups. This might be investigated by multiple testing adjustment, testing the ratio of $\frac{\beta_1}{\beta_2}$, etc. Results are summarized in Table 3.7 and distribution of regression parameter estimates are visualized in Figure 3.2. There is no sign of inflated Type-I error. Empirical power seem to be more sensitive to sample size and less sensitive to TPR or FC. Power is low when $n = 50$ for any combination of TPR and FC. One might consider sample size estimation and power analysis.

Table 3.7: Summary of simulations for testing a difference of composition of microbiome of two groups. There is no sign of inflated Type-I error. Empirical power seem to be more sensitive to sample size and less sensitive to TPR or FC.

	n_1, n_2	Mean			MSE		ESD		P($p < 0.05$)	
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
FC=1	50	0.837	-0.000	0.000	0.006	0.006	0.006	0.006	0.047	0.049
	100	0.837	-0.000	0.000	0.004	0.004	0.004	0.004	0.048	0.046
	200	0.836	0.000	-0.000	0.003	0.003	0.003	0.003	0.053	0.053
TPR=0.25 FC=1.5	50	0.847	-0.011	-0.013	0.006	0.006	0.007	0.008	0.374	0.501
	100	0.848	-0.011	-0.013	0.004	0.005	0.005	0.007	0.670	0.742
	200	0.848	-0.011	-0.012	0.003	0.003	0.004	0.006	0.921	0.872
TPR=0.25 FC=3	50	0.847	-0.011	-0.013	0.006	0.007	0.007	0.009	0.370	0.502
	100	0.848	-0.011	-0.013	0.004	0.005	0.005	0.007	0.689	0.742
	200	0.848	-0.011	-0.013	0.003	0.003	0.004	0.006	0.916	0.879
TPR=0.25 FC=5	50	0.848	-0.011	-0.013	0.006	0.007	0.007	0.008	0.397	0.476
	100	0.847	-0.011	-0.013	0.004	0.005	0.005	0.007	0.689	0.728
	200	0.848	-0.011	-0.013	0.003	0.003	0.004	0.006	0.911	0.893
TPR=0.5 FC=1.5	50	0.850	-0.013	-0.015	0.006	0.007	0.007	0.008	0.529	0.596
	100	0.849	-0.013	-0.014	0.004	0.005	0.005	0.006	0.800	0.843
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.956	0.971
TPR=0.5 FC=3	50	0.849	-0.013	-0.015	0.006	0.006	0.007	0.008	0.485	0.609
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.819	0.871
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.946	0.973
TPR=0.5 FC=5	50	0.850	-0.013	-0.014	0.006	0.006	0.007	0.008	0.517	0.590
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.786	0.859
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.964	0.966
TPR=0.75 FC=1.5	50	0.850	-0.013	-0.015	0.006	0.007	0.007	0.007	0.525	0.584
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.800	0.859
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.952	0.980
TPR=0.75 FC=3	50	0.849	-0.013	-0.014	0.006	0.006	0.007	0.008	0.531	0.578
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.800	0.858
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.948	0.974
TPR=0.75 FC=5	50	0.850	-0.013	-0.015	0.006	0.007	0.007	0.008	0.501	0.616
	100	0.849	-0.013	-0.015	0.004	0.005	0.005	0.006	0.801	0.866
	200	0.849	-0.013	-0.015	0.003	0.003	0.004	0.005	0.945	0.964



(a) Groups 1 and 2 independently simulated from identical distribution.



(b) Groups 1 and 2 independently simulated from when TPR·(100)% of relative mean abundance of group 2 is different from group 1 by multiplying a fold changes FC.

Figure 3.2: Box plots with added violin plots for visualizing the distribution of regression coefficients.

3.2.3 Comparison: run time and empirical power

In this section we illustrate run time of different methods. We used scenario VI and computed run time for different sample sizes and number of permutations. The R code was run on a personal computer with *Apple M1 chip* and *R version 4.1.1.* A Heat map plot labeled with the run times in seconds is displayed in Figure 3.3. We used functions *adonis {vegan}* for PERMANOVA and *MRM {ecodist}* for MRM. For each combination of sample size, method and number of permutations, a simulation run time can be approximated by multiplying the number of iterations and each run time added up over all combinations.

