7-2021

Statistical Modeling for High-dimensional Compositional data with Applications to the Human Microbiome

Thy Dao
University of Arkansas, Fayetteville

Statistical Modeling for High-dimensional Compositional data
with Applications to the Human Microbiome

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy in Mathematics

by

Thy Dao
Royal Melbourne Institute of Technology
Bachelor of Designs in Multimedia Systems, 2008
University of Arkansas
Master of Science in Operation Management, 2012
University of Arkansas
Master of Science in Computational Statistics and Analytics, 2016

July 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

_____
Qingyang Zhang, Ph.D.
Dissertation Director

_____
Tulin Kaman, Ph.D.
Committee Member

_____
Jung Ae Lee-Bartlett, Ph.D.
Committee Member

ABSTRACT

Compositional data refer to the data that lie on a simplex, which are common in many scientific domains such as genomics, geology, and economics. As the components in a composition must sum to one, traditional tests based on unconstrained data become inappropriate, and new statistical methods are needed to analyze this special type of data. This dissertation is motivated by some statistical problems arising in the analysis of compositional data. In particular, we focus on the high-dimensional and over-dispersed setting, where the dimensionality of compositions is greater than the sample size and the dispersion parameter is moderate or large. In this dissertation, we consider a general problem of testing for the compositional difference between K populations. We propose a new Bayesian hypothesis, together with a nonparametric and distance-based testing method. Furthermore, we utilize multiple variable-selecting models, including LASSO, elastic net, ridge regression and cumulative logit model, to identify the most important subset of variables. This dissertation is structured as follows:

Chapter 1 introduces the compositional microbiome data, and then briefly review different statistical tests and model to be used in our framework, including distance correlation, LASSO, Ridge regression, elastic net, cumulative logit and adjacent-category logit model.

Chapter 2 then presents our new statistical test together with two real world applications form human microbiome study. We first formulate a hypothesis from the Bayesian point of view and suggest a nonparametric test based on inter-point distance to evaluate statistical significance. Unlike most existing tests for compositional data, the distance-based method is more sensitive to the compositional difference than the mean-based method, especially when the data are over-dispersed or zero-inflated. It does not rely on any data transformation, sparsity assumption or

regularity conditions on the covariance matrix, but directly analyzes the compositions. The performance of this method is evaluated using simulation studies. We apply this new procedure to two human microbiome datasets including a throat microbiome dataset and an intestinal microbiome data.

In addition to the overall testing, we also want to identify a small subset of variables that distinguish different populations. Chapter 3 introduces the procedure to select most significant variables (bacteria or genus) using LASSO, ridge regression, elastic net, cumulative logit model and adjacent-category logit models. Chapter 4 validates our findings from Chapter 3 and presents visualizations using multi-dimensional scaling (MDS).

Chapter 5 discusses and concludes the dissertation with some future perspectives.

.

ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Dr. Qingyang Zhang for his support and continued help to my study, especially to my dissertation research.

In addition, special thanks to my parents who have been supporting my academic pursuit.

DEDICATION

To

Thu Van Dao

Hang Thi Anh Truong

Tho Anh Dao

Vinh Quang Dao

TABLE OF CONTENTS

LIST OF FIGURES

Number *Page*

LIST OF TABLES

Page

Chapter 1

INTRODUCTION

1.1 Compositional Microbiome Data

Microbiome is defined as a community of microorganisms such as bacteria, fungi, and viruses that inhabit a particular environment and especially the collection of microorganisms living in or on the human body. The human body is home to about 100 trillion bacteria and other microbes, collectively known as microbiome. It has been widely accepted that human gut microbiome plays important role in human health, and it can be considered as a newly identified organ that interacts with other organs and influences the development of various diseases including cancers. In microbiome and metagenomic research, the data are often compositional and high-dimensional, which poses great challenge to the statistical test and modeling.

Compositional data refer to data that lie on the simplex, which can be expressed as follows:

$$S^{d-1} = \{ x_1, x_2, \dots, x_d \}, \qquad s.t., \min_j x_j \geq 0 \text{ and } \sum_{j=1}^{d} x_j = 1 ,$$

where d is the number of compositions and dimension is $d - 1$ due to the unit constraint [2-5].

The microbiome data are generally compositional. Due to varying amounts of DNA generating material across different samples, sequencing read counts are often normalized to relative abundances, making the observed data compositional. The compositional data can be viewed as a type of partially missing data, where only the proportions or compositions are known but the true abundances are unknown.

One fundamental problem in microbiome data analysis is to test whether two populations have the same microbiome composition, which can be viewed as a two-sample testing. Since the components

of a composition must sum to one, some traditional tests intended for unconstrained data such as two-sample t-test and Hoteling's t-test may result in inappropriate or misleading inferences.

As pointed out in [6], dataset derived from microbiome has its compositional nature that should not be ignored throughout the analysis. Microbiome data are usually collected by high-throughput sequencing (HTS) technique, and one major limitation is that the sequencing instruments usually fail to quantify read counts that exceed the capacity, making the observed compositional data biased. In addition, the difference between absolute abundance and relative abundance after sequencing is unpredictable. This analogy, thus, extents to any fixed capacity instrument where the size of total read count observed in an HTS is constant, the total count is random sample of the relative abundance of the molecules, which has no relation to the absolute number of the input sample.

Aitchison (1986) and Gregory (2017) pointed out several problems of the traditional methods which overlook the unit sum constraint. First, the collection of samples having exactly same size was inadequate. To solve such a problem, one can subsample the read counts for each sample, but this method may lead to loss of information. Many normalization methods have been used including the trimmed mean of M values (TMM) [11] and the median-matching method [12]. These two methods of transformation, however, are not compatible for highly sparse data, thus inappropriate when the number of molecules in the environment is unknown or poorly estimated. One important transformation for compositional data is the log-ratio transformation. Ratio transformations fully describe the relationships between the features in the dataset and the logarithm creates a symmetric and linear space. The resulting log-ratio abundances may well represent the abundance of each variable relative to other features in the dataset, while greatly reducing the negative dependence. The most widely used log-ratio transformation is the centered

log-ratio (clr) transformation introduced by Aitchison (1986) [6], which will be discussed with details in Chapter 3. Another popular method for analyzing compositional data is the UniFrac distance based on Bray-Curtis and Jensen-Shannon divergences, developed by Lozupone et al. (2011). The weighted version of UniFrac approach has been discussed by Silverman et al. (2017) [15].

Next, I will briefly review the notion of distance correlation, which will be the basis of our analysis.

1.2 Distance Correlation

Pearson correlation coefficient is the most widely used measure of dependence between two random variables. However, the major limitation of Pearson's correlation is that it only targets linear dependence, therefore may overlook important nonlinear dependence. Spearman's correlation may work for some nonlinear cases, but it assumes monotonic relation. For such disadvantages, Székely et al. (2005) introduced the distance correlation for measuring dependence between two random vectors of arbitrary type and arbitrary dimension. Unlike the Pearson's correlation coefficient, the distance correlation equals zero if and only if two random vectors are statistically independent, indicating that distance correlation measures both linear and nonlinear association between two variables or random vectors.

Another advantage of distance correlation is that it is fully nonparametric and model-free. Most traditional tests such as z-test and t-test assume that data follow univariate or low-dimensional normal distributions, and only target the mean difference, hence inappropriate for high-dimensional and compositional data. These considerations lead us to use distance correlation method for analyzing high dimensional compositional dataset.

Distance correlation is similar to Pearson's correlation in spirit, which is derived from the distance variance and distance covariance. In [17], Szekely et al. proposed the concept of distance correlation in the continuous setting, which is later translated into categorical setting by Zhang [18].

We begin with some basic notions of distance correlation to be used in the subsequent chapters.

Notations

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be two random vectors, where p and q are positive integers, $f_X, f_Y \; and \; f_{X,Y}$ are the marginal characteristic functions and joint characteristic function of X and Y, respectively. The inner product of vectors t and s is denoted by $< t, s >$.

Let $|X|_p$ be Euclidean norm of X in $\mathbb{R}^p$. The data matrix is denoted by X_{n*p} with dimension $n * p$ and the sample vectors (rows) are labeled $X_1, \dots, X_n$. If X_1 is an independent copy of X; they are independent and identically distributed (i.i.d.)

Now, we consider the problem of testing the joint independence of random vectors. For all distributions with finite first moments, we are seeking a dependence measure $R(X, Y)$ such that:

i. $R(X, Y)$ is defined for X and Y in arbitrary dimension.

ii. $R(X, Y) = 0$ characterizes independence of X and Y.

The two conditions mentioned above are well met by the distance correlation (R). In fact, for two random vectors of any type and any dimension, we have:

i. $0 \leq R \leq 1$

ii. $R = 0$ if and only if X and Y are statistically independent.

It is noteworthy that in the bivariate normal case, R is a function of product-moment correlation ρ, and $R(X, Y) \leq |\rho(X, Y)|$ with equality when ρ = ±1.

Now we setup the null and alternative hypotheses for independence test as below:

$$H_0 : f_{X,Y} = f_X f_Y \; vs. H_1 : f_{X,Y} \neq f_X f_Y$$

It can be seen that the distance correlation R well reflect the distance $||f_{X,Y}(t,s) - f_X(t)f_Y(s)||$ between the joint characteristic function and the product of the marginal characteristic functions. Following the discussion in Szekely et al. (2007), the distance correlation hold the premise to be applied as a very general dependence measure without assuming normality for valid inferences. To begin with, we state some preparatory definitions to derive the distance correlation measure:

Definition 1.1.

For complex functions γ defined on $\mathbb{R}^p \times \mathbb{R}^q$ *the* $||\cdot||_w$*-norm in the weighted L₂ space of functions on* $\mathbb{R}^{q+q}$ *is defined by*

$$\|\gamma(t,s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(t,s)|^2 w(t,s)\, dt\, ds,$$

where w(t,s) is an arbitrary positive weight function for which the integral above exists.

We may use the $||\cdot||_w$ -norm to define a measure of dependence with any acceptable choice of weight $w(t,s)$.

Definition 1.2.

Given characteristic functions f_X, f_Y *and* $f_{X,Y}$ *with weight w(t, s) we define the measure* $V^2(X, Y; w)$ *by*

$$V^2(X,Y;w) = \left\|f_{X,Y}(t,s) - f_X(t)f_Y(s)\right\|_w^2 = \int_{R^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)dt\, ds,$$

where it can be seen that $V^2(X, Y; w)$ *vanishes if and only if X and Y are independent.*

As V is seen as the absolute value of the classical product-moment covariance, we can thus unsigned correlation $R_w = \frac{V(X,Y;w)}{\sqrt{V(X;w)V(Y;w)}}$, where

$$V^2(X;w) = \int_{R^{2p}} \left|f_{X,X}(t,s) - f_X(t)f_X(s)\right|^2 w(t,s)dt\, ds,$$

R_w is required to be positive for dependent variables and scale variant. For $\epsilon > 0$, if the weight function $w(t,s)$ is integrable and both X and Y have finite variance, then by Taylor expansions of the underlying characteristic functions, we have

$$\lim_{\epsilon \to 0} \frac{V^2(\epsilon X,\ \epsilon Y\ ; w)}{\sqrt{V(\epsilon X; w)V(\epsilon Y; w)}} = \rho^2(X,Y)$$

thus if $\rho = 0$, R_w approaches to zero even if X and Y are dependent for integrable w. Furthermore, by Szekely et al. (2007), R_w is scale invariant and cannot be zero for dependent X and Y by applying a nonintegrable weight function, and it leads to the following lemma.

Lemma 1.3.

If $0 < \alpha < 2$, then for all x in $\mathbb{R}^d$

$$\int_{R^d} \frac{1 - \cos(t,s)}{|t|_d^{d+\alpha}} dt\, ds = C(d,\alpha)|x|^\alpha,$$

where $C(d,\alpha) = \frac{2\pi^{\frac{d}{2}}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma(\frac{d+\alpha}{2})}$ and $\Gamma(\cdot)$ is the complete gamma function.

The integrals at 0 and ∞ are meant in the principal value sense.

For $\alpha = 1$, the constant $C_d = C(d,1) = \frac{\pi^{(\frac{1+d}{2})}}{\Gamma(\frac{d+1}{2})}$, then by lemma 1.3, it is natural to choose the weight function $w(t,s) = \left(c_p c_q |t|_p^{p+1} |s|_q^{q+1}\right)^{-1}$, thus $dw = \left(c_p c_q |t|_p^{p+1} |s|_q^{q+1}\right)^{-1} dtds$.

Given the weight function and the corresponding weighted $L_2 - norm\ ||.||$, the dependence measure is written as $V^2\ (X,Y)$:

$$V^2\ (X,Y) = \int_{R^{p+q}} \left|f_{X,Y}(t,s) - f_X(t)f_Y(s)\right|^2 dw,$$

It is sufficient that $E|X|_p < \infty\ and\ E|Y\ |_q < \infty$, then by the Cauchy-Bunyakovsky inequality

$$\left|f_{X,Y}\ (t,s) - f_X(t)f_Y(s)\right|^2 = \left[E\ \left(e^{\ i(t,X)} - f_X(t)\right)\left(e^{\ i(t,Y)} - f_Y(t)\right)\right]^2$$

$$\leq E\left[e^{i(t,X)} - f_X(t)\right]^2 E\left[e^{i(t,Y)} - f_Y(t)\right]^2$$

$$= (1 - |f_X(t)|^2)(1 - |f_Y(s)|^2).$$

If $E(|X|_p + |Y|_q) < \infty$, then by an application of Fubini's theorem,

$$\int_{\mathbb{R}^p} \left|f_{X,Y}(t,s) - f_X(t)f_Y(s)\right|^2 dw \leq \int_{\mathbb{R}^p} \frac{1-|\mathrm{f_X}(t)|^2}{c_p|t|_p^{p+1}} dt \int_{\mathbb{R}^q} \frac{1-|\mathrm{f_Y}(s)|^2}{c_q|t|_q^{q+1}} ds$$

$$= E\left[\int_{\mathbb{R}^p} \frac{1-\cos(t,X-X')}{c_p|t|_p^{p+1}} dt\right] E\left[\int_{\mathbb{R}^q} \frac{1-\cos(s,X-X')}{c_q|t|_q^{q+1}} ds\right]$$

$$= E|X-X'|_p \, E|X-X'|_q < \infty$$

This leads us to the next definition.

<u>*Definition 1.4.*</u>

The distance covariance (dCov) between random vectors X and Y with finite first moments is the nonnegative number V(X, Y) defined by

$$V^2(X,Y) = \left\|f_{X,Y}(t,s) - f_X(t)f_Y(s)\right\|^2 = \frac{1}{C_pC_q} \int_{R^{p+q}} \frac{\left|f_{X,Y}(t,s) - f_X(t)f_Y(s)\right|^2}{|t|_p^{1+p}|s|_q^{1+q}} dt\, ds,$$

Similarly, distance variance (dVar) is defined as the square root of

$$V^2(X) = V^2(X,X) = \left\|f_{X,Y}(t,s) - f_X(t)f_Y(s)\right\|^2$$

<u>*Definition 1.5.*</u>

The distance correlation (dCor) between random vectors X and Y with finite first moments is the nonnegative number R(X, Y) defined by

$$R^2(X,Y) = \begin{cases} \dfrac{V^2(X,Y)}{\sqrt{(V^2(X)V^2(Y)}}, & V^2(X)V^2(Y) > 0 \\ 0, & V^2(X)V^2(Y) = 0 \end{cases}$$

The explicit relation between $V, R, and\ \rho$ in the bivariate normal case can be derived though this definition of R. The next definition is the empirical estimate of pre-defined distance correlation measure.

Definition 1.6.

For an observed random sample $(X, Y) = \{(X_k, Y_k) : k = 1, \ldots, n\}$ *from the joint distribution of random vectors* $X\ in\ \mathbb{R}^p\ and\ Y\ in\ \mathbb{R}^q$ *, define*

$$a_{kl} = |X_k - X_l|_p, \quad and\ b_{kl} = |Y_k - Y_l|_q \ \ where\ k, l = 1, \ldots, n.$$

$$A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}, and\ B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$$

In addition, $\bar{a}_{k.}$ is the k-th row mean, $\bar{a}_{.l}$ is the l-th column mean, and $\bar{a}_{..}$ is the grand mean of the distance matrix of the *X* sample.

The empirical distance covariance $V_n(X, Y)$ *is the nonnegative number defined by*

$$V_n^2(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{k,l} B_{k,l}$$

Similarly, $V_n(X)$ *is the nonnegative number defined by*

$$V_n^2(X) = V_n^2(X, X) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{k,l}^2$$

With above setup, one can prove that $V_n^2(X, Y) \geq 0$.

Definition 1.7.

The empirical distance correlation $R_n(X, Y)$ *is defined by*

$$R_n^2(X, Y) = \begin{cases} \dfrac{V_n^2(X, Y)}{\sqrt{(V_n^2(X) V_n^2(Y)}}, & if\ V_n^2(X) V_n^2(Y) > 0 \\ 0, & if\ V_n^2(X) V_n^2(Y) = 0 \end{cases}$$

$V_n(X) = 0$ if and only if all the observed samples are identical, leads to $A_{kl} = 0\ for\ all\ k, l = 1, \dots, n$. Particularly, $A_{kk} = \bar{a}_{k.} - \bar{a}_{.k} + \bar{a}_{..}$ approaches zero, implying that $\bar{a}_{k.} = \bar{a}_{.k} = \bar{a}_{..}/2$; and $A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..} = a_{kl} = |X_k - X_l|_p\ so\ X_1 = X_n$. We then try to show that R_n is also a good empirical measure of dependence.

Now, from the above definitions, we could define $V_n(X, Y)\ as\ ||f^n_{X,Y}(t,s) - f^n_{X(t)}f^n_Y(s)||$, where $f^n_{X,Y}(t,s) = \frac{1}{n}\sum_{k=1}^n \exp\{i < t, X_k > + i < s, Y_k > \}$,

$$f^n_X(t) = \frac{1}{n}\sum_{k=1}^n \exp\{i < t, X_k > \}\ and\ f^n_Y(t) = \frac{1}{n}\sum_{k=1}^n \exp\{i < t, Y_k > \}$$

are the empirical characteristic functions and the marginal empirical characteristic functions from the sample data $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ respectively.

Theorem 1.8.

If (X, Y) is a sample from the joint distribution of (X,Y), then

$$V_n^2(X,Y) = \left\| f^n_{X,Y}(t,s) - f^n_X(t) f^n_Y(s) \right\|^2$$

The equivalence of two definitions can be proved under Theorem 1.8 in continuous settings. The following properties of distance covariance and correlations can be also established (see [17] for proofs).

Theorem 1.9 (Properties of distance covariance)

(i) If $E(|X|_p + |Y|_q) < \infty, then\ 0 \le R \le 1$, and $R(X,Y) = 0\ iff\ X$ and Y are independent.

(ii) If $E(|X|_p^2 + |Y|_q^2) < \infty$, then given three independent samples we have

$V^2(X,Y) = E\big(|X_1 - X_2|_p\,|Y_1 - Y_2|_q\big) + E\big(|X_1 - X_2|_p\big)E\big(|Y_1 - Y_2|_q\big) - 2E\left(|X_1 - X_2|_p |Y_1 - Y_3|_q\right)$

Because of the aforementioned nice properties of distance correlation, we will use distance correlation test to the microbiome composition testing problem, that will be discussed in Chapter 2.

1.3 Penalized Regressions

As mentioned earlier, compositional data are often high-dimensional. For example, it is not uncommon that there are hundreds to thousands of bacteria, but only tens of samples. For the sake the result interpretability, it is crucial to identify a reduced set of variables that could distinguish different populations. In this dissertation, we choose to use a consensus set of variables identified by multiple regression models, to obtain a robust variable selection. We here review some widely used regression models and penalized regression models, which are to be used in chapter 3 for identifying short list of important taxa.

