AN ASSOCIATION TEST BASED ON THE MIXTURE OF ZERO-INFLATED POISSON REGRESSION MODELS FOR DETECTING DIFFERENTIAL MICROBIAL ABUNDANCE IN CASE-CONTROL STUDIES

by

Maoyu Zhu

A Thesis

Presented to

The University of Guelph

In partial fulfilment of requirements

for the degree of

Master of Science

in

Mathematics and Statistics

Guelph, Ontario, Canada

ABSTRACT

An association test based on the mixture of zero-inflated Poisson regression models for detecting differential microbial abundance in case-control studies.

Maoyu Zhu

University of Guelph, 2017

Advisor:

Professor Z. Feng

Motivation: The human microbial communities play an important role in human health and disease because human metabolism, nutrient intake and energy generation fall under the influence of these communities. Association analysis concerning relative abundances among these communities with status-related outcomes can provide essential information, which can help us to understand the impact that changes in the relative abundances profile can have on disease status. Proper testing of overdispersion and zero-inflated microbiome data is challenging. Existing methods fail to pinpoint the degree of association.

Results: In this thesis, we propose a likelihood ratio test for testing the association between the relative abundance of bacteria and disease covariate for microbiome data while using a generalized zero-inflated Poisson regression mixture model. Simulation studies have shown that the likelihood ratio statistic, which examines the null hypothesis that the distribution of the bacterial count arises from healthy individuals and individuals with disease is the same versus the alternative hypothesis that the

distribution of the bacterial count arises from healthy individuals and individuals with disease are different, converges to a χ^2 distribution. The power of the likelihood ratio test is also evaluated by our simulation study. The application of our proposed method on the real microbiome data has shown that the associated bacteria at the genus level has different distributions of the bacteria counts between the healthy individuals and individuals with carcinoma. Our proposed method provides a useful tool for identifying differentiate taxonomic abundances underlying different disease status.

AKNOWLEDGEMENT

I would like to thank my advisor, Zeny Feng for her continuous support of my studies here at University of Guelph, also I would like to thank Ayesha Ali as my co-advisor. I would also like to thank Julie Horrocks, Jeremy Balka, and Edward Carter as my thesis defence examiner. I also appreciate the help from Susan McCormick in the department of Mathematics and Statistics. I sincerely appreciate the understanding and support from my family and colleagues.

Table Of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the human body, all microorganisms are included in human microbial communities. Human metabolism, nutrient intake and energy generation influence these communities; therefore the communities play important roles in human health and diseases, especially associated with eating habits. For example, obesity, diabetes and inflammatory bowel disease have been shown to be correlated with certain type gut microbiome communities [Manichanh et al., 2012; Qin et al., 2012; Turnbaugh et al., 2006]. Microbiome data involve the quantification of relative abundances of bacteria in the community, and they usually come from two high-throughput sequencing-based approaches. One approach profiles the bacterial taxonomic composition using the ubiquitous RNA marker gene, 16S ribosomal RNA (rRNA). Another approach depends on the shotgun metagenomic sequence, which sequences the microbial genomes contained in the sample. These two approaches provide the data regarding bacterial communities and are widely applied in microbiome studies, including, for example the

Human Microbiome Project (HMP) [Turnbaugh et al., 2007] and the Metagenomic of the Human Intestinal Tract (MetaHIT) project [Qin et al., 2010].

In order to quantify microbial abundances, some known sequences are used as references to identify the sequencing reads [Segata et al., 2012]. Since the materials carried by DNA differ between samples, the reads counts of microbial abundance are not comparable across samples. Thus, the reads counts are normalized to the relative abundances. Then processing software such as "mothur" [Schloss et al., 2009] can be applied to produce relative abundances of the bacterial counts at the different taxonomic rank levels which contains many zeros at the leaf levels such as genus or the strain of the phylogenetic tree.

It is worth investigating how microbial abundance is associated with various covariates such as disease status. Since the microbiome data consist of bacterial counts, a classical Poisson regression model can be used to analyze count data. However, the empirical count data often exhibit overdispersion and contain more zeros than expected under the Poisson model. A zero-inflated Poisson (ZIP) regression model has been widely used to deal with this problem. For example, a population with excess zeros is considered to have an extra proportion of zeros added to the proportion of zeros from the original Poisson distribution [Van den Broek, 1995]. This phenomenon often results from heterogeneous count data, which are commonly observed in many applications [Lim et al., 2014]. For example, in order to assess dental caries for each individual, the DMFT index, which is calculated by the number of filled, decayed and missing teeth, is used. [Lim et al., 2014].

If a population has excess zeros and several sub-populations have different means of counts, then a single Poisson component in the ZIP regression model may not be sufficient to describe the non-zero counts [Lim et al., 2014]. An alternative method such as a generalized zero-inflated Poisson (GZIP) regression mixture model can provide a better solution. This model was discussed by Lim et al. [2014] and proposed with covariates in the mixing proportion parameter and Poisson mean parameter so that each observation was allowed to have different mixing proportions and Poisson means. In our gut microbiome data which are analyzed in Chapter 4, the bacterial index indicates the number of counts of bacteria at genus level in gut microbial communities. As expected, a large number of subjects have no observed bacteria of certain types in the sample, which illustrates the zero-inflation. The samples of bacterial counts might be drawn from a population consisting of 2, 3 or 4 sub-populations, which illustrates one or more Poisson components. Therefore, to model such sparse bacterial data, a GZIP regression mixture model might be useful.