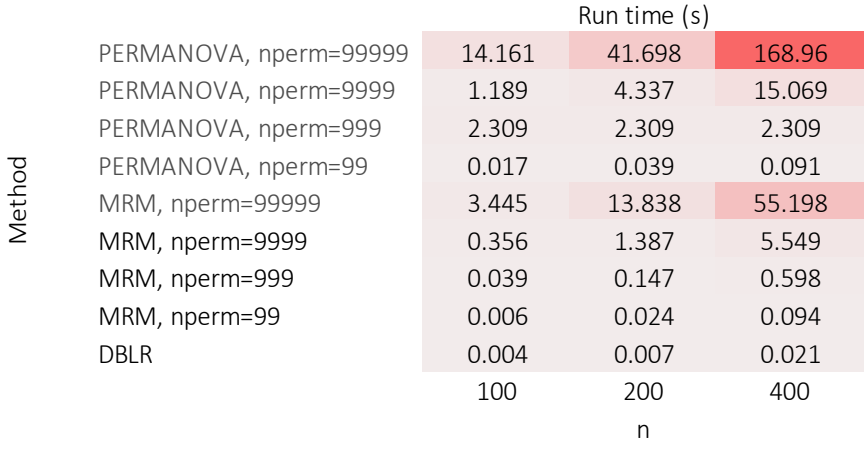


Figure 3.3: Values denote the run time in seconds for each method, number of permutations and sample size.

To compare the Empirical power of to methods MRM and DBLR, we used the setup of scenario III. Function f_1 and error terms drawn from $Norm(0, 1)$ and t_3 were used in this simulation. Results are summarized in Table 3.8.

Table 3.8: See scenario III where $Y = c * f_1(\boldsymbol{X}) + \boldsymbol{Z}'\boldsymbol{\beta} + \varepsilon$; REsults based on 1000 iterations and 999 permeation for MRM model.

		Pr(p.value<0.05)							
	c	0		0.1		0.3		0.5	
ε	n	DBLR	MRM	DBLR	MRM	DBLR	MRM	DBLR	MRM
	50	0.06	0.06	0.05	0.07	0.21	0.36	0.41	0.59
$N(0,1)$	200	0.06	0.05	0.08	0.11	0.78	0.87	0.98	0.99
	400	0.04	0.04	0.10	0.13	0.98	0.99	1.00	1.00
	50	0.05	0.04	0.04	0.05	0.11	0.17	0.25	0.39
t_3	200	0.05	0.05	0.06	0.06	0.32	0.37	0.73	0.79
	400	0.05	0.06	0.05	0.07	0.55	0.56	0.94	0.95

DBLR and MRM have same performance in terms of size and power. In both methods, power increases as sample size increases. This simulation was run on high performance computing cluster, *hpc*, utilizing 32 CPUs. The total run time for DBLR was about 500 seconds (less than 10 minutes), whereas the total run time for MRM (nperm=999) exceeded 6600 seconds (about 2 hours).

3.3 Real data example

Analyses in this section pertains to HMP2 data introduced in Chapter 2. As previously studied using intensive simulations in Chapter 3, distance based linear regression can serve as a tool to generate hypotheses about the community data composition. Distances at species level were computed and labeled as within and between UC, CD and nonIBD groups. Considering the within nonIBD distances as reference group, the DBLR model is

$$\begin{aligned} d_{ij} = \beta_0 &+ \beta_1 I_{\text{Within CD}} \\ &+ \beta_2 I_{\text{Between CD and UC}} \\ &+ \beta_3 I_{\text{Between CD and nonIBD}} \\ &+ \beta_4 I_{\text{ Within UC}} \\ &+ \beta_5 I_{\text{Between nonIBD and UC}} + \varepsilon_{ij}, \end{aligned} \tag{3.17}$$

where $I(\cdot)$ is the indicator function. The intercept parameter here is interpreted as mean distance of microbial composition of samples within the nonIBD group. Each other parameter is the difference of mean distances of other groups and nonIBD. Coefficient estimates, asymptotic standard errors and observed significance are summarized in Table 3.9 and Table 3.10. It's possible that DBLR has inflated Type-I error when Bray-Curtis distance method is used.

Table 3.9: DBLR summary of estimations based on Bray-Curtis distances of adult's gut microbiome species. Standard errors were computed via DBLR covariance estimation method.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7061	0.0091	77.5403	0.0000
Within CD	0.0622	0.0135	4.6172	0.0000
Between CD and UC	0.0742	0.0116	6.4048	0.0000
Between CD and nonIBD	0.0552	0.0075	7.3336	0.0000
Within UC	0.0699	0.0133	5.2431	0.0000
Between nonIBD and UC	0.0632	0.0075	8.4236	0.0000

Table 3.10: DBLR summary of estimations based on WUnif distances of adult's gut microbiome species. Standard errors were computed via DBLR covariance estimation method.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3174	0.0102	30.9843	0.0000
Within.CD	0.0525	0.0198	2.6473	0.0081
Between.CD.and.UC	0.0571	0.0160	3.5708	0.0004
Between.CD.and.nonIBD	0.0390	0.0097	4.0165	0.0001
Within.UC	0.0531	0.0198	2.6821	0.0073
Between.nonIBD.and.UC	0.0367	0.0099	3.7044	0.0002

For convenience, Table 2.7a and Table 2.7b are re-displayed here as Table 3.11 and Table 3.12.