The first type of models we consider is penalized regression model, which roots in LASSO regularization. The main idea is to shrink the coefficients of the less contributive variables toward zero to imposes a penalty to the logistic model for having too many variables. The three most commonly used penalized regression include ridge regression, LASSO regression and elastic net Regression.

1.3.0 Regularization:

To illustrate how these penalized regression models work, we use the common scenario where we have an output variable (or response) Y and many input variables (predictors) $X_1, X_2, \dots, X_p$.

Fitting a regression model with a continuous response Y using all available inputs:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_p X_p + \epsilon$$

There are several reasons (by the principle of parsimony) that we should not consider all possible inputs X_i

- Model interpretability
- Losing degrees of freedom for error (particularly when the sample size n is small)
- Multicollinearity (the covariates are highly correlated with each other).

In Gaussian linear regression case, we estimate the parameter vector β using the matrix equation (derived from least square estimate or maximum likelihood estimate)

$$\boldsymbol{b} = (X^T X)^{-1} X^T Y$$

When the design matrix X is nearly singular, this can lead to inaccurate estimates of the parameters and their standard errors. The predicted model has the form

$$E(Y) = f(x) = X^T \boldsymbol{b},$$

where we minimize the sum of squared errors, i.e., loss function

$$L = \sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} \big(Y_i - E(Y)\big)^2 = \sum_{i=1}^{n} \big(Y_i - f(x_i)\big)^2$$

Shrinkage is then applied to minimize the following objective function with constraint:

$$\sum_{\mathrm{i=1}}^{\mathrm{n}} (Y_i - \mu)^2, subject\ to\ \mu^2 < C$$

Using Lagrange multiplier, the above task is equivalent to minimizing

$$\sum_{\mathrm{i=1}}^{\mathrm{n}} (Y_i - \mu)^2 + \lambda_c \mu^2$$

By differentiation, we then get

$$-2 \sum_{\mathrm{i=1}}^{\mathrm{n}} (Y_i - E(\mu_c))^2 + 2\lambda_c E(\mu_c) = 0$$

Finally, $E(\mu_c) = \widehat{\mu_c} = \frac{\sum_{i=1}^{n} Y_i}{n+\lambda_c} = K_c\bar{Y}$, $for\ K_c < 1$. Thus,

$\lambda_C\ is\ unimportant\ as\ C \to 0\ then\ \widehat{\mu_c} \to \bar{Y}\ and\ when\ C \to\ \infty\ then\ \widehat{\mu_c} \to 0$

In summary, regularization is to minimize objective function with the following form

$$\sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda R(f)$$

where the first part is sum of square errors, representing modeling fitting, and $R(f)$ in the second part is a penalty that regularizes model complexity. The constant λ is set as a tuning parameter. In ordinary linear regression, λ=0 and thus the regularizer $R(f)$ is irrelevant.

We are going to consider three variations of penalized regressions that use different regularization functions $R(f)$. We will use Ridge regression as an example to illustrate who to automatically selecting an optimal model containing the most contributive variables.

1.3.1. Ridge Regression:

Ridge regression is one of the oldest least square regression, originated in the early 1960s (Arthur Hoerl, 1962), where variables with minor contribution have their coefficients close to zero. However, all the variables are incorporated in the model.

The regularization function R(f) used in ridge regression is the l^2 norm as follows

$$R(f) = \sum_{i=1}^{p} \beta_i^2 = \|\beta\|_2$$

The problem becomes minimizing

$$SSE_\lambda(\beta) = \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{p-1} X_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{p} \beta_i^2$$

Solving through Lagrange multiplier to a quadratic constraint on β's, we get:

$$\frac{\partial}{\partial \beta_l} SSE_\lambda(\beta) = -2 < Y - X\beta, X_l > \ +2\,\lambda\beta_l$$

By setting the derivation to zero, we have

$$-2 < Y - X\beta_\lambda, X_l > \ +2\,\lambda\beta_{l,\lambda} = 0; \ \ 1 \le l \le p-1$$

which is equivalent to

$$-Y^T X + \beta_\lambda^T (X^T X + \lambda I) = \ 0$$

The ridge regression estimator is

$$\widehat{\beta_\lambda^*} = (X^T X + \lambda I)^{-1} X^T Y,$$

which is identical to the previous $\hat{\mu}_c$ in the matrix form.

In short, the ridge regression estimator $\widehat{\beta_\lambda^*}$ is a shrunken estimator of β. The L_2- norm regularizer balances two aspects (1) minimizing the sum of squared residuals (2) minimizing the sum of squared coefficients, and it can be expected that many of the $\widehat{\beta^*}$ will get smaller, or approach to zero.

1.3.2. LASSO Regression:

Least Absolute Shrinkage and Selection Operator (LASSO), proposed by Robert Tibshirani, 1996, is a another widely used penalized regression. Similar to Ridge regression, it shrinks the beta coefficients using a penalty term. However, unlike Ridge regression, LASSO utilizes the l^1-norm rather than squares:

$$R(f) = \sum_{\mathrm{i}=1}^{\mathrm{p}} \beta_i = \|\beta\|_1$$

ith the objective function as above:

$$SSE_\lambda(\beta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} (Y_i - \sum_{\mathrm{j}=1}^{\mathrm{p}-1} X_{ij}\beta_j)^2 + \lambda \sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_i|$$

It is noteworthy that the estimation of LASSO coefficients $\widehat{\beta^*}$ will be largely shrunken to $\widehat{\beta_j^*} = 0$. With that being said, the coefficients of many less contributive variables are forced to be zero and the final model will contain only the most significant variables.

1.3.3. Elastic Net Regression:

The elastic net combines both the l^1 and l^2 regularization of Ridge and the LASSO regression into a hybrid one. It shrinks some coefficients toward zero (like ridge regression) and set some coefficients to exactly zero (like LASSO regression), depending on the contribution of each predictor.

The estimate for an elastic net regression can be obtain as follows

$$\widehat{\beta^*} = \text{ argmin } (\|Y - X\beta\|^2 + \lambda_1\|\beta\|_1 + \lambda_2 \|\beta\|_2)$$

More precisely, the form of $\widehat{\beta^*}$ has three parts including sum of squared residuals, the l^1-norm and the l^2-norm. Ridge regression is a special case when there is only l^2 regularization ($\lambda_1 = 0$); while the LASSO is another special case with only l^1 regularization ($\lambda_2 = 0$). The ordinary least squares is the special case when there is no regularization, where both $\lambda_1 = \lambda_2 = 0$.

1.3.4. K-fold Cross-Validation:

For the purpose of statistical validation, the original dataset can be divided into two parts, namely the training set and testing set. The training set is used to build a model or prediction rule, the test set will be for validating the model. A more general procedure is k-fold cross-validation (e.g., 5-fold or 10-fold cross validation), where the data set is randomly divided into k groups of approximately equal size.

Instead of refitting the model n (sample size) times, we only refit the model k times. First, we will fit the model with k-1 out of k groups, and then use the test fold to make predictions and

compute the MSE in that fold. Repeating the procedure for each fold, an overall estimate of the MSE is computed as

$$CV_{(k)} = \frac{1}{k}\sum_{\mathrm{i=1}}^{\mathrm{k}} MSE_i$$

In addition to its efficiency, k-folds Cross-Validation can well balance two aspects, namely, bias and variance. In general, k can be chosen from 2 to 10, and most popular choice are k=5 or k=10.

In the two real data applications in Chapter 2, we will use 10-fold cross validation to choose the best λ in LASSO model estimation. In chapter 3, in addition to the aforementioned models, we will also use multi-category logit models to select significant variables.

1.4 Multi-category Logit Models

Logistic regression models, also known as logit models, can be generally used to model the probability of a certain class or event. Logits models have attracted much attention in machine learning, social science and medical community. For example, the Trauma and Injury Severity Score (TRISS), originally developed by Boyd et al., used logistic regression to predict the death rate in injured patients.

Logistic regression can be easily adjusted to deal with high dimensional data, by adding a penalty term like in LASSO. As the response variable must be categorical, logistic regression can be used as a simple classifier using different types of predictors including continuous variables, categorical variables and possibly some interaction terms. Moreover, logistic regression is widely used to predict

the risk of developing a given disease (e.g. cancer), based on observed characteristics of the patient such as age, sex, body mass index or smoking status.

1.4.1 Binary Logistic Regression

Binary logistic regression is the simplest Logit model where the dependent variable has two levels (coded 0/1), for example, Pass versus Fail, Smoker versus Non-smoker, Male versus Female. It estimates the probability that an event occurs given the values of explanatory variables, for instance to estimate probability of being smoker or non-smoker given age, gender and education level of a patient.

Formulation:

Let $Y = \{0,1\}$ be a binary response variable and $X = (X_1, X_2, \ldots, X_k)$ be a set of explanatory variables which can be discrete or continuous. Let x_i be the observed value of the explanatory variables for observation i.

Binary Logit model with a single predictor can be written as

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

or equivalently,

$$\pi_i = \Pr(Y_i = 1 \mid X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Below are key assumptions for logistic model

- $\{Y_1, Y_2, \ldots, Y_n\}$ are independently distributed.
- Y_i is Bernoulli, $Y_i \sim Ber(\pi_i)$.
- Linear relationship between the logit of the response and the explanatory variables (NOT assume a linear relationship between the dependent variable and the independent variables).

Parameters can be easily estimated using maximum likelihood estimation (MLE) for β $(\beta 0, \beta 1)$.

One can maximizes the likelihood function

$$L(\beta_0, \beta_1) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{n_i y_i} = \prod_{i=1}^{N} \frac{\exp\{y_i(\beta_0 + \beta_1 x_i)\}}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

There are no closed-form solutions, the MLE are obtained by using iterative algorithms such as *Newton-Raphson* (NR), or *Iteratively re-weighted least squares* (IRWLS) [See [1] Agresti (2013), sections 5.5.4-5.5.5.]

1.4.2 Multinomial Logistic Regression

Logistic regression can be extended to multi-category response variable, that is, *Y* can take more than 2 categories (r > 2).

Multinomial logistic models explain how a multinomial response Y depends on a set of k explanatory variables $X = (X_1, X_2, \dots, X_k)$ by assuming that $Y \sim$ Multinomial (n, π) where π is a vector with probabilities of occurrence for each category, but unlike binary case, multinomial model may have different types of link functions for ordinal response (consisting of ordered categories) and nominal response (consisting of unordered categories). For ordinal responses, cumulative logits model, adjacent categories model and continuation-ratio model are recommended.

1.4.3 Baseline Category Logit Model

Baseline-category Logit model is an extension of binary logistic regression model, where we consider a simultaneous summary of the odds ($r-1$ non-redundant logits) of being in one category relative to being in a designated category, called the baseline category, for all pairs of categories. Suppose that a response variable Y, each $y_i = (y_{i1}, y_{i2}, \dots, y_{ir})^T \sim$ Multinomial (n_i, π_i) where

$$n_i = \sum_{j=1}^{r} y_{ij} \quad ; \text{and } \pi_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{ir})^T$$

A set of explanatory variables $X = (X_1, X_2, \dots, X_k)$ can be mixture of discrete and continuous variables. The baseline category logit model is equivalent to loglinear model if the predictor variables are all categorical.

Taking the last category as the baseline category, the model becomes

$$\log\left(\frac{\pi_{ij}}{\pi_{ir}}\right) = x_i^T \beta_j \,, \qquad j \neq r$$

where x_i is the vector of X predictors of length p, i.e., this model has $(r-1) \times p$ free parameters. Therefore, we can use matrix form and the coefficients are

$$\beta_j = [\beta_{1j}, \beta_{2j}, \dots, \beta_{pj}]$$

Interpretation of Parameter Estimates

The k^{th} element of β_j can be interpreted as the increase in log-odds of category j versus category r resulting from a one-unit increase in the k^{th} predictor term, while keeping the other terms constant. The baseline-category (j=r) probability is

$$\pi_{ir} = \frac{1}{1 + \sum_{k \neq r} \exp\,(x_i^T \beta_k)}$$

and the non-baseline categories probability $j \neq r$ is

$$\pi_{ir} = \frac{\exp\,(x_i^T \beta_j)}{1 + \sum_{k \neq r} \exp\,(x_i^T \beta_k)}$$

1.4.4 Adjacent Category Logits

Another popular multi-category logit model is Adjacent Category Logit model, which compares the probabilities of any two adjacent categories, e.g., category 1 vs category 2, category 2 vs category 3, etc, given the values of all predictors in the model.

In other words, Adjacent Category Logits is similar to Baseline-Category Logit Model, but the baseline changes from one category to the next. Assuming that the response categories 1, 2, …, r are ordered, this comparison of adjacent-categories give more straightforward interpretation, e.g., young group vs middle-aged group, middle-aged group vs old group.

The adjacent-category logit models are defined as

$$L_1 = \log\left(\frac{\pi_1}{\pi_2}\right) = \beta_{10} + \beta_{11}X_1 + \cdots + \beta_{1p}X_p$$

$$L_2 = \log\left(\frac{\pi_2}{\pi_3}\right) = \beta_{20} + \beta_{21}X_1 + \cdots + \beta_{2p}X_p$$

…

$$L_{r-1} = \log\left(\frac{\pi_{r-1}}{\pi_r}\right) = \beta_{r-1,0} + \beta_{r-1,1}X_1 + \cdots + \beta_{r-1,p}X_p$$

Parameters in the models can be interpreted (e.g. the coefficient for β_1) in a similar way to the baseline-category model, i.e., by changing the value of X_1 by 1 unit while all the other variables X's remain constants, the odds changes by a factor of $\exp(\beta_1)$. In another word, β_1 is the change in the log-odds of category j + 1 versus category j when X_1 increases by one unit, holding all the other X-variables constant.

1.4.5 Proportional-Odds Cumulative Logit Model

Proportional-odds cumulative logit model is a frequently used model for ordinal response. This model uses cumulative probabilities up to a given thresholding category, thereby making the whole range of ordinal categories binary at that category.

Let $Y = 1, 2, \dots, J$ be the response where the ordering is natural. The associated probabilities are $\{\pi_1, \pi_2, \dots, \pi_j\}$, and a cumulative probability of a response less than equal to j is

$$P(Y \leq j) = \pi_1 + \cdots + \pi_j$$

Cumulative logit is then defined as

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \log\left(\frac{\pi_1 + \cdots . + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right)$$

and more precisely,

$$L_1 = \log\left(\frac{\pi_1}{\pi_2 + \cdots + \pi_r}\right) = \beta_{10} + \beta_{11}X_1 + \cdots + \beta_{1p}X_p$$

$$L_2 = \log\left(\frac{\pi_1 + \pi_2}{\pi_3 + \cdots + \pi_r}\right) = \beta_{20} + \beta_{21}X_1 + \cdots + \beta_{2p}X_p$$

$$\dots$$

$$L_{r-1} = \log\left(\frac{\pi_1 + \cdots + \pi_{r-1}}{\pi_r}\right) = \beta_{r-1,0} + \beta_{r-1,1}X_1 + \cdots + \beta_{r-1,p}X_p$$

The cumulative logit model has $(r - 1)$ intercepts plus $(r - 1)p$ slopes, for a total of $(r - 1)(p + 1)$ parameters to be estimated. In practice, we can assume the slopes are same for all equations, which gives us the proportional-odds cumulative logit models.

For simplicity, let us consider only one predictor

$$\text{logit}\,[P(Y \leq j)] = \alpha_j + \beta x$$

then the cumulative probabilities are

$$P(Y \leq j) = \frac{\exp\,(\alpha_j + \beta x)}{1 + \exp\,(\alpha_j + \beta x)}$$

The intercept α_j is the log-odds of all categories 1 to j when $X_1 = X_2 = \cdots = 0$, and β_k is the increase in log-odds of falling into or below any category associated with a one-unit increase in X_k,

while all the other *X*-variables remain unchanged. The odds-ratio is proportional to the difference between X_1 and X_2 where β is the constant of proportionality.

Chapter 2

DISTANCE CORRELATION FOR COMPOSITIONAL DATA ANALYSIS

2.1 Introduction

Compositional data refer to data vectors that lie on the simplex $S^{d-1} = \{ x_1, x_2, \dots, x_d \}$, $s.t.\,, \min_j x_j \geq 0$. As the components in a composition must sum to one, many classical statistical tests including two sample t-test and Wilcoxon rank-sum test became inappropriate. Two sample t-test target the mean difference, and relies on several assumptions including normal distribution, equal variance, unconstrained data and the independence of populations. Therefore, directly applying these standard methods to compositional data could result in misleading inference.

To overcome this difficulty, Aitchison (1982) proposed to use log-ratio transformation to relax the unit-sum constraint. However, his test only be applied to low dimensional settings where the dimensionality is less than sample size. Various methods have been developed since the work of Aitchison. In 2017, Cao et al. developed a powerful two-sample test for high-dimensional means using centered log-ratio transformation with statistical satisfactory under some regularity conditions and sparsity assumption. However, Cao et al.'s test has several shortcomings. For instances, it can only deal with two sample comparison, and its validity depends on a list of regularity conditions on the underlying covariance matrices, and its performance relies on the sparsity assumption, i.e., only a small proportion of components in the composition are different across groups.

To handle high-dimensionality and over-dispersion that are commonly seen in recent microbiome data, we consider a general problem of testing for the compositional difference between multiple populations. We formulated a new hypothesis from a Bayesian point of view, suggesting a non-parametric test based on inter-point distance to evaluate significance. Unlike most existing tests for compositional data, our method does not rely on any data transformation, sparsity assumption or

regularity conditions on the covariance matrix, but directly analyzes the compositions. The performance of the proposed method is tested by simulated high-dimensional, over-dispersed and zero-inflated. The proposed method is applied in two human microbiome data to test the association microbiome composition and the phenotype of interest.

2.2 Problem formulation

In this part, we will first briefly review the test by Cao et al. (2017).

Let $k \in \{1,2,\ldots,K\}$ be the group index and $j \in \{1,2,\ldots,p\}$ be the index of components in the composition, then the observed $n_k * p$ data matrix for group k can be denoted as

$X^{(k)} = \left(X_1^{(k)},\ldots,X_{n^k}^{(k)}\right)^T$, where $X_i^{(k)} = \left(X_{i1}^{(k)},\ldots,X_{ip}^{(k)}\right)^T$ represents the composition for subject i that lie on the $(p-1)$-dimensional simplex.

We assume that the observed compositional data $X^{(k)}$ arise from a latent matrix $W^{(k)} = \left(W_1^{(k)},\ldots,W_{n^k}^{(k)}\right)^T$ by normalization

$$X_{ij}^{(k)} = \frac{W_{ij}^{(k)}}{\sum_{h=1}^{p} W_{ih}^{(k)}}$$

where $W^{(k)}$ refers to the true abundance of bacterial taxa. Since the true abundances $W^{(k)}$ are unknown, Cao et al. (2017) formulated a new hypothesis for testing the difference between two groups:

$$H_0: E\left(\log\left(W_1^{(1)}\right)\right) = E\left(\log\left(W_1^{(2)}\right)\right) + c1_p, \text{for some } c \in \mathbb{R},$$

$$H_\alpha: E\left(\log\left(W_1^{(1)}\right)\right) \neq E\left(\log\left(W_1^{(2)}\right)\right) + c1_p, \text{for any } c \in \mathbb{R},$$

where 1_p stands for the vector of p ones. The above is mean-based hypothesis and can be tested using the centered log-ratio transformation

$$Y_{ij}^{(k)} = \log \frac{X_{ij}^{(k)}}{(\prod_{h=1}^{p} X_{ih}^{(k)})^{\frac{1}{p}}}, k = 1,2; i = 1, \dots, n_k.$$

The centered log-ratio variables $Y_{ij}^{(k)}{}'s$ can be shown to be only weakly dependent and satisfy certain desired statistical properties, resulting the following test statistics

$$M_n = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \le j \le p} \frac{(\bar{Y}_j^{(1)} - \bar{Y}_j^{(2)})^2}{\hat{\gamma}_{jj}},$$

$where\ \bar{Y}_j^{(k)} = \sum_{i=1}^{n_k} \frac{Y_{ij}^{(k)}}{n_k}, and\ \hat{\gamma}_{jj} = \sum_{k=1}^{2} \sum_{i=1}^{n_k} \frac{(Y_{ij}^{(k)} - \bar{Y}_j^{(2)})^2}{n_1 + n_2}$ and the *p*-value then be obtained through Gumbel distribution (also known as the log-Weibull and the double exponential distribution) [5]

$$p\_value = 1 - \{\exp[\exp(-.5M_n - 2\log p + \log\log p + \log\pi)]\}^{-1}.$$

It can be seen that Cao et al.'s test targets the mean difference in high-dimensional settings, and its validity relies on several assumptions on the underlying covariance matrices, which is hard to check in practice. Therefore in this work, we considered a different hypothesis on the distribution of composition instead of means. We assume $W_i^{(k)}$ follows the multinomial distribution,

$$W_i^{(k)} \sim \text{Multinomial}(N_i^{(k)}, \pi_i^{(k)}),$$

where $N_i^{(k)}$ represents the total abundance of bacterial taxa for sample *i* from group *k*, and $\pi_i^{(k)}$ represents the true composition.