Table 3.11: MRM coefficients and p-values with 9999 permutations when Bray-Curtis distance is used.

	Distance	pval
Int	0.7061	1.0000
Within.CD	0.0622	0.0001
Between.CD.and.UC	0.0742	0.0001
Between.CD.and.nonIBD	0.0552	0.0001
Within.UC	0.0699	0.0002
Between.nonIBD.and.UC	0.0632	0.0001

Table 3.12: MRM coefficients and p-values with 9999 permutations when WUnifrac distance is used.

	Distance	pval
Int	0.3174	0.9972
Within.CD	0.0525	0.0295
Between.CD.and.UC	0.0571	0.0118
Between.CD.and.nonIBD	0.0390	0.0017
Within.UC	0.0531	0.0460
Between.nonIBD.and.UC	0.0367	0.0076

As it is expected, the coefficients estimates are identical. However, the p-values are different due to different methods of evaluating observed significance in MRM and DBLR. Both methods are sensitive to the choice of distance measure. It seems that the use of Bray-Curtis distance my cause lower p-values that might be a sign of high false positive rate.

3.4 List of R packages

A list of R packages that we often used for simulation and data analysis is shown in Table 3.13.

Table 3.13: List of R packages

Data preparation	dplyr 1.0.7, reshape2 1.4.4, tidyr 1.1.4, stringr 1.4.0
Parallel computation	parallel, doParallel 1.0.16, foreach 1.5.1
Distance construction	vegan 2.5-7, ecodist 2.0.7, usedist 0.4.0
Microbiome data tools	microbiome 1.14.0, HMP 2.0.1, phyloseq 1.36.0, SummarizedExperiment 1.22.0, curatedMetagenomicData 3.0.10, SimSeq 1.4.0
Distributions	combinat 0.0-8, mvtnorm 1.1-3, dirmult 0.1.3-4
Graphics	ggplot2 3.3.5, ggpubr 0.4.0, gridExtra 2.3, latex2exp 0.5.0, lefser 1.2.0, plotly 4.10.0
Documentation	kableExtra 1.3.4, knitr 1.36, markdown 1.1, rmarkdown 2.11, xtable 1.8-4

CHAPTER 4: DISCUSSION

As mentioned in Chapter 1, pairwise distance based statistical methods have been developed to study the functional relationship between these multivariate variables such as ecological community composition data. These methods fall into different categories such as distance based linear correlation association test (e.g. Mantel test), multiple regression on distance matrices (e.g. MRM), distance based multivariate analysis of variance (e.g. PERMANOVA). However, a substantial amount of research argue that these methods suffer from inflated Type-I error and lack of power due to naive application of permutation testing.

In Chapter 2, after a brief review of what "microbiome" is, we borrowed the metagenomic data from iHMP study and implemented some preliminary analysis and exemplified the applications of distance based multivariate techniques in microbiome data analysis, including alpha and beta diversity and further analysis of beta diversity using Mantel test, PERMANOVA and MRM. This demonstrates the complexity of properly finding the relationships between diseases due to differential composition of microbiome, for example the minor abundances of rare microbiota causing noise in analysis. The diversity of the microbiome is of utmost interest due to the high fluctuation in the measurement of the raw data. The measurements of community diversity, alpha and beta diversity primarily, once again present an opportunity to evaluate popular ratification methods used in real data of this kind. From the analy-

ses performed no presented method offered a robust accuracy in the inference between disease groups and control healthy group in the cohorts. Metagenomics literature is still heavily interested in new methods to analyze this new kind of high dimensional data reliably, as clear inference and extrapolation of important variable features between groups is not always achievable.

In Chapter 3, we proposed a distance based linear regression model (DBLR) that is a linear regression on distance matrices. Estimates of parameters of DBLR are no different from ordinary least squares regression, but p-values are rather computed using our proposed method of estimation of variance-covariance matrix on a large sample theory basis. This leverages high performance computing maximally to reduce the intensive permutation computations in exchange for a parametric assumption to assess the data. The performance of this model was assessed through simulation studies and finished by an example of application of DBLR on HMP metagenomic data set. A thousand of duplicates of each size data set was evaluated for sensitivity.