In order to model over-dispersion, we assumed random parameters, $N_i^{(k)} \sim f_N(\alpha) and\ \pi_i^{(k)} = \left(\pi_{i1}^{(k)}, \dots, \pi_{ip}^{(k)}\right) \sim f_\pi\left(\Theta^{(k)}\right)$ where α and $\Theta^{(k)}$ are hyper-parameters.

We then define the compositional equivalence between two groups based on the distribution of parameter π:

Definition 2.1

Two groups k and k′ are said to be compositionally equivalent if

$$f_\pi\left(\Theta^{(k)}\right) = f_\pi(\Theta^{(k')}).$$

By definition 2.1, we formulate the null and alternative hypotheses among K groups:

$$H_0: f_\pi\left(\Theta^{(1)}\right) = \ldots = f_\pi\left(\Theta^{(k)}\right),$$

$$H_\alpha: f_\pi\left(\Theta^{(k)}\right) \neq f_\pi(\Theta^{(k')}) \; for \; some \; k \; and \; k'$$

In this framework, we assume that the total abundance $N_i^{(k)}$ is independent of $\pi_i^{(k)}$, and $N_i^{(k)} \sim f_N(\alpha)$ for $i \in \{1, \ldots, n_k\}$ and $k \in \{1,2, \ldots, K\}$, therefore testing H_0 is to test for distributional homogeneity of the compositions between K groups. Let $X_i^{(k)} \sim f_X^{(k)}(x)$, then we have the following equivalent hypothesis

$$H_0^*: f_X^{(1)}(x) = \cdots = f_X^{(k)}(x) \text{ for all x,}$$

$$H_\alpha^*: f_X^{(k)}(x) = f_X^{(k')}(x) \text{ for some x, k and k′,}$$

Where it can be seen that H_0^* is equivalent to testing the independence between the composition X and the grouping variable k ∈{1,2,..., K} (i.e., phenotype), i.e., testing the independence between a continuous random vector and a categorical variable.

2.3 Distance Based Test

This part presents the distance-based method [1] that we proposed to test the hypothesis H_0^* . Using the notion of distance covariance between two random vectors X and Y (see definition 1.4 in Chapter 1)

$$dCov^2(X,Y) = \int_{\mathbb{R}^{dx+dy}} \frac{\|\phi_{x,y}(t,s) - \phi_x(t)\phi_y(s)\|^2}{C_{d_x} C_{d_y} \|t\|_{d_x}^{1+d_x} \|s\|_{d_y}^{1+d_y}} dtds, \qquad (1)$$

where $\phi(\cdot)$ is a characteristic function, d_x and d_y are the dimensions of X and Y, $C_{d_x} = \frac{\pi^{\frac{1+d_x}{2}}}{\Gamma\left\{\frac{1+d_x}{2}\right\}}$ and $C_{d_y} = \frac{\pi^{\frac{1+d_y}{2}}}{\Gamma\left\{\frac{1+d_y}{2}\right\}}$, $\left||z|\right|_{d_z}$ denotes the Euclidean norm of $z \in \mathbb{R}^{d_z}$, and $\| \phi \|^2 = \phi \ \bar{\phi}$ for the complex-valued function ϕ and its conjugate $\bar{\phi}$.

It is noteworthy to mention that $dCov(X,Y) = 0$ if and only if X and Y are statistically independent (property of distance correlation discussed in Chapter 1), so this feature will capture any form of association between a continuous random vector X and a categorical variable Y. Szekely et al. (2007) has provided an alternative and equivalent definition of distance covariance based on Euclidean distance (Theorem 1 in [17])

$$dCov^2(X,Y) = Cov(\| X_1 - X_2 \|, \| Y_1 - Y_2 \|) - 2Cov(\| X_1 - X_2 \|, \| Y_1 - Y_3 \|),$$

where (X_1,Y_1), (X_2,Y_2) and (X_3,Y_3) be three independent copies of (X,Y).

This alternative definition is used to derive the explicit formula of distance covariance between composition X and phenotype Y. For $Y \in \{1,2,\ldots,K\}$ with probabilities $\{p_1,\ldots,p_K\}$ and $X = \{X_1,\ldots,X_p,$ we assume Y is nominal (without ordering between categories) for illustration purpose, with (X_1,Y_1), (X_2,Y_2) and (X_3,Y_3) be three independent copies of (X,Y). We define

$$\| Y_1 - Y_2 \| = 1, if\ Y_1 \neq Y_2\ and\ 0\ otherwise.$$

In addition, we define the expected inter-point distance as

$$D_{ij} = E(\| X_1 - X_2 \ \|_{Y_1=i,Y_2=j}), i,j = 1,\ldots,K.$$

Using the definitions above, the distance covariance between Y and X can then be derived as follows

$$E(\| Y_1 - Y_2 \|) = 1 - \sum_{i=1}^{K} p_i^2,$$

$$E(\| X_1 - X_2 \|) = \sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j D_{ij},$$

$$E(\| X_1 - X_2 \|\| Y_1 - Y_2 \|) = \sum_{i\neq j} p_i p_j D_{ij} = \sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j D_{ij} - \sum_{i=1}^{K} p_i^2 D_{ii}$$

$$E(\| X_1 - X_2 \|\| Y_1 - Y_3 \|) = \sum_{j=1}^{K}\sum_{i\neq 1}^{K} p_i p_j p_l D_{ij} = \sum_{i=1}^{K}\sum_{j=1}^{K} p_i(1-p_i) p_j D_{ij}$$

Summarizing the above, we get

$$d\,\widehat{Cov^2(X,Y)} = 2\sum_{i=1}^{K}\sum_{j=1}^{K} \hat{p}_i^2 \hat{p}_j \widehat{D}_{ij} - \sum_{i=1}^{K} \hat{p}_i^2 \widehat{D}_{ii} - (\sum_{i=1}^{K} \hat{p}_i^2)(\sum_{i=1}^{K}\sum_{j=1}^{K} \hat{p}_i \hat{p}_j \widehat{D}_{ij})$$

By Cauchy-Schwarz inequality, it can be shown that $dCov(X,Y) \geq 0$ and the equality holds if and only if $D_{ii} = D_{jj} = D_{ij}\ for\ all\ (i,j)'s$.

The maximum likelihood estimate (MLE) of p_i is $\hat{p}_i = \frac{n_i}{n}$, where n_i stands for the sample size of group i. The sample inter-point distance can be computed as follows

$$\widehat{D}_{ij} = \frac{1}{n_i n_j}\sum_{m=1}^{n_i}\sum_{l=1}^{n_j} ||X_m^{(i)} - X_l^{(i)}||\,, \qquad (2)$$

$$\widehat{D}_{ii} = \frac{2}{n_i(n_i-1)}\sum_{m=1}^{n_i}\sum_{l=1}^{n_j} ||X_m^{(i)} - X_l^{(i)}||\,, \qquad (3)$$

where $\left\{X_1^{(i)}, \dots, X_{n_i}^{(i)}\right\}\ and\ \left\{X_1^{(j)}, \dots, X_{n_j}^{(j)}\right\}$ are samples of X_i and X_j respectively. Finally, p-values are obtained using a simple permutation procedure described in [17].

Our proposed method was evaluated by an extensive simulation study where all the settings are high dimensional and over-dispersed. Using a fixed dimension $p = 200$ and different sample sizes $n_1 = n_2 = 50$ and $n_1 = n_2 = 100$, we generated the abundance $W_{ij}^{(k)}$ from 3 different settings as below:

Setting 1: $W_{ij}^{(k)} \sim \text{NegBin}\left(\mu_j^{(k)}, r_j^{(k)}\right)$, for $i = 1, \dots, n_i$, $j = 1, \dots, p$, $r_j^{(1)} \sim \text{Unif}(0.1,1), r_j^{(1)} = r_j^{(2)}$ and $\mu_j^{(1)} \sim \text{Unif}(10,15)$. Let $I = \{I_+, I_-\}$ be the set of taxa with different abundances in two conditions, $\mu_j^{(2)} = \mu_j^{(1)} + \Delta \; for \; j \in I_+ \; and \; \mu \; \mu_j^{(2)} = \mu_j^{(1)} - \Delta \; for \; j \in I_-$, $\mu_j^{(2)} = \mu_j^{(1)}$ for $j \notin I, |I_+| = |I_-| = dp$, where $|\cdot|$ represents set cardinality, and *d* is the proportion of differential means. Given *d*=5%,20%, representing relatively sparse and dense signals in mean difference, we used Δ={0.5,1.0,1.5,2.0,2.5,3.0}.

Setting 2: Same as Setting 1, but $\mu_j^{(1)} \sim Unif(5,10)$

Setting 3: (Negative binomial model with excess zeros): $W_{ij}^{(k)} = 0$ with probability $\pi = (10\%, 20\%)$, $W_{ij}^{(k)} \sim NegBin\left(\mu_j^{(k)}\right), r_j^{(k)}$) with probability $1 - \pi$, let *d*=10%, Δ={0.5,1.0,1.5,2.0,2.5}. Other settings remain same as in Setting 1.

The simulation abundance $W_{ij}^{(k)}$ are then normalized to the composition $X_{ij}^{(k)}$ to perform the test for the null hypothesis at the significant level of 0.05. We calculated test statistics M_n and p-value from Gumbel distribution for the log-ratio based method (Cao et al.'s test, [17]), then compute p-value from 5,000 permutations for our distance correlation test. For each setting, we simulate 1,000 datasets, and compare the true positive rates (TPRs) by the two tests. By comparing the true positive rates (TPRs) of the two tests, as shown in Figure 2.0.1-2.0.3, our distance-based method outperforms the Cao et al.'s method throughout all settings.

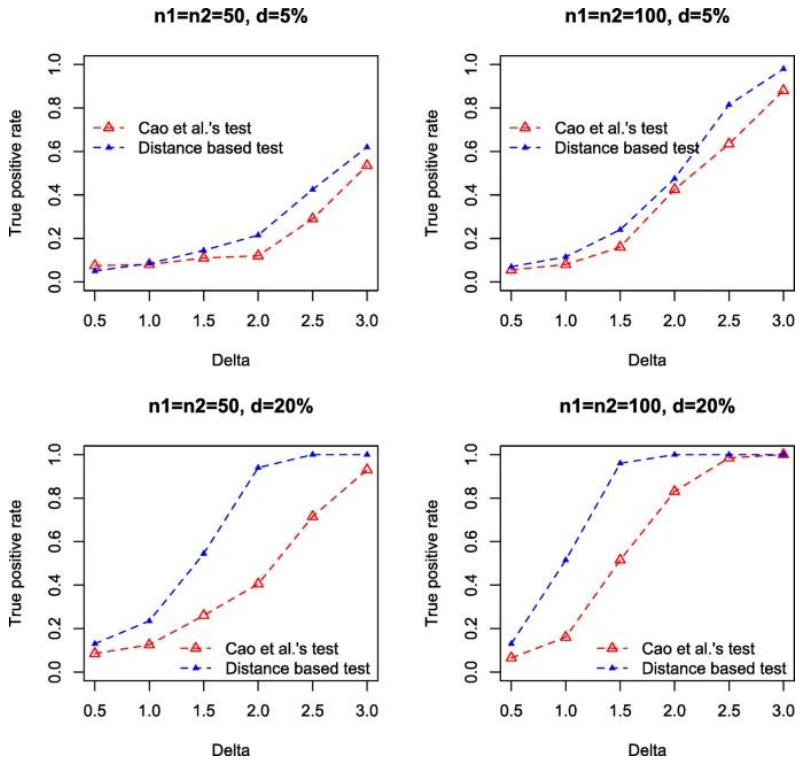


Figure 2.0.1. TPR comparison in setting 1

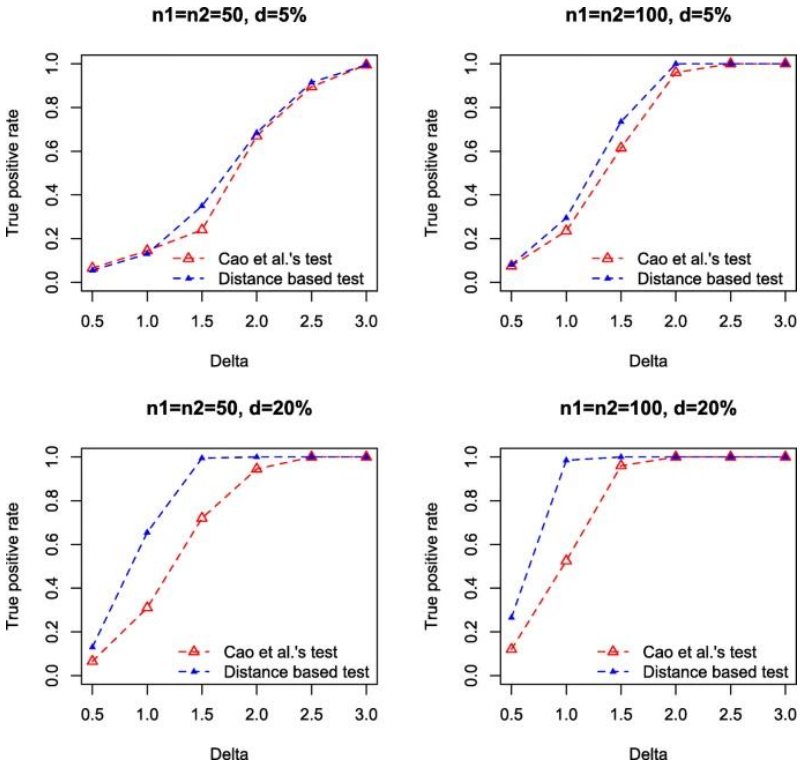


Figure 2.0.2. TPR comparison in setting 2

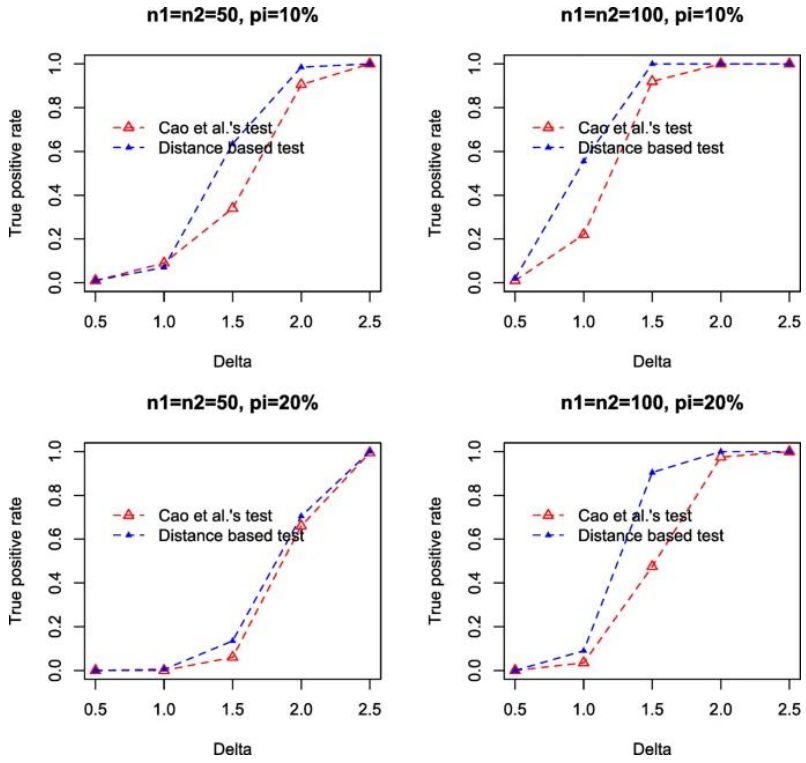


Figure 2.0.3. TPR comparison in setting 3

It can be seen that our distance based-test consistently outperforms the log-ratio based method in all above settings. Particularly, in the dense setting (d=20%), our test achieves substantially higher TPR than the log-ratio test. For instance, in Setting 1, when Δ=2.0, n_1=n_2=50, our test achieves a high TPR of 0.97 while the TPR by log-ratio test is only 0.41. However, when Δ is subtle, e.g., Δ=0.50, both tests fail to detect the difference, even for relatively large sample size, e.g., n_1=n_2=100.

In the second simulation study, we change the dimension p from 100 to 500. The sample size is fixed at $n_1 = n_2 = 100$. We investigate the effect of dimension on the true positive rate. The abundance $W_{ij}^{(k)}$ is generated from 2 settings:

Setting 4: Same as Setting 1, except for fixed Δ=1.5 and dp =10.

Setting 5: (Negative binomial model with excess zeros): $W_{ij}^{(k)} = 0\ with\ probability\ \pi = 10\%$, $W_{ij}^{(k)} \sim NegBin\left(\mu_j^{(k)}\right), r_j^{(k)}$) $with\ probability\ 1-\pi$. Other settings remain the same as in Setting 4.

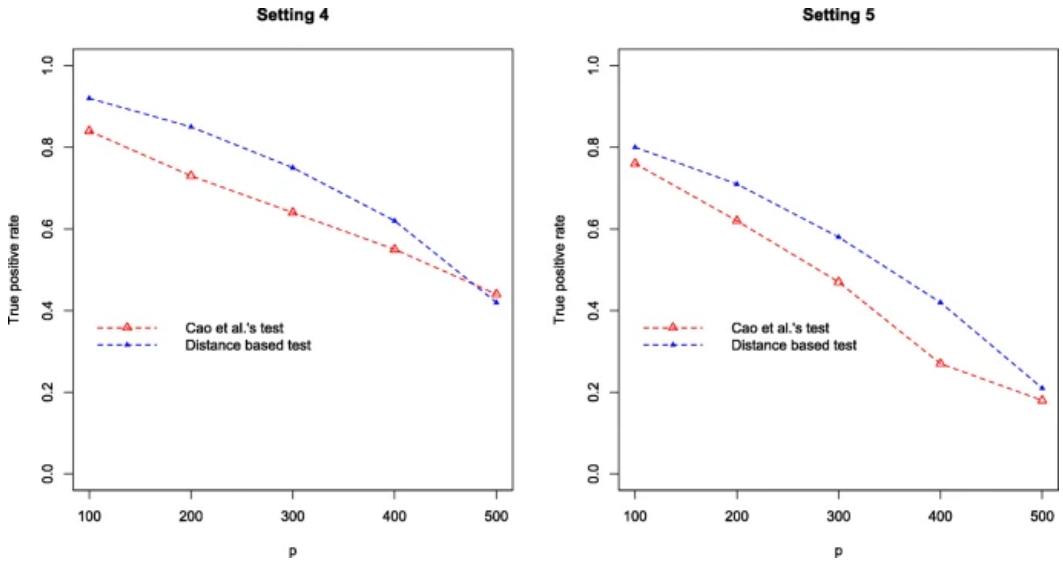


Figure 2.0.4. TPR comparison in setting 4 and 5

Figure 2.0.4 shows that the distance-based test outperforms the log-ratio test especially when the dimension is relatively low. When the dimension is high, for instance *p*=500, the two tests are comparable. More importantly, there is a substantial decrease of TPR as *p* increases, indicating that a feature screening could improve the test performance when *p* is large.