Our method substantially reduces the needed computing power, since it does not rely on permutation. However, one of the challenges in our method is sensitivity to sample size considering asymptotic convergence of distribution. The classical issue in all large variable data sets is observations of possible combinations describing classifications within the data. Practically this is very important as not all measurements in metagenomics have small variance or are precisely reproducible due to the nature of the method. Thus evaluating multiple sample sizes was of utmost importance. In this section the connection between pairwise distance regression and regression model for raw data was explained by equivalency of kernel machine regression and kernel distance covariance. Overall high accuracy was obtained in most tests performed here, typically over 0.9 percent. This method has shown a predilection to operate well as relatively low observations per variable, as well as being able to operate over multiple variants of the data set.

This method offers a unique way to evaluate large data. It would be of interest to expand this evaluation to new questions. Research questions that could be asked include the relationship between group centroids and dispersion, multiple testing adjustment, classification and model selection, study design and sample size determination. In reducing the data set to measurements of data between centroids and dispersion of the measured variables often allows for easier feature importance extraction. But the DBLR method could be applied to extremely large variable data sets such as metagenomics or proteomics where widely used methods fail to refine any accurate inference from the data. The adjustments to multiple testing would assist in multi-class containing data sets. This becomes of interest when trying to relate features of importance between geographical groups or in finely differentiated diseases, such as mental disorders. While large popularity of machine learning methods has taken over many big data applications, the computational intensity required to deploy and train these methods is often due to the internal evaluation of the highly dimensional data sets. Many machine learning methods could benefit from methods such as DBLR by incorporating it into a prepossessing or post processing step. As is the case with most methods of data inference or manipulation, clear understandings of how the operation affects the data must be clear. Overall, the work here demonstrates the need for tractable methods of analysis in large data sets. As the age of informatics expands large dimensional and hierarchical data sets will become more normal. Need of manageable methods to extract important features of the data sets will become invaluable, especially when considered the ruse if personalized medicine and subjective kinds of classifications.

REFERENCES

[1] H. Smith, R. Gnanadesikan, and J. B. Hughes, "Multivariate Analysis of Variance (MANOVA)," *International Biometric Society*, vol. 18, no. 1, pp. 22–41, 1962.

[2] R. F. Haase and M. V. Ellis, "Multivariate Analysis of Variance," *Journal of Counseling Psychology*, vol. 34, pp. 404–413, 10 1987.

[3] B. H. McArdle and M. J. Anderson, "Fitting multivariate models to community data: A comment on distance-based redundancy analysis," *Ecology*, vol. 82, no. 1, pp. 290–297, 2001.

[4] M. J. Anderson, "A new method for non-parametric multivariate analysis of variance," *Austral Ecology*, vol. 26, no. 1, pp. 32–46, 2001.

[5] N. MANTEL, "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research*, vol. 27, no. 2, pp. 209–220, 1967.

[6] E. J. Dietz, "Permutation Tests for Association between Two Distance Matrices," *Systematic Zoology*, vol. 32, no. 1, pp. 21–26, 1983.

[7] P. E. Smouse, J. C. Long, and R. R. Sokal, "Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence," *Systematic Zoology*, vol. 35, no. 4, pp. 627–632, 1986.

[8] P. LEGENDRE and M.-J. FORTIN, "Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data," *Molecular Ecology Resources*, vol. 10, pp. 831–844, 9 2010.

[9] J. W. Lichstein, "Multiple regression on distance matrices: A multivariate spatial analysis tool," *Plant Ecology*, vol. 188, no. 2, pp. 117–131, 2007.

[10] H. Hotelling, "The Economics of Exhaustible Resources," *Journal of Political Economy*, vol. 39, 4 1931.

[11] S. S. Wilks, "Certain Generalizations in the Analysis of Variance," *Biometrika*, vol. 24, 11 1932.

[12] R. A. FISHER, "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS," *Annals of Eugenics*, vol. 7, 9 1936.

[13] M. S. Bartlett, "A note on tests of significance in multivariate analysis," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 35, 4 1939.

[14] D. N. Lawley, "A Correction to A Generalization of Fisher's z Test"," *Biometrika*, vol. 30, 1 1939.

[15] K. C. S. Pillai, "Some New Test Criteria in Multivariate Analysis," Tech. Rep. 1, 1955.

[16] M. J. R. Healy, C. R. Rao, and S. K. Mitra, "Generalized Inverse of Matrices and its Applications.," *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, 1972.

[17] Y. Fujikoshi, T. Himeno, and H. Wakaki, "Asymptotic Results of a High Dimensional MANOVA Test and Power Comparison When the Dimension is Large Compared to the Sample Size," *JOURNAL OF THE JAPAN STATISTICAL SOCIETY*, vol. 34, no. 1, 2004.

[18] A. P. Dempster, "A High Dimensional Two Sample Significance Test," Tech. Rep. 4, 1958.

[19] A. P. Dempster, "A Significance Test for the Separation of Two Highly Multivariate Small Samples," Tech. Rep. 1, 1960.