In the third study, we consider testing the compositional difference between multiple groups. We set *K*=4 with sample sizes $n_1 = n_2 = n_3 = n_4 =$ 50, fixed dimension p at 200. The abundance $W_{ij}^{(k)}$ are generated from the negative binomial model with excess zeros. testing the compositional difference between multiple groups.

The abundance $W_{ij}^{(k)}$ are generated with excess zeros $\pi = P(W_{ij}^{(k)}=0)$, with probability 1- π. $W_{ij}^{(k)} \sim \text{NegBin}\left(\mu_j^{(k)}, r_j^{(k)}\right)$, for $i = 1, \ldots, n_i,\ j = 1, \ldots, p,\ r_j^{(1)} \sim \text{Unif}(0.1,1)$, $r_j^{(3)} \sim \text{Unif}(0.1,1),\ r_j^{(1)} = r_j^{(2)}, r_j^{(4)} = r_j^{(3)}$, and $\mu_j^{(1)} \sim \text{Unif}(10,15)$, and $\mu_j^{(3)} \sim \text{Unif}(10,15)$. Let $I = \{I_+, I_-\}$ be the set of taxa with different abundances in two conditions, $\mu_j^{(2)} = \mu_j^{(1)} + \Delta \text{ and } \mu_j^{(4)} = \mu_j^{(3)} + \Delta \text{ for } j \in I_+ \text{ and } \mu_j^{(2)} = \mu_j^{(1)} - \Delta \text{ and} \mu_j^{(4)} = \mu_j^{(3)} - \Delta \text{for } j \in I_-,\ \mu_j^{(4)} = \mu_j^{(3)} = \mu_j^{(2)} = \mu_j^{(1)}$ for $j \notin I,\ |I_+| = |I_-| = 20$, where $|\cdot|$ represents set cardinality, with given d=10%,20% and Δ={0.5,1.0,1.5,2.0,2.5}.

We calculated *p*-value based on 5,000 permutations for distance based-test and for Cao et al.'s test, *p*-values computed by Gumbel distribution from six pairwise comparisons, and use the smallest *p*-value for decision-making. Figure 2.0.5 summarizes the TPRs by the two tests, where it can be seen that our proposed test performs consistently better than the log-ratio based test. Notably, in the setting π=20% and Δ=2.0, the distance correlation test achieves a TPR of 0.83, compared to the TPR of 0.46 by the log-ratio test.

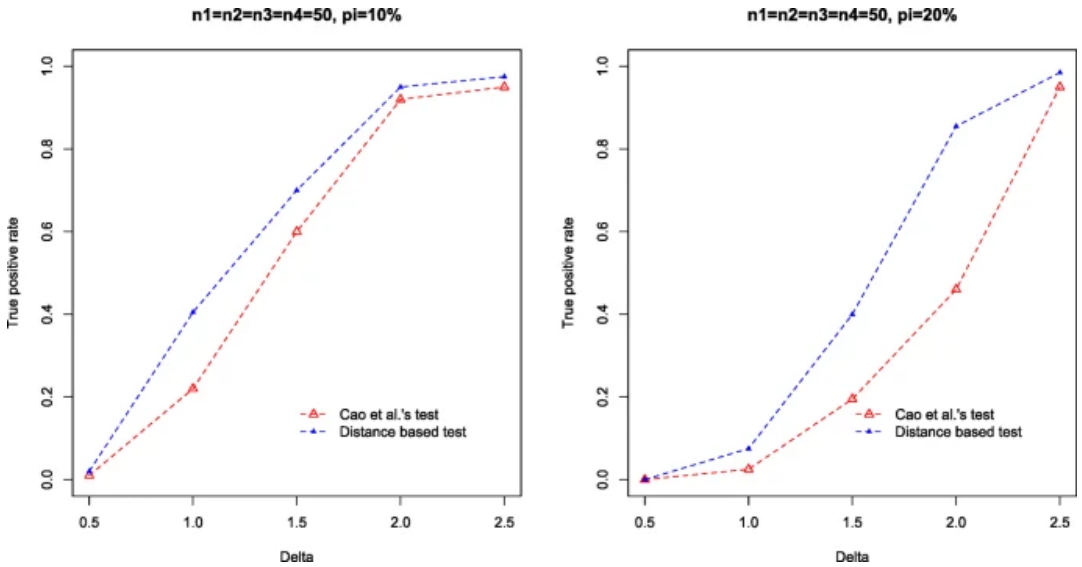


Figure 2.0.5. TPR comparison in the third study

Study the TPRs of the two tests at significant level α=0.05 shows that the distance-based test again outperforms the log-ratio test especially when the dimension is relatively low. When the dimension is high, TPR decrease as p increases, which also improving test performance.

To evaluate the robustness of the proposed method, we conducted a sensitivity analysis for setting 2. For each simulation run, we randomly select 50% of taxa and calculate the p-value using distance correlation test. The empirical true positive rate is summarized in Figure 2.0.6.

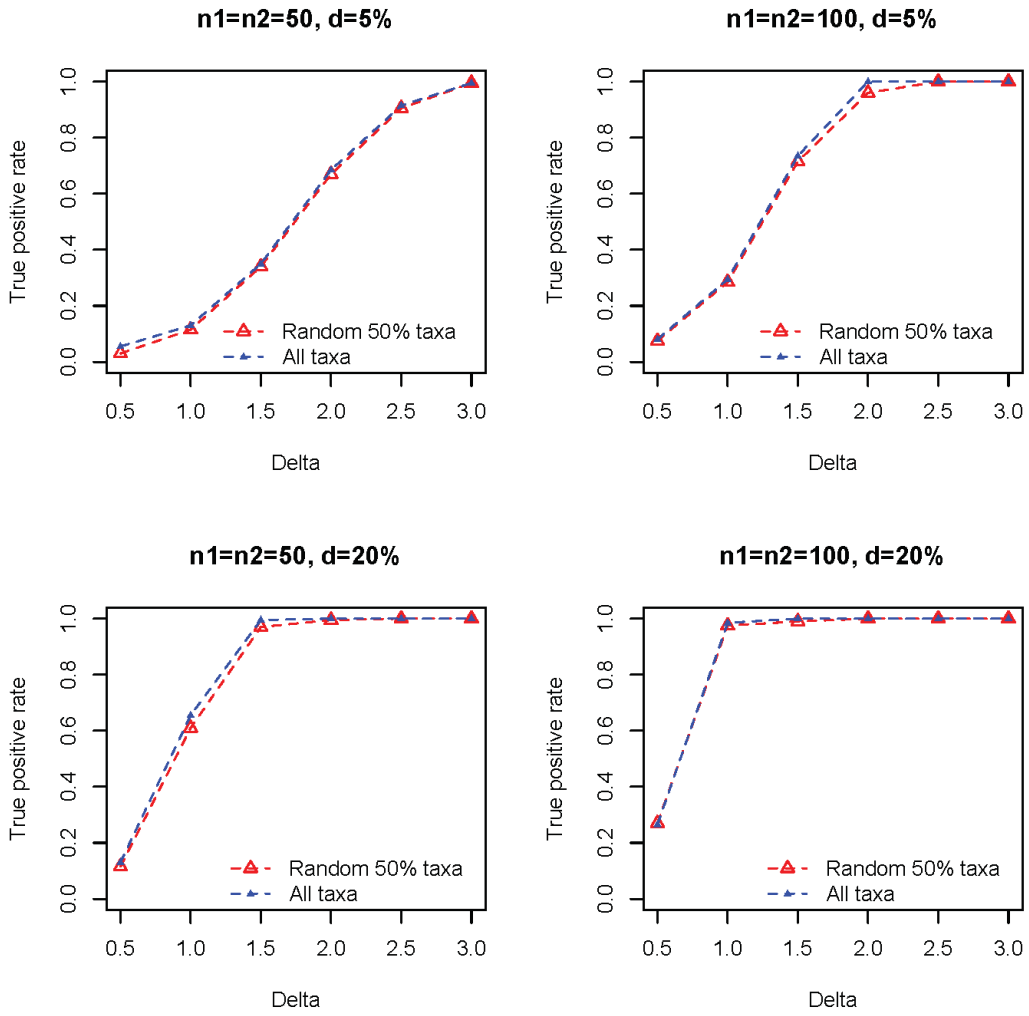


Figure 2.0.6. Sensitivity analysis for setting 2

It can be seen that under all sample sizes and signal densities (d), the empirical true positive rate using randomly selected 50% taxa is comparable to the one with complete variable set, indicating that the distance-based method is fairly robust.

In addition to the simulation study, we have applied this method to analyze two microbiome datasets.

2.4 Applications

2.4.1 Analysis of Throat microbiome data:

Data description: The download data (observed abundance) retrieved from *GUniFrac* R package [19], including two parts which are throat.otu.tab and throat.meta, supporting the study of smoking effect on the upper respiratory tract microbiome. The dataset contains read counts of 856 predefined operational taxonomic units (OTUs, or phylotypes) on 62 samples from the throat microbiome of left body side. There are total 60 subjects (patients) consisting of 32 nonsmokers and 28 smokers.

Applying both tests including the log-ratio based test and distance-based test in the analysis of the throat dataset, we are interested in testing whether there is any significant difference in microbial compositions between smokers and non-smokers. First, we need to clean the data by deleting OTUs with extremely small number of reads (less than 20 reads in total), resulting a final set of 190 OTUs. Next, to perform our distance correlation test, we normalized the abundance W to get the composition for each sample, then calculate the sample proportions $\hat{p}_i$, and the inter-group distances $\widehat{D}_{ij}$ for $i, j = 1, \ldots, K$ (applied Eqs. (2) and (3)). And the last step is to compute the permutation *p*-value based on the distance covariance $d\ \widehat{Cov(X, Y)}$. The results yield a *p*-value of 0.0027, indicating a significant difference between smokers and non-smokers in microbial

composition. However, the test performed by Cao et al.'s gives a *p*-value of 0.098, thus fails to reject the null hypothesis of equal means at the level of 0.05 of significance.

The different results between these two tests alerts the existence of nonlinear effects and over-dispersion, since our methods focus on the distributional difference while the log-ratio test only targets on the means. We also implement additional analyses by taking two examples of bacteria 2434 and bacteria 2831 (see Fig 2.0) to enforce the significant difference in distribution using the centered log-ratios plots; however, Cao et al.'s test gives the insignificant result by the mean difference due to the nonlinear effect and heavy tails, which inflates the variance estimates.

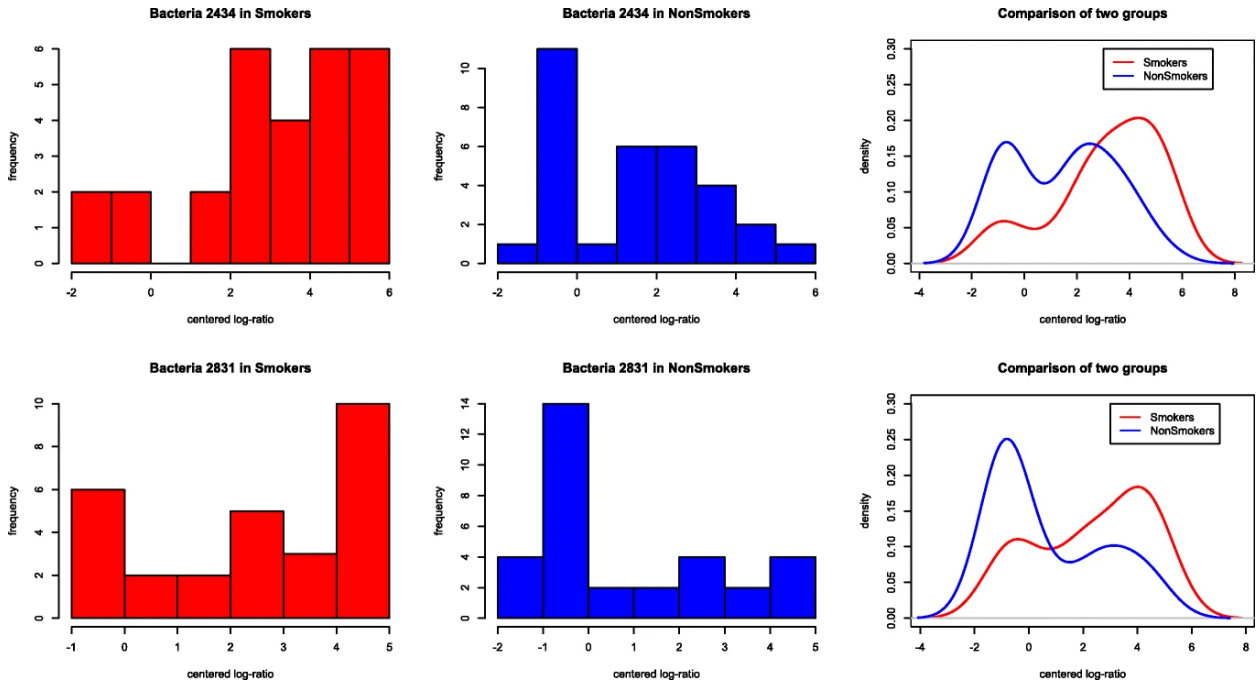


Fig 2.0 Comparison of two groups in bacteria 2434 and bacteria 2831 of throat microbiome data.

The distributions of inter-point distance and 3-minimum spanning tree (3-MST) were also used to compare the difference between smokers and non-smokers. According to Szekely et al. (2007), if two multivariate distributions are identical, the inter-point distances within each group have the

same distribution [17]. The below illustration (Fig 2.1) have shown there is some differences in the inter-point distance distributions of two groups of smokers and non-smokers.

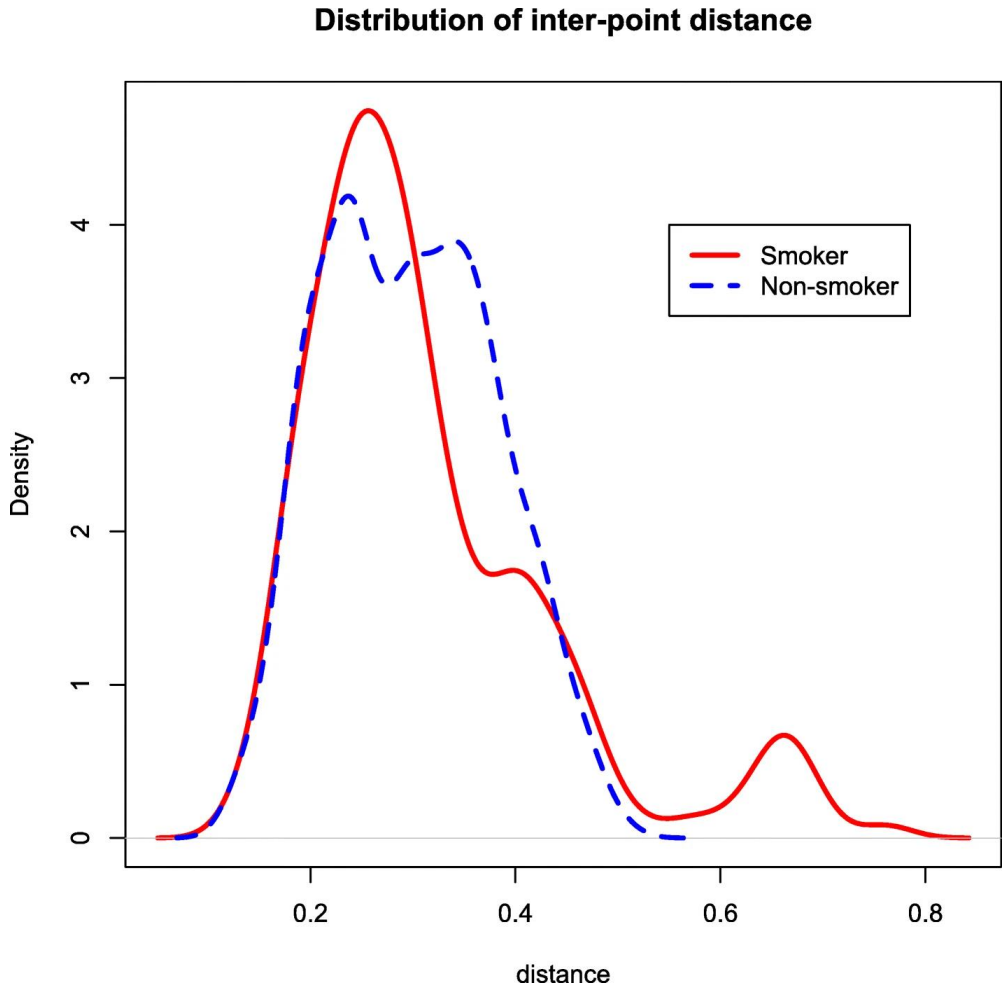


Fig 2.1. Distribution of inter-point distance of throat microbiome

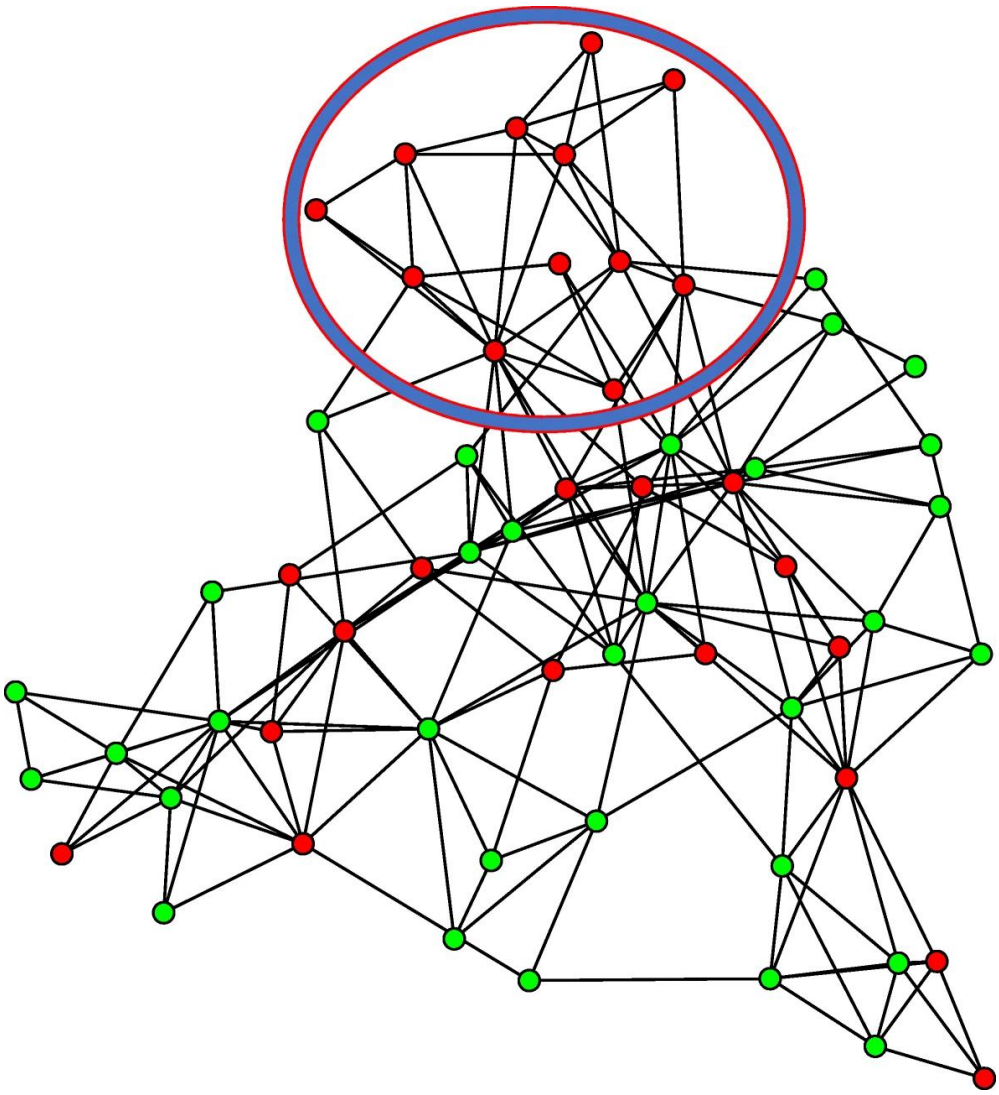


Fig 2.2. 3-MST of throat microbiome

In addition, the tree-based visualization (Fig 2.2) shows a set of 12 smokers (circled) that are highly connected to each other (a connection in the network represents compositional similarity between samples), but with very few connections with non-smokers. If two samples do not show equal chance to connect with any other sample, they do not have the same distribution. Therefore this 3-MST again confirm a distributional difference that found between these two groups.