[20] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties," *Biometrics*, vol. 27, p. 857, 12 1971.

[21] J. Gower, "Properties of Euclidean and non-Euclidean distance matrices," *Linear Algebra and its Applications*, vol. 67, 6 1985.

[22] J. C. Gower and P. Legendre, "Metric and Euclidean Properties of Dissimilarity Coefficients," tech. rep., 1986.

[23] J. C. Gower and W. J. Krzanowski, "Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance," *Journal of the Royal Statistical Society. Series C: Applied Statistics*, vol. 48, no. 4, pp. 505–519, 1999.

[24] M. J. Anderson and D. C. I. Walsh, "PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing?," Tech. Rep. 4, 2013.

[25] D. V. Lindley and M. R. Novick, "The Role of Exchangeability in Inference," *Source: The Annals of Statistics*, vol. 9, no. 1, pp. 45–58, 1981.

[26] M. J. Anderson, "Permutational Multivariate Analysis of Variance (PERMANOVA)," *Wiley StatsRef: Statistics Reference Online*, pp. 1–15, 2017.

[27] P. Legendre, M. Fortin, and D. Borcard, "Should the Mantel test be used in spatial analysis?," *Methods in Ecology and Evolution*, vol. 6, pp. 1239–1247, 11 2015.

[28] D. A. Griffith, "Spatial Autocorrelation," *Encyclopedia of Social Measurement*, pp. 581–590, 1 2004.

[29] F.-J. Lapointe and P. Legendre, "Comparison tests for dendrograms: A comparative evaluation," *Journal of Classification*, vol. 12, pp. 265–282, 9 1995.

[30] P. Legendre, "Comparison of permutation methods for the partial correlation and partial Mantel tests," *Journal of Statistical Computation and Simulation*, vol. 67, no. 1, pp. 37–73, 2000.

[31] L. J. Harmon and R. E. Glor, "Poor statistical performance of the mantel test in phylogenetic comparative analyses," *Evolution*, vol. 64, no. 7, pp. 2173–2178, 2010.

[32] D. I. Warton, S. T. Wright, and Y. Wang, "Distance-based multivariate analyses confound location and dispersion effects," *Methods in Ecology and Evolution*, vol. 3, pp. 89–101, 2 2012.

[33] S. M. Bloom, V. N. Bijanki, G. M. Nava, L. Sun, N. P. Malvin, D. L. Donermeyer, W. M. Dunne, P. M. Allen, and T. S. Stappenbeck, "Commensal Bacteroides Species Induce Colitis in Host-Genotype-Specific Fashion in a Mouse Model of Inflammatory Bowel Disease," *Cell Host & Microbe*, vol. 9, pp. 390–403, 5 2011.

[34] B. Hamidi, K. Wallace, C. Vasu, and A. V. Alekseyenko, "W_d* -test: Robust distance-based multivariate analysis of variance," *Microbiome*, vol. 7, no. 1, pp. 1–9, 2019.

[35] R. P. Franckowiak, M. Panasci, K. J. Jarvis, I. S. Acuña-Rodriguez, E. L. Landguth, M.-J. Fortin, and H. H. Wagner, "Model selection with multiple regression on distance matrices leads to incorrect inferences," *PLOS ONE*, vol. 12, p. e0175194, 4 2017.

[36] F. Zou, Z. Xu, and T. Vision, "Assessing the Significance of Quantitative Trait Loci in Replicable Mapping Populations," *Genetics*, vol. 174, pp. 1063–1068, 10 2006.

[37] G. A. Churchill and R. W. Doerge, "Naive Application of Permutation Testing Leads to Inflated Type I Error Rates," *Genetics*, vol. 178, pp. 609–610, 1 2008.

[38] Y. Huang, H. Xu, V. Calian, and J. C. Hsu, "To permute or not to permute," *Bioinformatics*, vol. 22, pp. 2244–2248, 9 2006.

[39] W. F. Christensen and B. N. Zabriskie, "When Your Permutation Test is Doomed to Fail," *The American Statistician*, pp. 1–11, 4 2021.

[40] L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight, "Defining the human microbiome," *Nutrition Reviews*, vol. 70, no. SUPPL. 1, 2012.

[41] J. Lederberg and A. McCray, "Ome Sweet Omics: a genealogical treasury of words," *The Scientist*, vol. 15, no. 7, pp. 8–8, 2001.

[42] J. Chen and D.-G. Chen, *Statistical Analysis of Microbiome Data with R - ICSA Book Series in Statistics*. Springer US, 2018.

[43] G. D. Wu, F. D. Bushmanc, and J. D. Lewis, "Diet, the human gut microbiota, and IBD," *Anaerobe*, vol. 24, pp. 117–120, 2013.