2.4.2 Analysis of intestinal microbiome data

Data description: The data is collected by Lahti et al. (2014) [20] to study the microbial communities living in the human intestine, which have a big impact on our well-being and health. The data consists of phylogenetic intestinal microbiota of 1,006 western adults from Europe and the United States with altogether 1172 samples and 130 genus-like phylogenetic groups. The clinical data contains many variables including age, sex, nationality, BMI, DNA extraction method etc., and we will take Age as the outcome variable to test our method.

To test whether there is any difference in microbiome composition among different age groups, here we define 3 different age groups: young (<40), middle (41–60) and old (>61), as suggested by Lahti et al. (2014). The distance-based test results in a *p*-value of 3.0×10^{-6}, moreover, the *p*-values from three pairwise comparisons are: 8.2×10^{-5} for young vs middle, 2.2×10^{-5} for young vs old, and 0.081 for middle vs old, indicating a significant difference in microbiome compositions between young and other groups.

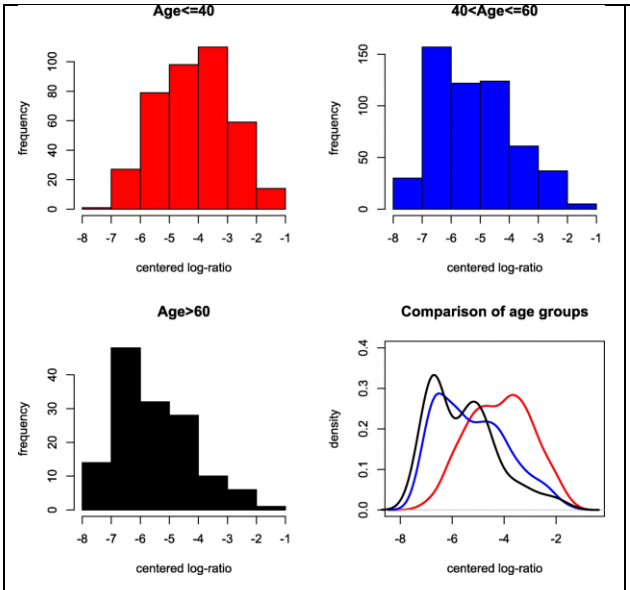


Fig 2.3. Center log-ratio of group 25

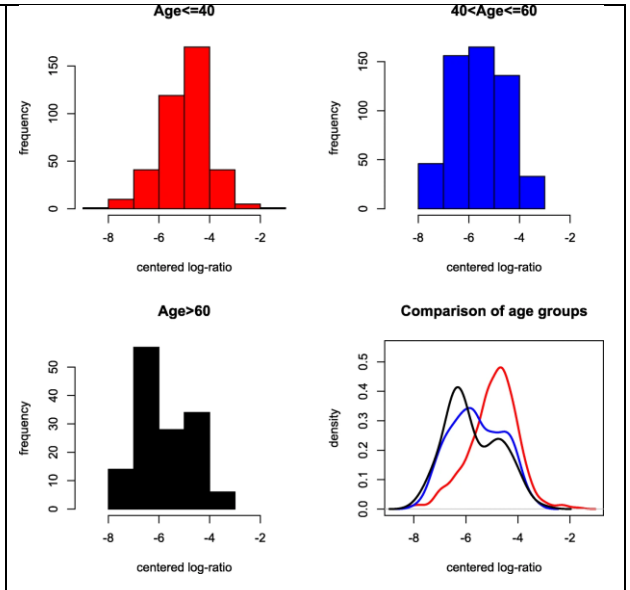


Fig 2.4. Center log-ratio of group 60

Some additional visualizations were performed to confirm our findings. Two examples of group 25 and group 60 are shown in Fig 2.3 and Fig 2.4), where the discrepancy is observed between young and middle/old subjects (see Fig 2.5) from the distribution difference of inter-point distance.

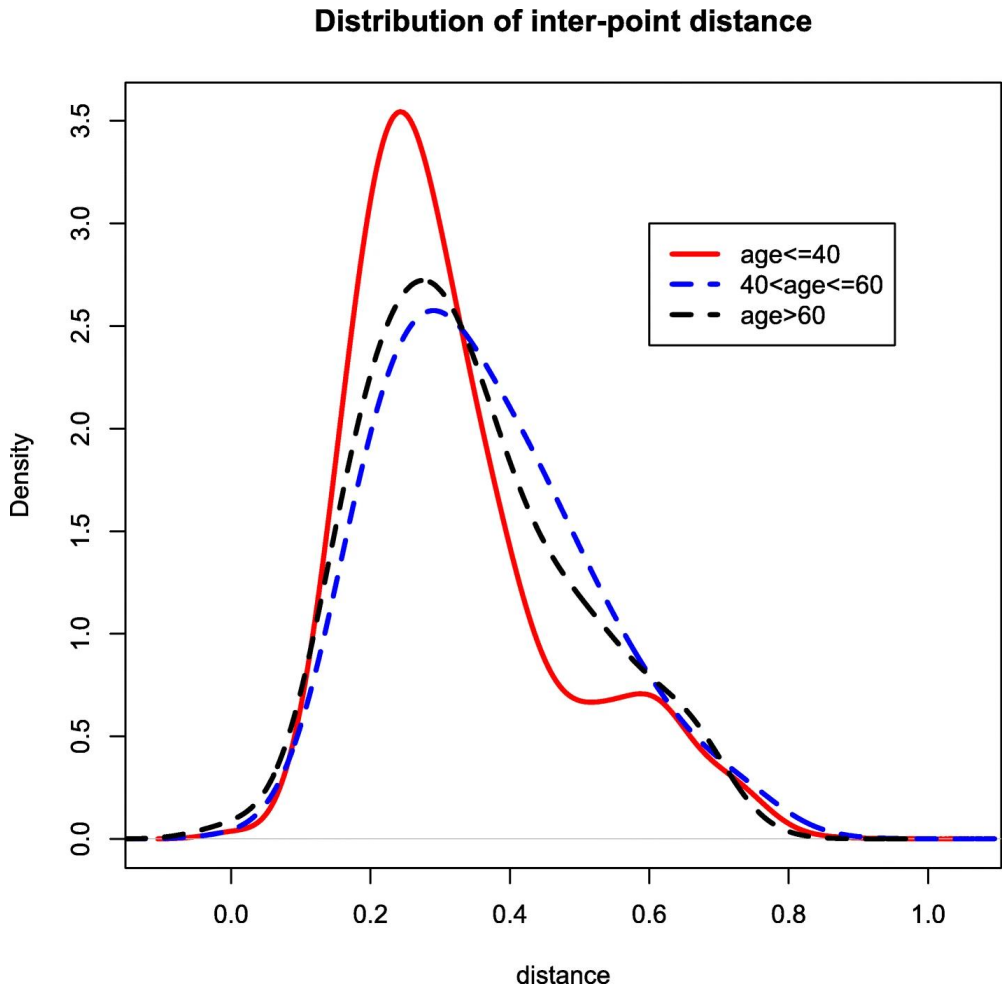


Fig 2.5. Distribution of inter-point distance of intestinal microbiome data.

In summary, our distance-based method is more sensitive to compositional difference. It is easy to implement and computationally efficient, compatible to high dimensional, compositional, over-dispersed or zero-inflated data.

In the next chapter, we will conduct model-based procedure to select the subset of significant variables contributing to the differences between groups. The identifies variables set can be used as a microbiome signature that differentiate phenotypic groups.

Chapter 3

VARIABLES SELECTIONS

Here we present variable selection results from different statistical models. We will define a consensus set of variables as the final set.

3.1 Summary Table:

Our summary tables list the most significant genus in the throat microbiome and the intestinal data. The count tables show the number of genus shared by multiple penalized regressions including Ridge, LASSO, Elastic Net regression and the Multicategories logits models.

3.1.1 Variables selections of intestinal microbiome data

We performed the analysis utilizing the most significant ones among 130 bacteria. We selected the 50 most statistically significant genus from each method Ridge, LASSO, Elastic Net Regression, cumulative Logits and Adjacent-category Logit models.

The cumulative Logits and the Adjacent-category Logit model select significant variable based on their adjusted p-value (< .05). The penalized regressions choose variables based on their beta values, the larger the beta value is, the more significant the genus is.

Table 1. Frequency table of significant intestinal bacteria shared among regression models

	Cumulative Logit	Adjacent-Category Logit	LASSO Regression	Ridge Regression	Elastic Net Regression
Cumulative Logits	50	49	18	27	20
Adjacent-Category Logit	49	50	19	27	21
LASSO Regression	18	19	50	28	41
Ridge Regression	27	27	28	50	32
Elastic Net Regression	20	21	41	32	50

From the most 50 significant bacteria, we generate a count table that summarize the number of genus identified by some of all of the 5 methods. From the above table, we can see that among 50 most significant selected from 5 methods, cumulative Logits and Adjacent-category Logit method agree perfectly well (49/50), LASSO and Elastic Net regression share 41 significant genus, LASSO and Ridge share 27 significant genus, while Cumulative Logits model and LASSO regression have 18 significant bacteria in common.

We also produce the following table that lists all the bacteria with occurrence (5,4,3,2,1) from the count table.

Table 2. Names of significant throat bacteria shared among regression models

5	**4**	**3**	**2**	**1**
V108, V110, V114, V121, V26, V37, V38, V43, V48, V49, V80, V9.	V111, V12, V16, V18, V21, V46, V54, V95.	V100, V102, V11, V115, V119, V120, V126, V128, V13, V131, V2, V22, V27, V3, V30, V31, V32, V41, V42, V50, V60, V61, V64, V68, V76, V88, V93, V96.	V10, V101, V112, V116, V123, V14, V19, V20, V23, V24, V25, V29, V35, V44, V45, V52, V53, V55, V6, V69, V7, V77, V78, V8, V81, V83, V86, V89, V90, V97.	V106, V127, V17, V39, V47, V5, V58, V62, V63, V73, V82, V87, V94, V99.

This table shows that among 50 bacteria, there are 12 having statistically significant effect on the age group difference. There are 8 bacteria that are selected in 4 methods. We then selected the 20 most significant ones based on this table, which appeared in all five or four variables selection methods for the use in Chapter 4.

3.1.2 Variables selections of throat microbiome data

The throat microbiome dataset shows that the smoker and non-smoker status of patient have significant difference in term of throat microbiome.

We utilized four different methods to select variables, including Logit Model, Ridge, LASSO, Elastic Net Regression to find the 50 most significant genus from 190 bacteria.

The Logit model selects significant variable based on their p-value (< .05) and we found 13 bacteria with significance. The penalized regressions evaluate variables based on their beta values, the larger the beta value is, the more significant the genus is.

We use the following table to summarize the number of genus selected by different methods:

Table 3. Frequency table of significant throat bacteria shared among regression models

	Logit Model	LASSO Regression	Ridge Regression	Elastic Net Regression
Logit Model	50	24	23	29
LASSO Regression	24	50	20	36
Ridge Regression	23	20	50	29
Elastic Net Regression	29	36	29	50

As seen from this table, for the top 50 selected bacteria which have smallest p-value or biggest beta value from LASSO and Elastic Net regression, Logit Model shared 24, 23 and 29 significant bacteria with LASSO, Ridge and Enet regression respectively. LASSO and Ridge has the least in common (only 20 bacteria), while LASSO and Elastic Net regression have the most in common (36/50).

We also produce a table that lists all the bacteria with occurrence (4,3,2,1) from the above count table. Table 4 shows there are 14 significant bacteria shared the significant effect on human throat

selected from 3 regressions models. Also, there are 16 bacteria have significant effect on human throat were appeared in all four models, they will be chosen to represent the throat microbiome data for further analysis in Chapter 4.

Table 4. Names of significant throat bacteria shared among regression models

4	**3**	**2**	**1**
X1478 X1490	X1371 X1540	X1024 X1280	X1036 X1204 X154 X1633 X1936
X2382 X3246	X2047 X2132	X1511 X1596	X2082 X2300 X24 X2434 X2572
X3839 X4194	X2705 X2831	X1618 X181	X2621 X2718 X2839 X3026
X4223 X4363	X3954 X4036	X2046 X257	X3105 X3147 X3276 X3418
X44 X4422	X4243 X4912	X2928 X3227	X3427 X3538 X3878 X3943
X4457 X4703	X501 X5129	X332 X3391	X3945 X3957 X3988 X4131
X4707 X5287	X5160 X689	X392 X4321	X4248 X444 X4793 X4813 X4816
X548 X93		X4608 X483	X4871 X4964 X5111 X5273
		X4966 X5045	X5308 X5313 X5394 X5460
		X5414 X5496	X5468 X5583 X58 X618 X625
		X5563 X5661	X667 X760 X772 X898
		X990	

3.2 Venn Diagram:

In addition to the summary tables, we have produced the Venn to illustrate the agreement between different variable selection methods.

3.2.1 Venn diagrams of selected variables of intestinal microbiome dataset

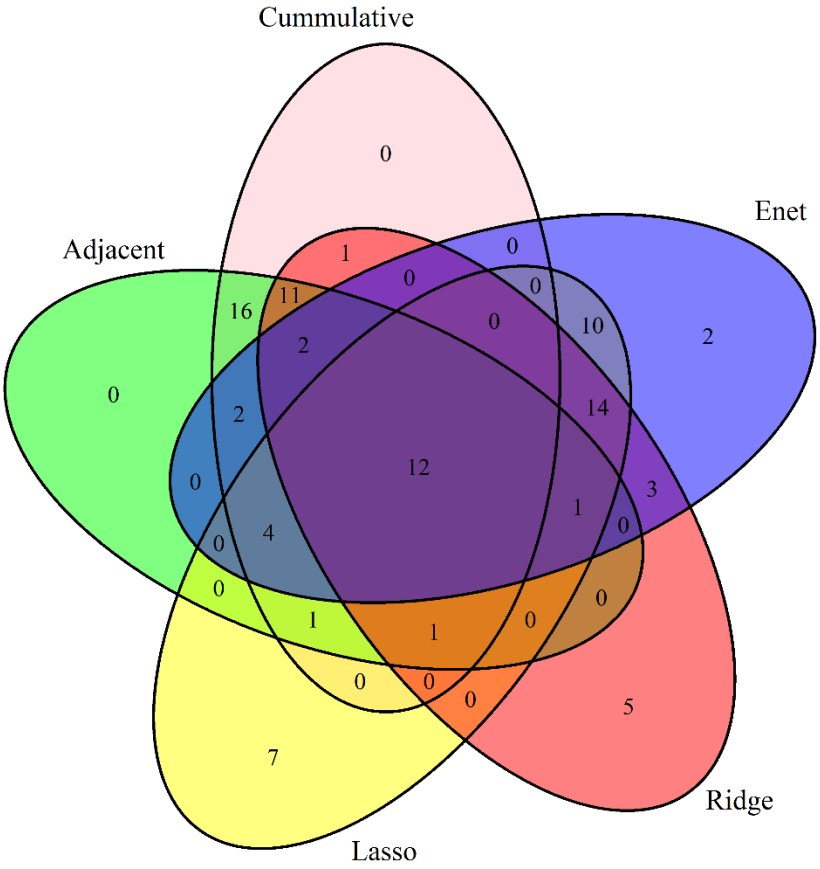


Fig 3.1. Venn Diagram of 5 methods for intestinal dataset (50 genus)

The Fig 3.1. is a Venn diagram shows all the possible relationship among 5 different variables selection methods applied for the intestinal microbiome dataset. There are 12 significant bacteria in the difference among three different age groups (young < 40 years, 40<middle<60, old >60 years) found in all five alternatives. We can also read that there are 4 bacteria found significant by all four methods (except for Ridge regressions).

The below pair Venn diagram illustrate a very nice result from Cumulative and Adjacent-category Logit models, that share perfectly 49 out of 50 most significant bacteria in common.

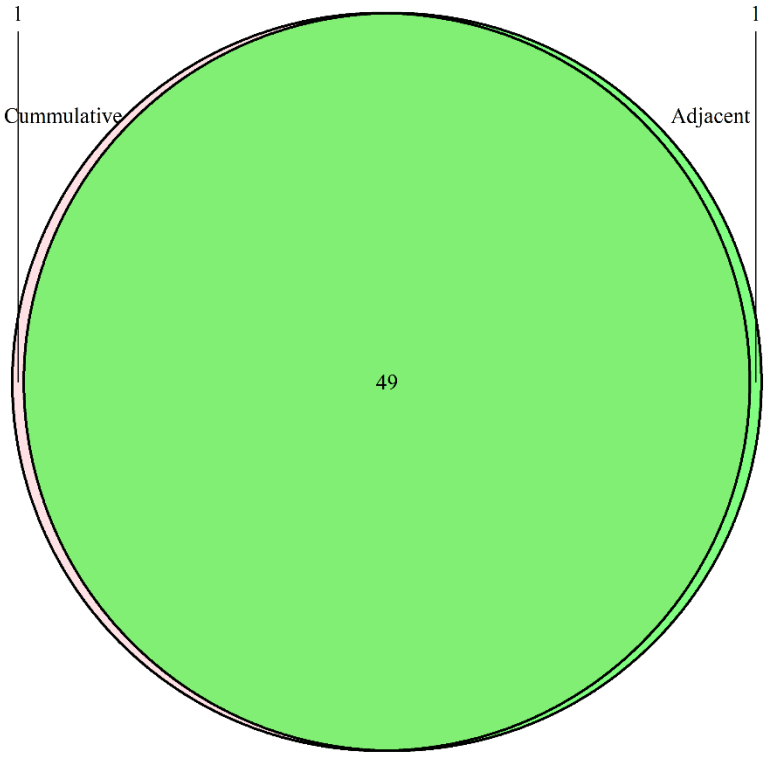


Fig 3.2. Pair Venn Diagram of Cumulative Logits and Adjacent-category Logits for intestinal dataset (50 genus)

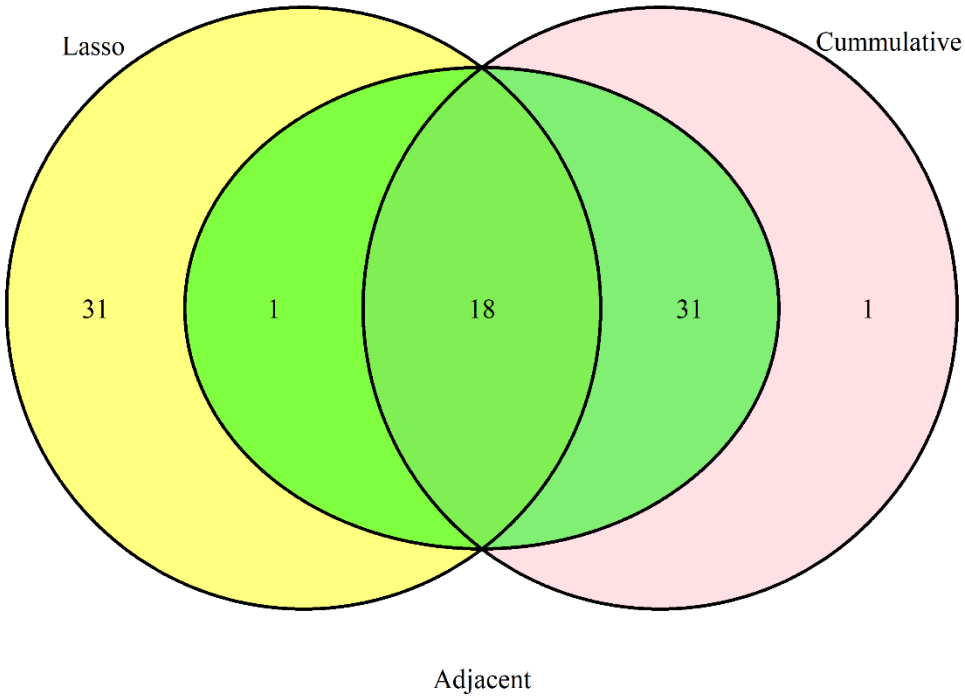


Fig 3.3. Tri Venn Diagram of Cumulative Logits, Adjacent-category Logits and LASSO regression for intestinal dataset (50 genus)

The TriVenn diagram above shows that there are 18 common bacteria found in three different methods, including LASSO regression, Cumulative and Adjacent-category Logits. There is only 1 bacteria is not in common as significance found by Cumulative and Adjacent-category Logit methods, which made them perfectly well sharing the same bacteria with significant effect by different age groups.

Another TriVenn diagram (see Fig 3.4.) illustrate the three penalized regressions have share 27 bacteria in common in contributing in making the difference in three groups of age of intestinal microbiome data.

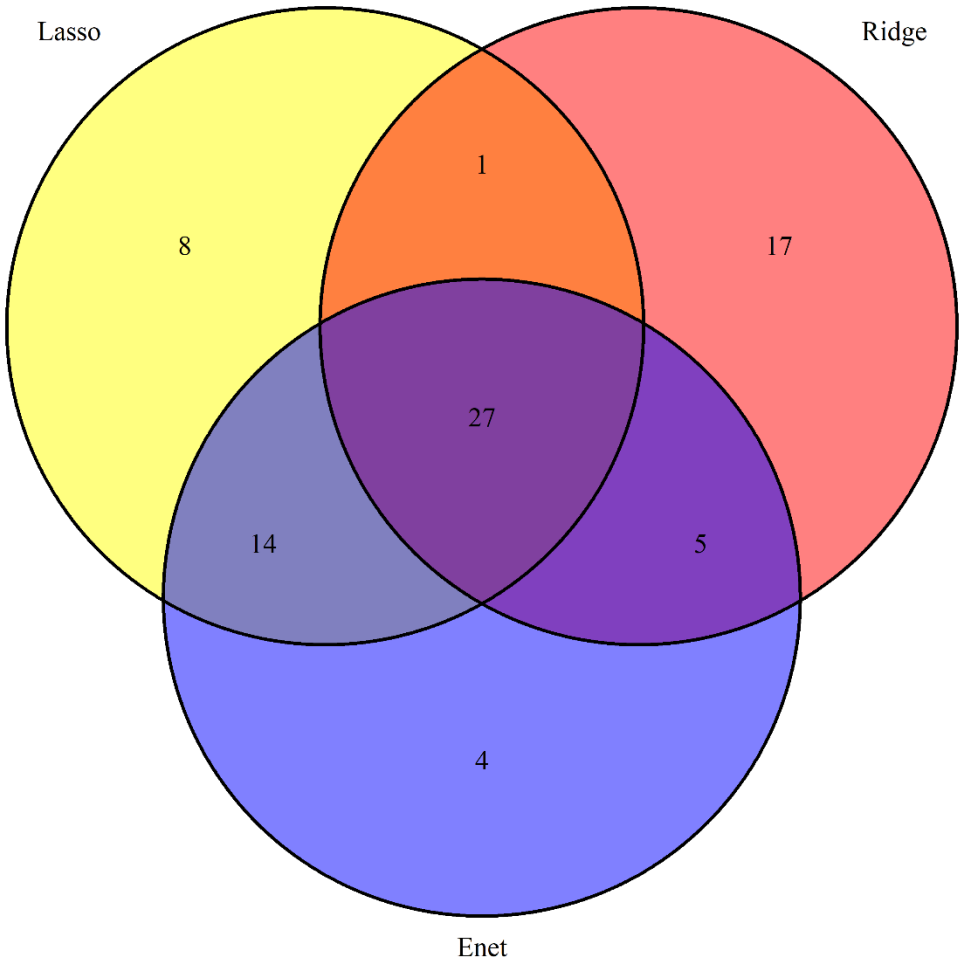


Fig 3.4. Tri Venn Diagram of 3 penalized regressions for intestinal dataset (50 genus)

3.2.2 Venn diagrams of selected variables of throat microbiome dataset

The Venn diagram of all four methods used in selecting variables which have significant effect by smoking status in throat microbiome data is shown in Fig 3.5. There are 16 bacteria are found significant in all 4 methods among the selected 50 genus from each method.

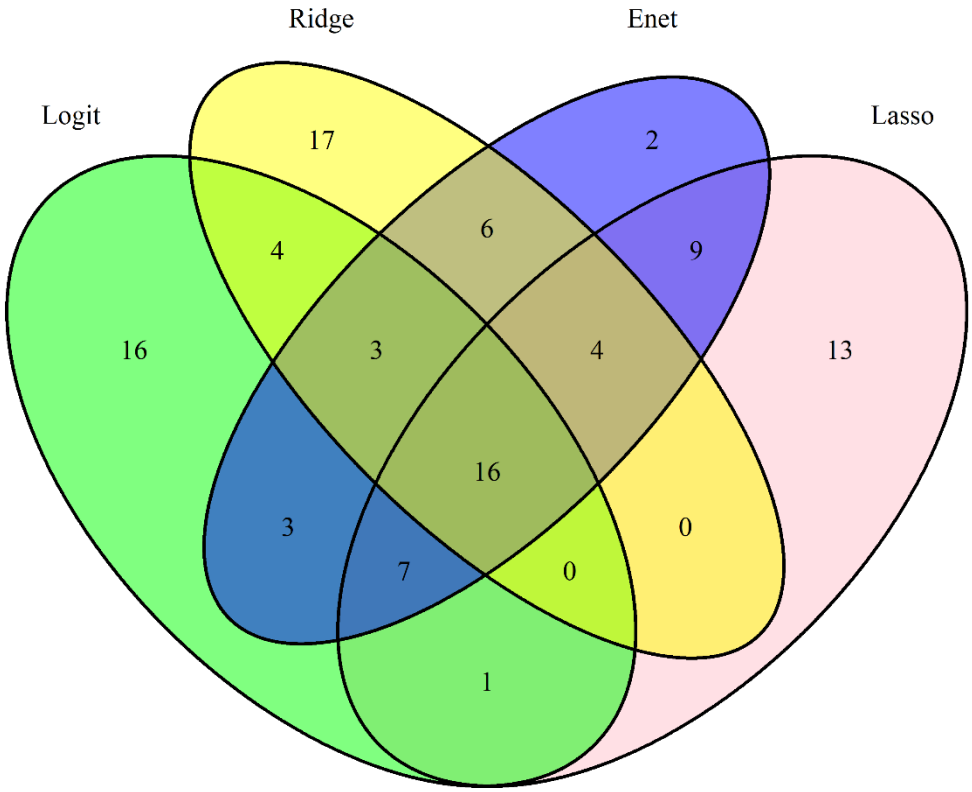


Fig 3.5. Venn Diagram of 4 methods applied for throat microbiome data (50 bacteria)

Figure 3.5 shows that there are 4 bacteria shared the role of significance from 3 regression models, including LASSO, Ridge and Elastic Net. This is magnified in the next illustration Fig 3.6.

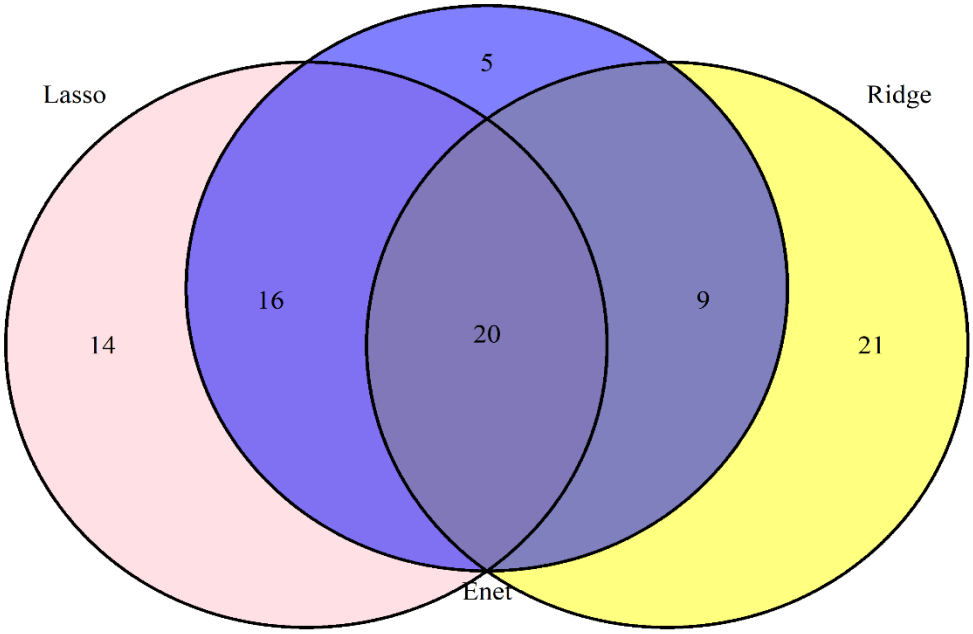


Fig 3.6. Venn Diagram of 3 penalized regression for throat microbiome data (50 bacteria)

The tri-Venn diagram in Fig 3.6 shows 20 common significant bacteria shared by all 3 penalized regressions methods; while there are 16 bacteria shared solely between LASSO and Elastic Net, which made a total of 36 significant bacteria found by both these two regression methods.

The next illustration is pair diagram showing that Logit and LASSO have 23 most significant bacteria being affected by smoking status in the throat microbiome data, which is the least common comparing to other pair comparisons between Logit model with Ridge and Elastic net regressions.

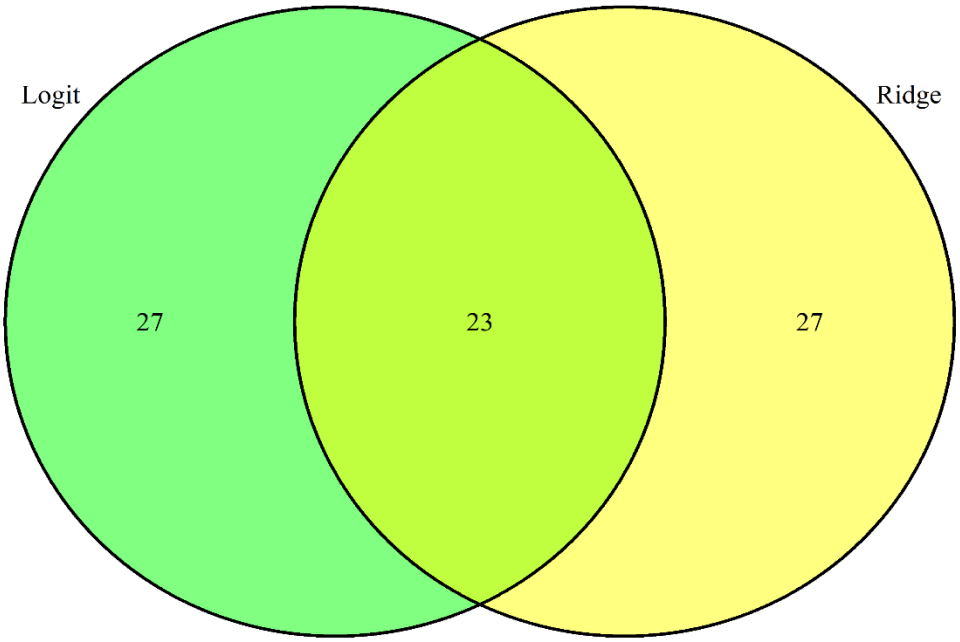


Fig 3.7. Pair Venn Diagram of Logits model and Ridge regression for throat microbiome data (50 bacteria)

In the next chapter, we will illustrate these selected variables are important in making the significant effect on different smoking status in throat microbiome data or among multiple age groups in intestinal microbiome data.

Chapter 4

VISUALIZATION & VALIDATION OF THE RESULTS FROM CHAPTER 3

To illustrate the difference among groups in terms of microbiome composition, we use some routine visualization including boxplots and density plots, as well as a dimensional reduction technique, namely, the multi-dimentional scaling (MDS) plot which is briefly introduced below.

4.1 Multi-dimentional Scaling (MDS) plot

MDS plot is a widely used visualization method to display separation of multiple groups in a multi-dimensional space. MDS arranges the points into an abstract Cartesian space, where the points are computed based on the distance matrix. First, distances between points on the plot approximates their multivariate dissimilarity as closely as possible. In the context of compositional data, the points located closely (or as a cluster) on the MDS plot indicates samples that have similar species composition.

Like other popular dimension reduction methods or visualizations such as Principal Component Analysis (PCA), MDS takes the matrix of pairwise dissimilarities between samples of input and outputs a coordinate matrix whose configuration minimizes a pre-defined loss function.

To be specific, a collection of objects (patients, bacteria, etc) on which a distance function (dissimilarity matrix) is defined as

$$D \coloneqq \left(\begin{bmatrix} d_{1,1} & \cdots & d_{1,M} \\ \vdots & \ddots & \vdots \\ d_{M,1} & \cdots & d_{M,M} \end{bmatrix} \right)$$

where $d_{i,j}$ is distance between i-th and j-th objects.

The goal of MDS is formulated as an optimization problem, to find M vectors $x_1, x_2, \dots x_M \in \mathbb{R}^N$ such that $\left|\left|x_i - x_j\right|\right| \sim d_{i,j}\ for\ all\ i,j \in (1, \dots, M)$. The solution of this problem

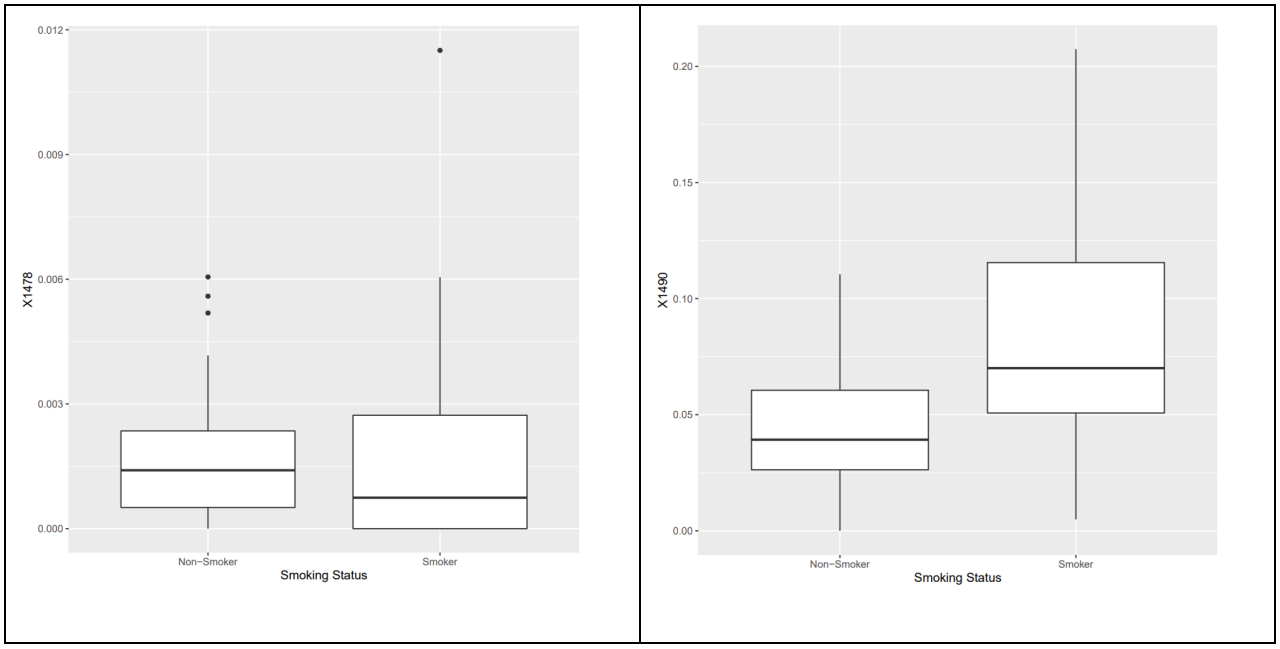
$$\min_{x_1, x_2, \ldots x_M} \sum_{i<j} \left(\left| |x_i - x_j| - d_{i,j} \right| \right)^2$$

is then found by numerical optimization techniques.

This technique is then applied in both throat and intestinal microbiome datasets.

4.2 Throat Microbiome Data

Fig 4.1 are examples of boxplots of bacteria X1478, X1490, X4703, X3246.



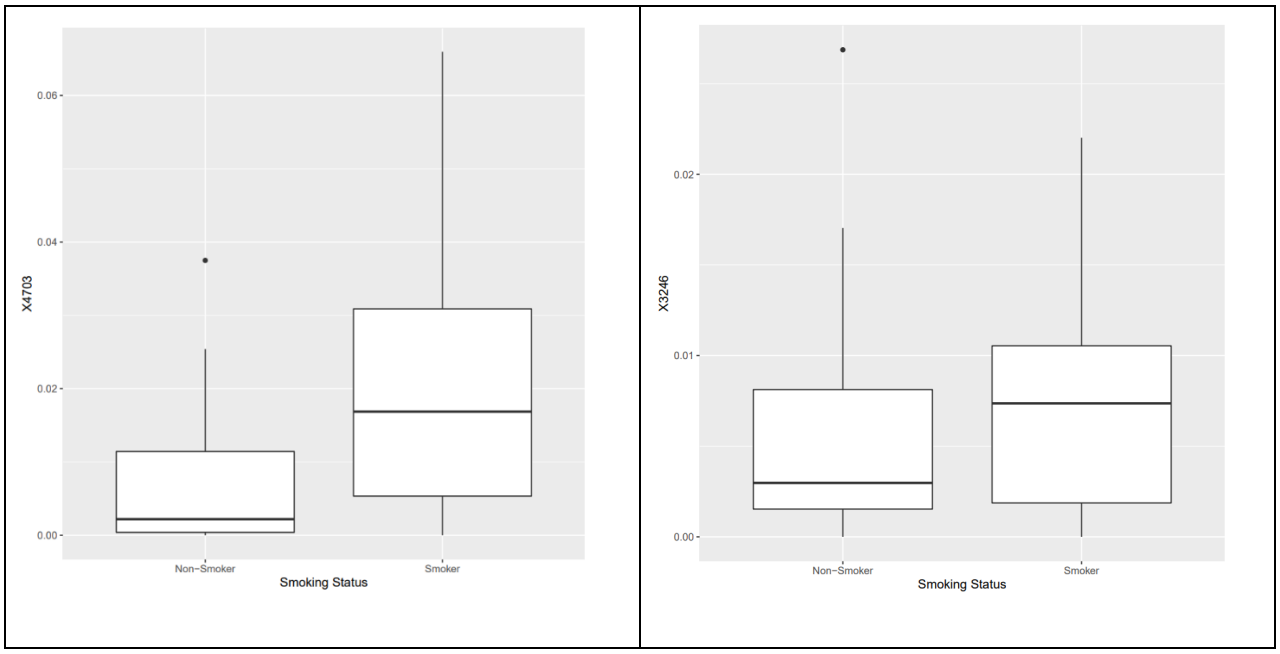


Fig 4.1. Boxplots of some significant bacteria of throat microbiome dataset

By looking at these above boxplots, we have seen the quartiles of the Smoker groups in bacteria X1478 is slightly different; bacteria X1490, X4703 and X3246 show obvious difference of the quartiles between Smokers and Non-Smokers. These depict the effect of the smoking status in throat microbiome.