[44] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider, "An estimation of the number of cells in the human body," *Annals of Human Biology*, vol. 40, pp. 463–471, 11 2013.

[45] R. D. Berg, "The indigenous gastrointestinal microflora," 11 1996.

[46] D. C. Savage, "Microbial Ecology of the Gastrointestinal Tract," *Annual Review of Microbiology*, vol. 31, pp. 107–133, 10 1977.

[47] R. Sender, S. Fuchs, and R. Milo, "Revised Estimates for the Number of Human and Bacteria Cells in the Body," *PLOS Biology*, vol. 14, p. e1002533, 8 2016.

[48] J. R. Kelsen and G. D. Wu, "The gut microbiota, environment and diseases of modern society," *Gut Microbes*, vol. 3, pp. 374–382, 7 2012.

[49] G. G. Kaplan, "The global burden of IBD: from 2015 to 2025," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, pp. 720–727, 12 2015.

[50] R. B. Sartor, "Therapeutic correction of bacterial dysbiosis discovered by molecular techniques," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 16413–16414, 10 2008.

[51] T. Integrative, "The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease corresponding author," *Cell Host and Microbe*, vol. 16, no. 3, pp. 276–289, 2014.

[52] L. M. Proctor, H. H. Creasy, J. M. Fettweis, J. Lloyd-Price, A. Mahurkar, W. Zhou, G. A. Buck, M. P. Snyder, J. F. Strauss, G. M. Weinstock, O. White, and C. Huttenhower, "The Integrative Human Microbiome Project," *Nature*, vol. 569, pp. 641–648, 5 2019.

[53] J. M. Fettweis, M. G. Serrano, J. L. P. L. Brooks, D. J. Edwards, P. H. Girerd, H. I. Parikh, B. Huang, T. J. Arodz, L. Edupuganti, A. L. Glascock, J. Xu, N. R. Jimenez, S. C. Vivadelli, S. S. Fong, N. U. Sheth, S. Jean, V. Lee, Y. A. Bokhari, A. M. Lara, S. D. Mistry, R. A. Duckworth, S. P. Bradley, V. N. Koparde, X. V. Orenda, S. H. Milton, S. K. Rozycki, A. V. Matveyev, M. L. Wright, S. V. Huzurbazar, E. M. Jackson, E. Smirnova, J. Korlach, Y.-C. Tsai, M. R. Dickinson, J. L. P. L. Brooks, J. I. Drake, D. O. Chaffin, A. L. Sexton, M. G. Gravett, C. E. Rubens, N. R. Wijesooriya, K. D. Hendricks-Muñoz, K. K. Jefferson, J. F. Strauss, and G. A. Buck, "The vaginal microbiome and preterm birth," *Nature Medicine*, vol. 25, pp. 1012–1021, 6 2019.

[54] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R. A. White, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, C. Huttenhower, J. Bishai, K. Bullock, A. Deik, C. Dennis, J. L. Kaplan, H. Khalili, L. J. McIver, C. J. Moran, L. Nguyen, K. A. Pierce, R. Schwager, A. Sirota-Madi, B. W. Stevens, W. Tan, J. J. ten Hoeve, G. Weingart, R. G. Wilson, V. Yajnik, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, pp. 655–662, 5 2019.

[55] W. Zhou, M. R. Sailani, K. Contrepois, Y. Zhou, S. Ahadi, S. R. Leopold, M. J. Zhang, V. Rao, M. Avina, T. Mishra, J. Johnson, B. Lee-McMullen, S. Chen, A. A. Metwally, T. D. B. Tran, H. Nguyen, X. Zhou, B. Albright, B.-Y. Hong, L. Petersen, E. Bautista, B. Hanson, L. Chen, D. Spakowicz, A. Bahmani, D. Salins, B. Leopold, M. Ashland, O. Dagan-Rosenfeld, S. Rego, P. Limcaoco, E. Colbert, C. Allister, D. Perelman, C. Craig, E. Wei, H. Chaib, D. Hornburg, J. Dunn, L. Liang, S. M. S.-F. Rose, K. Kukurba, B. Piening, H. Rost, D. Tse, T. McLaughlin, E. Sodergren, G. M. Weinstock, and M. Snyder, "Longitudinal multi-omics of host–microbe dynamics in prediabetes," *Nature*, vol. 569, pp. 663–671, 5 2019.

[56] H. Mallick, E. A. Franzosa, L. J. Mclver, S. Banerjee, A. Sirota-Madi, A. D. Kostic, C. B. Clish, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences," *Nature Communications*, vol. 10, p. 3136, 12 2019.

[57] E. A. Franzosa, L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, and C. Huttenhower, "Species-level functional profiling of metagenomes and metatranscriptomes," *Nature Methods*, vol. 15, pp. 962–968, 11 2018.