Next, to emphasize the effect of these bacteria on the age group, I also conducted the density plots of bacteria X1478, X1490, X4703 and X3246, chosen from the most 16 significant bacteria of the throat microbiome data selected from 4 selection methods in chapter three.

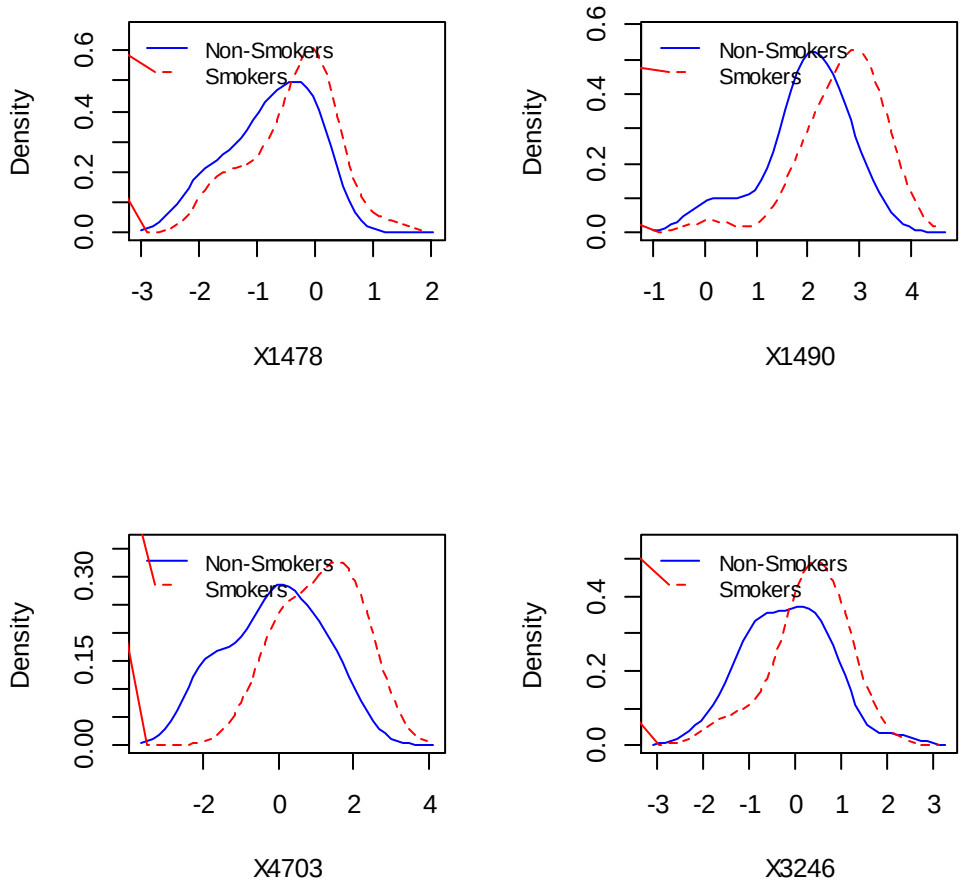


Fig 4.2. Density plots of some significant bacteria of throat microbiome dataset

In fig 4.2, the red and blue curves represent the density distribution of Smokers and Non-Smoker of these defined throat bacteria. They depict the different shapes between Smokers and Non-Smokers, emphasize the difference among these two groups, which tells us smoking status has statistically significant effect on the throat microbiome.

The Multi-dimentional scaling plot is used for this dataset to visualize the relationship between Smokers and Non-Smokers in throat microbiome dataset. Each dot represents an individual participated in the throat data study, there are 60 individuals which include 32 Smokers (red dots) and 28 Non-Smokers (blue dots).

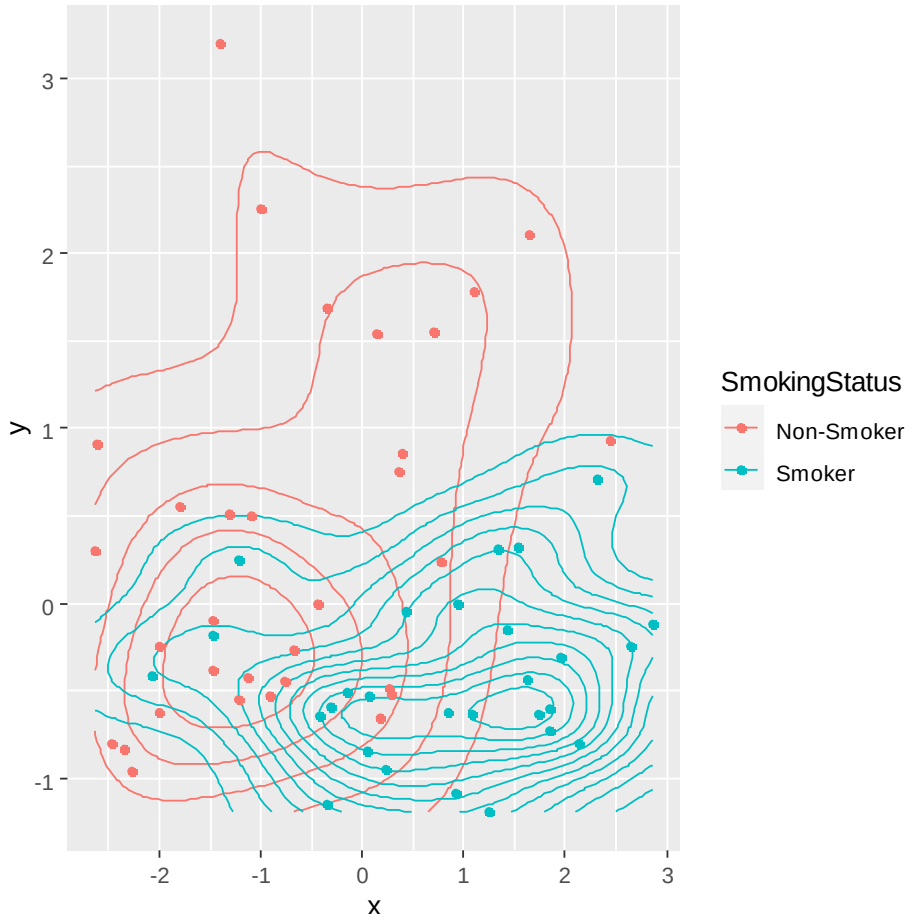


Fig 4.3. MDS plot of throat microbiome dataset

In the Fig 4.3, there are red and blue contours separating two clusters forming from the Smoker and Non-Smoker groups, indicating that Smoking status have significant effect on throat disease. The x-axis and y-axis are optimized values calculated from the dissimilarity function of MDS method as described above.

4.3. Intestinal Microbiome Data

Different from analysis of the throat microbiome data where we studied the difference between two groups of smoking status, the intestinal microbiome data are applied with the study among three age groups.

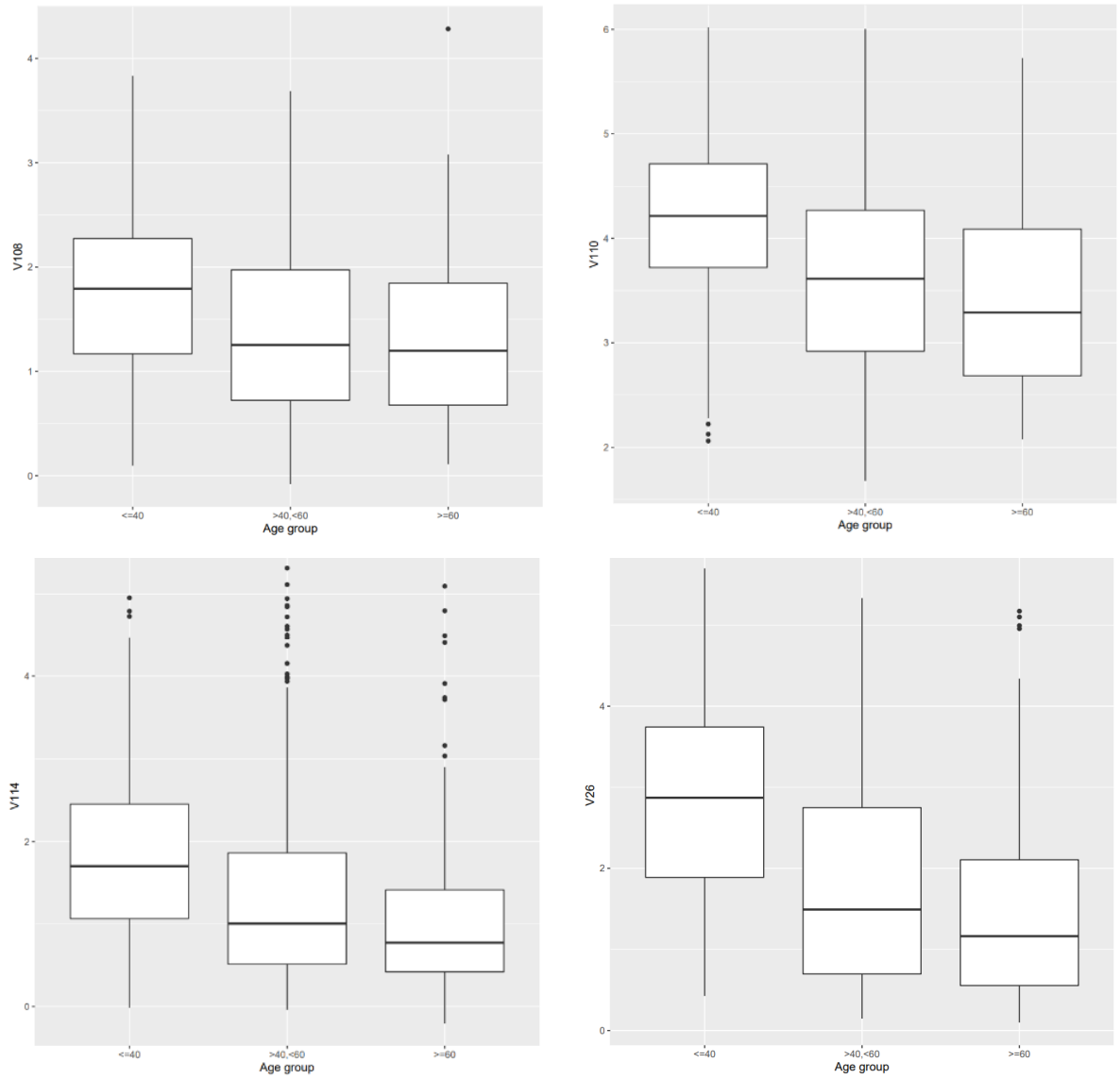


Fig 4.4. Boxplots of some significant bacteria of intestinal microbiome dataset

Above are examples of the bacteria V108, V110, V114 and V261 among the most significant bacteria in intestinal microbiome dataset, which are selected from five different variable selection methods described in chapter three. Fig 4.4 are the boxplots of these bacteria, show that the young age group has the median is higher than those of the middle and old age groups, which indicate these bacteria are significant in making effect among age groups.

We also plot the density distribution for three age groups of the bacteria V108, V110, V114 and V26, among the twenty most significant bacteria of the intestinal dataset selected from five different variable selection methods described in chapter three in order to see their effect on age.

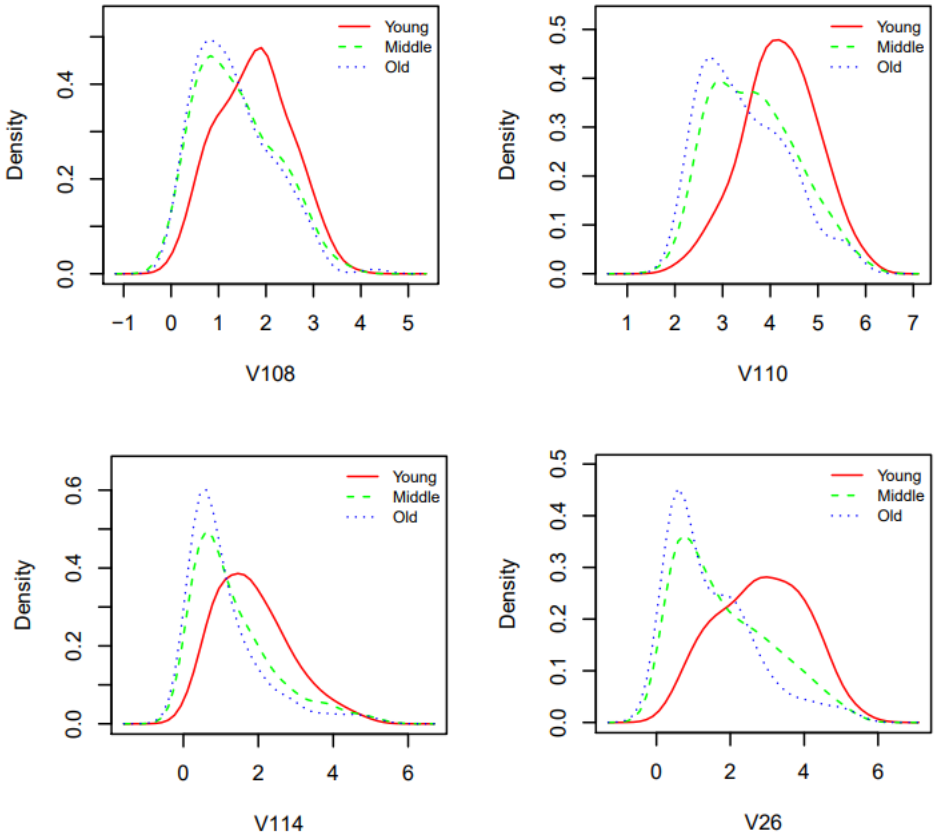


Fig 4.5. Density plots of some significant bacteria of the intestinal microbiome dataset

In figure 4.5, three color lines (red, green and blue) represent the distribution of the young, middle and old age group respectively. We could have seen that the red line shape is obviously to the right of the other two lines, indicating young age group is different than the middle and old age group. In other words, age have significant effect on these important intestinal bacteria.

Similarly, we conducted MDS plot for intestinal microbiome dataset to visualize the difference among three different age groups.

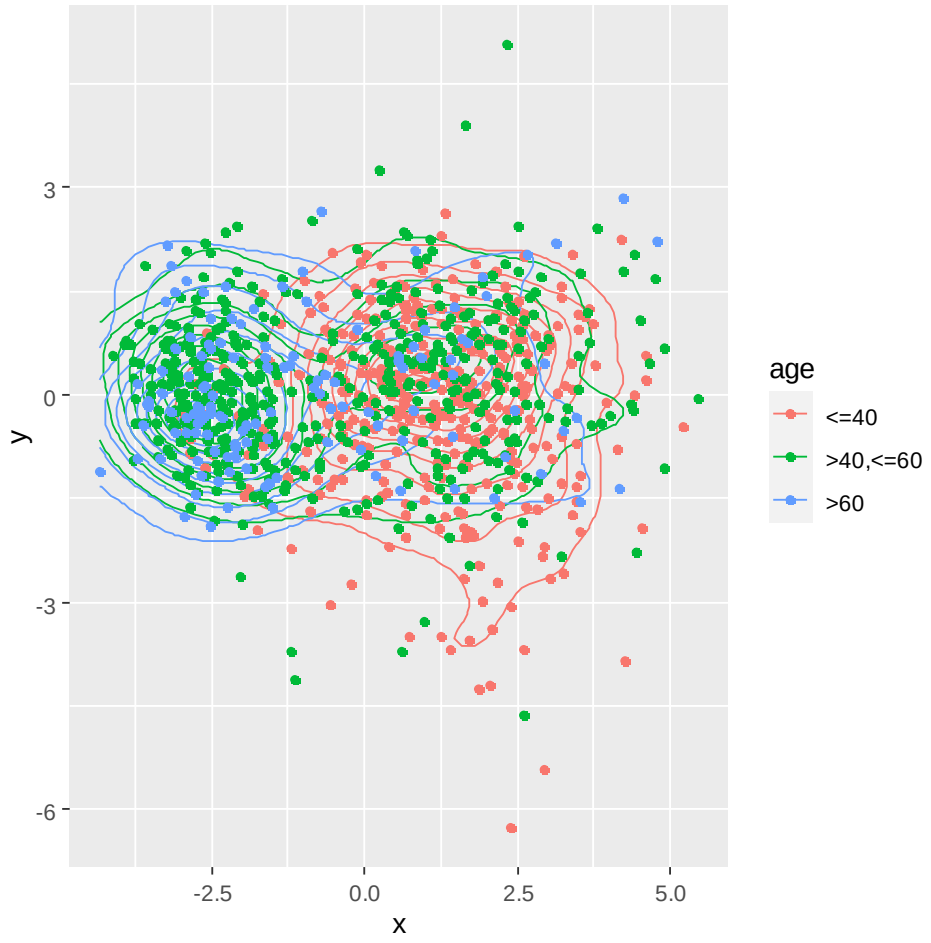


Fig 4.6. MDS plot of intestinal microbiome dataset

In the Fig 4.6, there are three colors red, green and blue represent three different age groups young, middle and old age. There are 950 individuals shown as dots. The three different colored contours formed clusters of those three age groups.

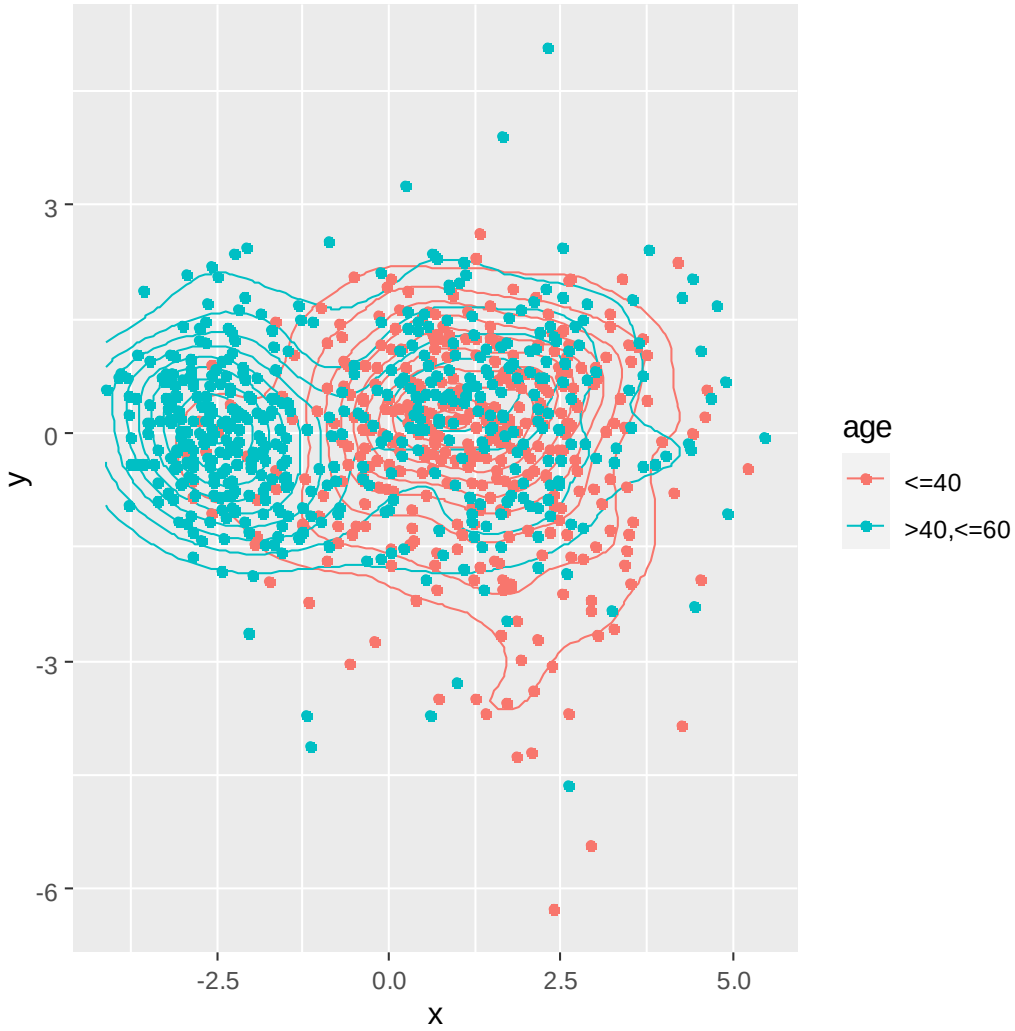


Fig 4.7. MDS plot of two age groups (young vs. middle) in intestinal microbiome dataset

As shown in the previous part through the statistical methods in chapter 2, interpoint distance plot in chapter 3, as well as the boxplots and density plots, the young group is obviously significant different than the middle age group and the old age group. In addition to that, we have created pair comparisons for the young versus middle group (see Fig 4.7) and young versus old age group (see Fig 4.8) utilizing multidimensional scaling.

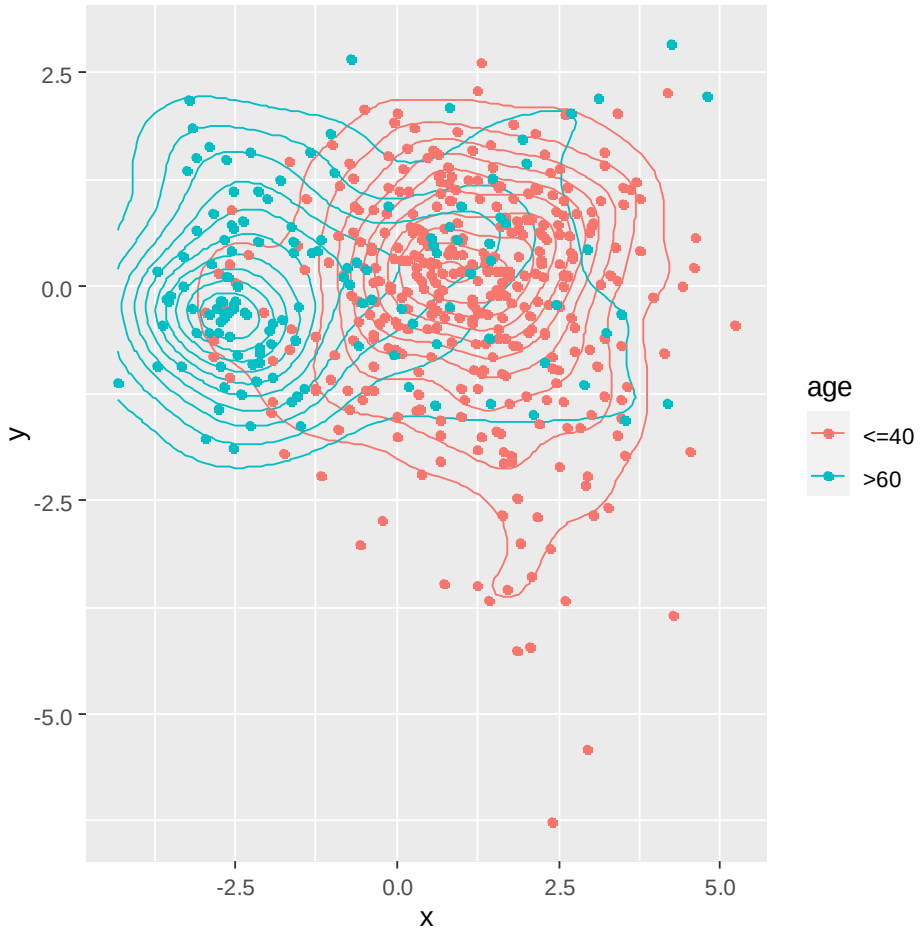


Fig 4.8. MDS plot of two age groups (young vs. old) in the intestinal microbiome dataset

In the two above MDS plots, could have seen that clearly there are difference between red and cyan contours forming clusters for the difference two age groups in these two MDS plots, showing that age have significant effect in the intestinal microbiome dataset, especially the young age group is more different than the others.

However, there is not statistically significant difference between middle age group and old age group, which is illustrated in Fig 4.9. The red and cyan contours in this plot seem mixing together and do not separate from each other.

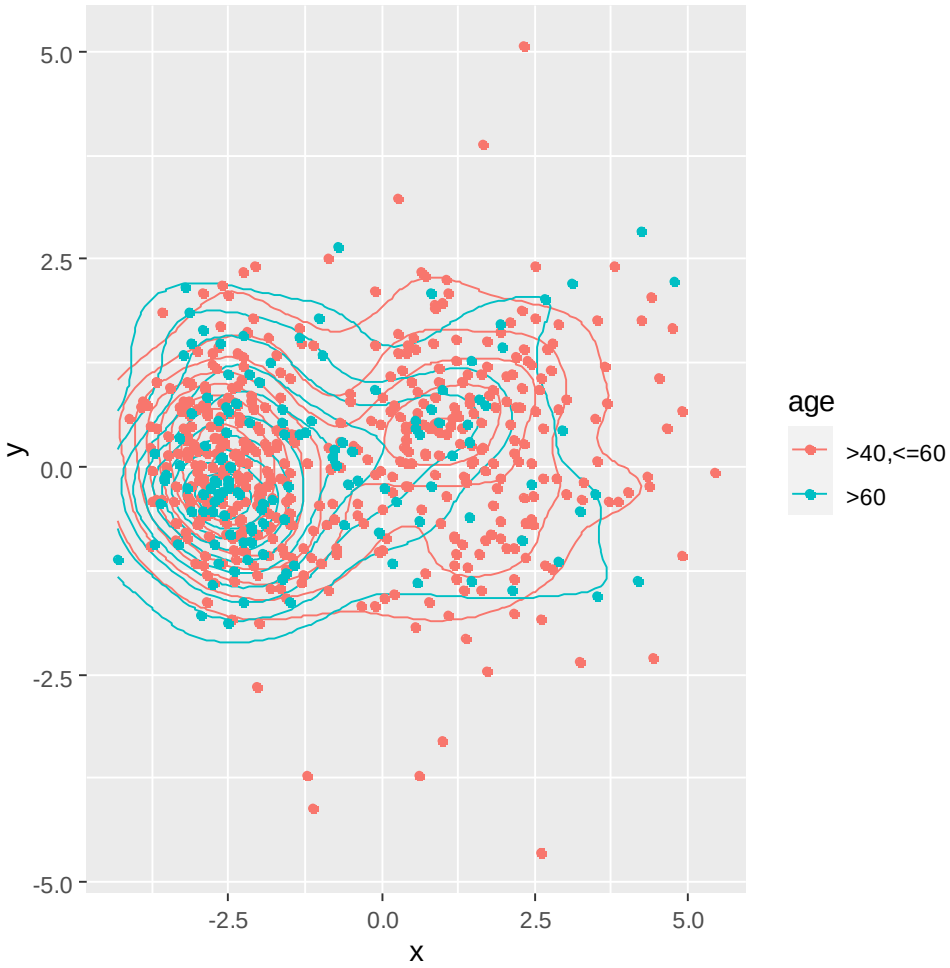


Fig 4.9. MDS plot of two age groups (middle vs. old) in the intestinal microbiome dataset

Chapter 5

CONCLUSION AND DISCUSSION

5.1 Conclusion

In this dissertation, we formulate a Bayesian testing framework to identify the compositional differences between multiple populations. In addition, we propose to use the distance correlation measure to test the null hypothesis. Simulation studies and two real applications in the human microbiome demonstrate that our test is more sensitive to the compositional difference than the mean-based method, especially when the data are over-dispersed or zero-inflated. The proposed test is easy to implement and computationally efficient, facilitating its application to large-scale datasets. Moreover, we conducted variable selection to select the most significant variables using multiple methods such as multi-category logistic models and penalized regression models. We illustrated the significance difference using various visualizations including density plots and the multidimensional scaling plots.

As a whole, we put together a simple but powerful statistical framework to test the compositional difference between multiple populations and identify a small subset of taxa that drive the separation of different populations.

5.2 Discussion

Microbiome data are often compositional, high-dimensional and over-dispersed, which poses great challenges to the statistical analysis. To overcome these obstacles, we have formulated a new testable hypothesis from a Bayesian point of view and suggested a nonparametric test to detect the compositional difference between multiple populations. Compared to the existing tests, our method has several advantages. First, the distance-based test is free of parametric assumptions but directly targets the distributional difference, therefore it is capable of detecting nonlinear effects.

The application in throat microbiome provided a good example, where the new test successfully captured the difference between two phenotypes, while the mean based test failed to do so. In addition, our method can deal with multiple groups, while most of existing methods are only for two-group comparison. Third, our test does not require sparsity assumption on the mean differences as in Cao et al.'s test, and in our simulation study, the new test worked quite well against both sparse and relatively dense alternatives.

There are several possible extensions of the proposed test. First, the distance based method can be readily extended to ordinal phenotypes (or conditions), although we have been using nominal phenotypes for illustrative purpose. For ordinal phenotype, $Y \in \{1,2,...,K\}$, where there is a natural ordering $1<2...<K$, (e.g., {mild, moderate, severe} for severity of a disease, {I, II, III, IV} for cancer stage, or {non-smoking, light smoking, heavy smoking} for smoking status), we need predefine the distance matrix between categories i and j, for instance, $d_{ij}=|i-j|$, or $d_{ij}=|i-j|^2$. The distance covariance between composition X and ordinal phenotype Y has the following expression

$$dCov^2(\boldsymbol{X},Y) = \left(\sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j d_{ij}\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j D_{ij}\right) + \sum_{i=1}^{K}\sum_{j=1}^{K} p_i p_j d_{ij} D_{ij} - 2\sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{l=1}^{K} p_i p_j p_l d_{il} D_{ij},$$

and one may use the same permutation procedure to obtain p-values. In practice, the distance matrix d_{ij} should be carefully chosen to reflect the true spacings between categories. An inappropriate choice of d_{ij} may result in misleading conclusions. Second, our test might be improved by incorporating more information about bacteria taxa. For instance, one can assign different weights for different bacterial taxa based on their position in the polygenetic tree [28], and use weighted Euclidean distance to construct the test statistic.

In addition to the microbiome application that we illustrated in this paper, the proposed test can be readily applied to several other fields. For instance, the market share data in economics are compositional and often high-dimensional [57]. One may apply our test to detect the market share difference between multiple countries. In geology, it is often of interest to study the compositions of species in sediment [58] and it is possible to apply our test to detect the difference in species compositions between multiple locations.

BIBLIOGRAPHY

[1] A. Agresti. An Introduction to Categorical Data Analysis. Wiley-Interscience, Hoboken, NJ, second edition, 2007.

[2] Q. Zhang, T. Dao. A Distance Based Multi-Sample Test for High Dimensional Compositional Data with Applications to the Human Microbiome. BMC Bioinformatics 21, 205 (2020). https://doi.org/10.1186/s12859-020-3530-x.

[3] J. Aitchison. The Statistical Analysis of compositional data. J R Stat Soc Ser B. 1982; 44(2): 139-77.

[4] J. Fry, T. Fry, K. McLauren. Compositional data analysis and zeros in micro data. Appl Econ, 32(8):953-9, 2010.

[5] Cao Y, Lin W, Li H. Two-sample tests of high-dimensional means for compositional data. Biometrika. 2017; 105(1):115–32.

[6] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017 Nov 15; 8:2224. doi: 10.3389/fmicb.2017.02224. PMID: 29187837; PMCID: PMC5695134.

[7] K.B. Gregory, R.J. Carroll, V. Baladandayuthapani, and S.N Lahiri. A two-sample test for equality of means in high dimension. Journal of the American Statistical Association, 110(510):837–849, 2015.

[8] J. Aitchison. The statistical analysis of compositional data. Caldwell: Blackburn Press; 2003.

[9] P. J. McMurdie, S. Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 2014

[10] M. L. Shaffer. Minimum population sizes for species conservation. *BioScience* 31, 131–134. (1981).

[11] Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25.1–R25.9. doi: 10.1186/gb-2010-11-3-r25.

[12] Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106.

[13] Martín-Fernández, J., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G., and Potenza, R. (1998). Measures of difference for compositional data and hierarchical clustering methods. *Proc. IAMG*. 98, 526–531.

[14] Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). Unifrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133.

[15] Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* 6:21887. doi: 10.7554/eLife.21887.

[16] Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data.* London: John Wiley & Sons.

[17] G. Szekely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6):2769–2794, 2007.

[18] Q. Zhang. Independence test for large sparse contingency tables based on distance correlation. Statistics and Probability Letters, 148:17–22, 2019.

[19] Chen J, Bittinger K, Charlson E, Hoffmann C, Lewis J, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012; 28(16):2106–13.

[20] Lahti L, Salojarvi J, Salonen A, Scheffer M, de Vos W. Tipping elements in the human intestinal ecosystem. Nat Commun. 2014; 5(4344):1–10.

[21] Josse J, Holmes S. Measures of dependence between random vectors and tests of independence: a survey. 2014. arXiv:1307.7383.

[22] Charlson E, Chen J, Custers-Allen R, Bittinger K, Li H, et al. Disordered microbial communities in the upper respiratory tract of cigarette smokers. PLoS ONE. 2010; 5(12):e15216.

[23] L.Song, P. Langfelder, S. Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics 13, 328 (2012). https://doi.org/10.1186/1471-2105-13-328.

[24] Zhu L, Xu K, Li R, Zhong W. Projection correlation between two random vectors. Biometrika. 2018; 104(4):829–43.

[25] D.Y. Chao, Y. Chen, J. Chen, S. Shi, Z. Chen, C. Wang, et al. Genome-wide Association Mapping Identifies a New Arsenate Reductase Enzyme Critical for Limiting Arsenic Accumulation in Plants. PLoS Biol 12(12): e1002009, 2014. https://doi.org/10.1371/journal.pbio.1002009.

[26] V. Graziano and M. Nakai. A geometrical framework for covariance matrices of continuous and categorical variables. Sociological Methods & Research, 44(1):48–79, 2015.

[27] W. Luo, M.S. Friedman, K. Shedden, K.D. Hankenson, and P.J. Woolf. Gage: Generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics, 10(1):161–161, 2009.

[28] Tang Y, Ma L, Nicolae D. A phylogenetic scan test on Dirichlet-tree multinomial model for microbiome data. Ann Appl Stat. 2018; 12(1):1–26.

[29] A.P. Smith, K. Hoek, and D. Becker. Whole-genome expression profiling of the melanoma progression pathway reveals marked molecular differences between nevi/melanoma in situ and advanced-stage melanomas. Cancer Biology and Therapy, 4(9):1018, 2005.

[30] Y. Xia, J. Sun. Hypothesis Testing and Statistical Analysis of Microbiome. Genes & Diseases. 4. 10.1016/j.gendis.2017.06.001.

[31] M.S. Srivastava, S. Katayama, and Y. Kano. A two-sample test in high dimensional data. Journal of Multivariate Analysis, 114:349–358, 2013.

[32] J. Morais, C. Thomas-Agnan, M. Simioni. Using compositional and Dirichlet models for market share regression. Journal Applications Statistics; 45(9):1670–89; 2018.

[33] S. Kim and D.J. Volsky. Parametric analysis of gene set enrichment. BMC Bioinformatics, 6(1):144–144, 2005.

[34] Flood R, Bloemsma M, Weltje G, Barr I, O'Rourke S, et al.Compositional data analysis of Holocene sediments from the West Bengal Sundarbans, India: Geochemical proxies for grain-size variability in a delta environment. Appl Geochem. 2016; 75:222–35.

[35] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S. Pomeroy, T.R. Golu, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43):15545– 15550, 2005.

[36] J. Hardin, A. Mitani, L. Hicks, B. VanKoten. A Robust Measure and testing dependence by correlation between two genes on a microarray. BMC Bioinformatics, 8:220; 2007. doi: 10.1186/1471-2105-8-220.
[37] Szekely G, Rizzo M. Energy statistics: A class of statistics based on distances. J Stat Plan Infer. 2013; 143(8):1249–72.

[38] P.R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(4):515–530, 2005.

[39] H. Chai, H. Jiang, L. Lin, L. Liu. A marginalized two-part Beta regression model for microbiome compositional data. PLoS Computational Biology, 14(7): e1006329; 2018. https://doi.org/10.1371/journal.pcbi.1006329.

[40] J. Szkely and L. Maria. Partial distance correlation with methods for dissimilarities. The Annals of Statistics, 42(6):2382–2412, 2014.

[41] X. Wang, W. Pan, W. Hu, Y. Tian, and H. Zhang. Conditional distance correlation. Journal of the American Statistical Association, 110:0–0, 01 2015.

[42] Q. Zhang, G. Mahdi, J. Tinker, and H. Chen. A graph-based multi-sample test for identifying pathways associated with cancer progression. Computational Biology and Chemistry, 87:107285, 2020.

[43] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, A. Valencia. EnrichNet: Network-based gene set enrichment analysis. Bioinformatics (Oxford, England), 28(18): i451, 2012.

[44] T.T. Cai, W. Liu, Y. Xia. Two-sample test of high dimensional means under dependence. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 76(2):349–372, 2014.

[45] K. Tu, H. Yu, and M. Zhu. MEGO: Gene functional module expression based on gene ontology. BioTechniques, 38(2):277–283, 2005.

[46] W. Lin, P. Shi, R. Feng, H. Li. Variable selection in regression with compositional covariates. Biometrika. 101(4): 785-797, Dec 2014. https://doi.org/10.1093/biomet/asu031.

[47] M. Tsagris, S. Preston and A. Wood. Nonparametric hypothesis testing for equality of means on the simplex. Journal of Statistical Computation and Simulation 87 (2016): 406 - 422.

[48] Y. Wang, Y. Li, H. Cao et al.. Efficient test for nonlinear dependence of two continuous variables. BMC Bioinformatics 16, 260 (2015). https://doi.org/10.1186/s12859-015-0697-7.

[49] R. Li, W. Zhong & L. Zhu. Feature Screening via Distance Correlation Learning, Journal of the American Statistical Association, 107:499, 1129-1139, 2012.

[50] Q. Zhang, J. Burdette, and J. Wang. Integrative network analysis of TCGA data for ovarian cancer. BMC Systems Biology, 8:1338, 12 2014.

[51] D. Cheriton and R.E. Tarjan. Finding minimum spanning trees. SIAM Journal on Computing, 5(4):724–742, 1976.

[52] H.Chen and Jerome H. Friedman. A new graph-based two-sample test for multivariate and object data. Journal of the American Statistical Association, 112(517):397–409, 2017;2013.

[53] J.A. Hoeting, R.A. Davis, A.A. Merton, S.E. Thompson. Model Selection For Geostatistical Models. Ecological Applications, 16: 87-98, 2006. https://doi.org/10.1890/04-0576.

[54] Matteson D, James N. A Nonparametric Approach for Multiple Change Point Analysis of Multivariate Data. J Am Stat Assoc. 2014; 109(505):334–45.

[55] R. Smith. A mutual information approach to calculating nonlinearity. *STAT*, 4: 291– 303, 2015. doi: 10.1002/sta4.96.

[56] Shen C, Priebe C, Vogelstein J. From Distance Correlation to Multiscale Graph Correlation; 2019. In Press. https://doi.org/10.1080/01621459.2018.1543125.

[57] Morais J, Thomas-Agnan C, Simioni M. Using compositional and Dirichlet models for market share regression. J Appl Stat. 2018; 45(9):1670–89.

[58] Flood R, Bloemsma M, Weltje G, Barr I, O'Rourke S, et al.Compositional data analysis of Holocene sediments from the West Bengal Sundarbans, India: Geochemical proxies for grain-size variability in a delta environment. Appl Geochem. 2016; 75:222–35.