[58] L. J. McIver, G. Abu-Ali, E. A. Franzosa, R. Schwager, X. C. Morgan, L. Waldron, N. Segata, and C. Huttenhower, "bioBakery: a meta'omic analysis environment," *Bioinformatics*, vol. 34, pp. 1235–1237, 4 2018.

[59] M. Schirmer, E. A. Franzosa, J. Lloyd-Price, L. J. McIver, R. Schwager, T. W. Poon, A. N. Ananthakrishnan, E. Andrews, G. Barron, K. Lake, M. Prasad, J. Sauk, B. Stevens, R. G. Wilson, J. Braun, L. A. Denson, S. Kugathasan, D. P. B. McGovern, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "Dynamics of metatranscription in the inflammatory bowel disease gut microbiome," *Nature Microbiology*, vol. 3, pp. 337–346, 3 2018.

[60] E. A. Franzosa, A. Sirota-Madi, J. Avila-Pacheco, N. Fornelos, H. J. Haiser, S. Reinker, T. Vatanen, A. B. Hall, H. Mallick, L. J. McIver, J. S. Sauk, R. G. Wilson, B. W. Stevens, J. M. Scott, K. Pierce, A. A. Deik, K. Bullock, F. Imhann, J. A. Porter, A. Zhernakova, J. Fu, R. K. Weersma, C. Wijmenga, C. B. Clish, H. Vlamakis, C. Huttenhower, and R. J. Xavier, "Gut microbiome structure and metabolic activity in inflammatory bowel disease," *Nature Microbiology*, vol. 4, pp. 293–305, 2 2019.

[61] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower, "Metagenomic microbial community profiling using unique clade-specific marker genes," *Nature Methods*, vol. 9, pp. 811–814, 8 2012.

[62] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L.-I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein, "Best practices for analysing microbiomes," *Nature Reviews Microbiology*, vol. 16, 7 2018.

[63] R. H. Whittaker, "EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY," *TAXON*, vol. 21, pp. 213–251, 5 1972.

[64] T. M. DeJong, "A Comparison of Three Diversity Indices Based on Their Components of Richness and Evenness," *Oikos*, vol. 26, no. 2, p. 222, 1975.

[65] C. H. Heip, P. M. J. Herman, and K. Soetaert, "Indices of diversity and evenness Project DISCLOSE View project Respiration in ocean margin sediments View project," *Oceanis*, vol. 24, no. 4, pp. 61–87, 1998.

[66] J. J. Sepkoski, "Alpha, beta, or gamma: where does all the diversity go?," *Paleobiology*, vol. 14, pp. 221–234, 2 1988.

[67] G. Stirling and B. Wilsey, "Empirical Relationships between Species Richness, Evenness, and Proportional Diversity," *the american naturalist*, vol. 158, no. 3, pp. 286–299, 2001.

[68] N. J. Gotelli and R. K. Colwell, "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness," *Ecology Letters*, vol. 4, 7 2001.

[69] L. Jost, "The Relation between Evenness and Diversity," *Diversity*, vol. 2, pp. 207–232, 2 2010.

[70] K. Li, M. Bihan, S. Yooseph, and B. A. Methé, "Analyses of the Microbial Diversity across the Human Microbiome," *PLoS ONE*, vol. 7, p. e32118, 6 2012.

[71] E. Pielou, "The measurement of diversity in different types of biological collections," *Journal of Theoretical Biology*, vol. 13, 12 1966.

[72] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association*, vol. 47, 12 1952.

[73] J. R. Bray and J. T. Curtis, "An Ordination of the Upland Forest Communities of Southern Wisconsin," *Ecological Monographs*, vol. 27, 10 1957.

[74] C. Lozupone and R. Knight, "UniFrac: a New Phylogenetic Method for Comparing Microbial Communities," *APPLIED AND ENVIRONMENTAL MICROBIOLOGY*, vol. 71, no. 12, pp. 8228–8235, 2005.

[75] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight, "Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities," *Applied and Environmental Microbiology*, vol. 73, pp. 1576–1585, 3 2007.

[76] M. Hamady, C. Lozupone, and R. Knight, "Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data," *The ISME Journal*, vol. 4, pp. 17–27, 1 2010.

[77] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight, "UniFrac: an effective distance metric for microbial community comparison," *The ISME Journal*, vol. 5, pp. 169–172, 2 2011.

[78] J. C. Gower, "Principal Coordinates Analysis," in *Wiley StatsRef: Statistics Reference Online*, Wiley, 3 2015.

[79] W. Hoeffding, "A Class of Statistics with Asymptotically Normal Distribution," *The Annals of Mathematical Statistics*, vol. 19, pp. 293–325, 9 1948.

[80] J. N. Arvesen, "Jackknifing U-Statistics," *The Annals of Mathematical Statistics*, vol. 40, no. 6, pp. 2076–2100, 1969.

[81] G. G. Gregory, "Large Sample Theory for U-Statistics and Tests of Fit," *The Annals of Statistics*, vol. 5, no. 1, pp. 110–123, 1977.

[82] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-statistics," *Annals of Statistics*, vol. 36, no. 2, pp. 844–874, 2008.

[83] B. E. Honoré and J. L. Powell, "Pairwise difference estimators of censored and truncated regression models," *Journal of Econometrics*, vol. 64, no. 1-2, pp. 241–278, 1994.

[84] D. Kong, A. Maity, F. C. Hsu, and J. Y. Tzeng, "Testing and estimation in marker-set association study using semiparametric quantile regression kernel machine," *Biometrics*, vol. 72, no. 2, pp. 364–371, 2016.

[85] H. Callaert and P. Janssen, "The Berry-Esseen Theorem for U-Statistics," *The Annals of Statistics*, vol. 6, no. 2, pp. 417–421, 1978.

[86] B. E. Honoré and J. L. Powell, "Pairwise Difference Estimators for Nonlinear Models," in *Identification and Inference for Econometric Models*, pp. 520–553, Cambridge University Press, 6 2005.

[87] P. Kumar Sen, "Robust Statistical Inference for High-Dimensional Data Models with Application to Genomics," Tech. Rep. 2&3, 2006.

[88] P. S. Zhong and S. X. Chen, "Tests for high-dimensional regression coefficients with factorial designs," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 260–274, 2011.

[89] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "EQUIVALENCE OF DISTANCE-BASED AND RKHS-BASED STATISTICS IN HYPOTHESIS TESTING," *The Annals of Statistics*, vol. 41, no. 5, pp. 2263–2291, 2013.

[90] W. Y. Hua and D. Ghosh, "Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies," *Biometrics*, vol. 71, no. 3, pp. 812–820, 2015.

[91] Q. He, T. Cai, Y. Liu, N. Zhao, Q. E. Harmon, L. M. Almli, E. B. Binder, S. M. Engel, K. J. Ressler, K. N. Conneely, X. Lin, and M. C. Wu, "Prioritizing individual genetic variants after kernel machine testing using variable selection," *Genetic Epidemiology*, vol. 40, no. 8, pp. 722–731, 2016.

[92] Z. Z. Tang, G. Chen, and A. V. Alekseyenko, "PERMANOVA-S: Association test for microbial community composition that accommodates confounders and multiple distances," *Bioinformatics*, vol. 32, no. 17, pp. 2618–2625, 2016.

[93] O. Paliy and V. Shankar, "Application of multivariate statistical techniques in microbial ecology," *Molecular Ecology*, vol. 25, no. 5, pp. 1032–1057, 2016.

[94] X. Zhan, X. Tong, N. Zhao, A. Maity, M. C. Wu, and J. Chen, "A small-sample multivariate kernel machine test for microbiome association studies," *Genetic Epidemiology*, vol. 41, no. 3, pp. 210–220, 2017.

[95] H. Koh, M. J. Blaser, and H. Li, "A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping," *Microbiome*, vol. 5, no. 1, pp. 1–15, 2017.

[96] M. Luz Calle, "Statistical analysis of metagenomics data," *Genomics and Informatics*, vol. 17, no. 1, 2019.

[97] T. He, S. Li, P. S. Zhong, and Y. Cui, "An optimal kernel-based U-statistic method for quantitative gene-set association analysis," *Genetic Epidemiology*, vol. 43, no. 2, pp. 137–149, 2019.

[98] N. B. Larson, J. Chen, and D. J. Schaid, "A review of kernel methods for genetic association studies," *Genetic Epidemiology*, vol. 43, pp. 122–136, 3 2019.

[99] M. Fuchs, R. Hornung, A. L. Boulesteix, and R. De Bin, "On the asymptotic behaviour of the variance estimator of a U-statistic," *Journal of Statistical Planning and Inference*, vol. 209, pp. 101–111, 2020.

[100] J. Aitchison, "The Statistical Analysis of Compositional Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, 1 1982.

[101] Y. Cao, W. Lin, and H. Li, "Two-sample tests of high-dimensional means for compositional data," *Biometrika*, vol. 105, no. 1, pp. 115–132, 2018.

[102] S. Hawinkel, F. Mattiello, L. Bijnens, and O. Thas, "A broken promise: microbiome differential abundance methods do not control the false discovery rate," *Briefings in Bioinformatics*, vol. 20, pp. 210–221, 1 2019.