Since the number of components is unknown in the microbiome data, several methods have been used to determine the number of components in a mixture distribution; for instance, the Bayesian information criterion (BIC) are discussed by Lim et al. [2014]. When the number of components is known, we are interested in testing for each bacterial type at the genus level to determine whether the distribution of the bacterial counts is the same between healthy individuals and the individuals with carcinoma. Under regularity conditions for each given bacteria, the likelihood ratio statistic is used to test whether the distribution of the bacteria count in a sample from the healthy individuals is the same as those from the carcinoma individuals.

Notably, under the null hypothesis, the likelihood ratio statistic has an asymptotic chi-square (χ^2) distribution with the appropriate degrees of freedom.

The likelihood ratio statistic is calculated using estimated parameters, making a parameter estimation necessary. Maximum likelihood (ML) is a standard method for estimating parameters. However, the classic ML is of limited use when unobserved data are involved. One approach to estimate ML from incomplete data is to use an expectation maximization (EM)algorithm, as discussed by Dempster et al. [1977]. Bacterial counts are unclassified in terms of the corresponding component, resulting in the unobserved data; thus, the EM algorithm is useful in finding the maximum likelihood estimations (MLEs) of parameters n this case. The EM algorithm follows the natural framework of maximum likelihood estimates but before it does that, the missing data is inserted into the original data in E-step. It is essentially an iterative computation algorithm that will converge to parameter values at a local maximum of the likelihood function [Collins, 1997]. If the solution to the M-step is not available in the closed form of differentiating the likelihood function with respect to parameters, other techniques, such as iteratively reweighted least squares (IRWLS), can be used in the inner loop of the M-step to obtain the updated ML estimators [Lim et al., 2014]. The IRWLS algorithm has been developed by Carroll and Ruppert [1988].

Dempster et al. [1977] proved that the log-likelihood of the complete data at each iteration of the algorithm is non-decreasing and that the log-likelihood converges to some global maximum. Jamshidian and Jennrich [1993] emphasized that the EM algorithm often converges slowly. In order to speed up convergence, it was suggested to use the Aitken acceleration-based stopping criterion (ACC), based on a multi-

class classification of the EM algorithm [Ng et al., 2012]. The ACC was suggested by McNicholas et al. [2010].

The focus of this thesis is to compare the distributions of bacterial counts under different disease status while using the GZIP regression mixture model. Our model handles excess zero counts and overdispersion which can be specified by several sub-components. It provides flexibility to various compositional structures among the abundance of bacteria counts, and it potentially carries out more information from the original data. Our test allows for the judging of evidence, pinpointing the actual probability that evidence from the microbial data will support the assumed existence of an association between the abundance of a bacteria and disease status.

The thesis is organized as follows. In chapter 2, we present the GZIP regression mixture model and the likelihood ratio test based on the GZIP regression mixture model. In chapter 3, we describe the procedure for the simulation study, and examine the appropriateness of using the χ^2 distribution to approximate the likelihood ratio statistic under the null hypothesis, and assess the power of the proposed method through simulations. In chapter 4, we compare bacterial distributions under different disease status by applying our proposed method to the gut mucosal microbiome data. Finally, we conclude with a discussion in chapter 5.

Chapter 2

Methodology

In this chapter, we briefly review the finite mixture model, a zero-inflated Poisson regression mixture model underlying the likelihood ratio test; methods for fitting these models are also discussed.

2.1 Finite Mixture Model

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a random sample of size n from a finite Poisson mixture distribution with probability distribution function given by:

$$f(\boldsymbol{y}; \boldsymbol{\Omega}) = \sum_{j=1}^{J} \boldsymbol{\pi}_j \phi_j (\boldsymbol{y}; \boldsymbol{\theta}_j) \tag{2.1}$$

where $\boldsymbol{\Omega} = (\pi_1, \ldots, \pi_j, \theta_1, \ldots, \theta_j)$ is the vector of unknown parameters in which, $0 < \pi_j < 1$ is the mixing proportion which corresponds to the j^{th} component with restrictions that $\sum_{j=1}^{J} \pi_j = 1$, J is the number of components, θ_j is the mean of the Poisson distribution, and $\phi_j(.)$ is the Poisson probability mass function for the j^{th}

component.

2.2 Zero-Inflated Poisson Regression Model

If the mixture distribution in Eq.(2.1) has two components where one component is for a Poisson distribution and another component is for a population of zeros. This is called the zero-inflated Poisson distribution and $\boldsymbol{\Omega} = (\pi, \theta)$:

$$\begin{cases} y_i \sim Poiss\big(\theta\big) & \text{with probability } \pi \\ y_i = 0 & \text{with probability } 1-\pi \end{cases}$$

The zero-inflated Poisson $\big(\text{ZIP}\big)$ distribution can be rewritten as:

$$\boldsymbol{y} = \begin{cases} y & \text{with probability } \pi\frac{e^{-\theta}\theta^y}{y!}, \quad y = 1, 2, 3, \ldots \\ 0 & \text{with probability } \pi e^{-\theta} + \big(1-\pi\big) \end{cases}$$

If the mixture distribution in Eq.(2.1) has J components and the last component is for the population of zeros only, then $\boldsymbol{\Omega} = (\pi_1, \ldots, \pi_{J-1}, \theta_1, \ldots, \theta_{J-1})$ and $\pi_J = 1 - \sum_{j=1}^{J-1} \pi_j$. Thus, the generalized zero-inflated Poisson $\big(\text{GZIP}\big)$ mixture model can be formulated as below:

$$f\big(\boldsymbol{y}; \boldsymbol{\Omega}\big) = \sum_{j=1}^{J-1} \boldsymbol{\pi}_j Pois\big(\boldsymbol{y}; \boldsymbol{\theta}_j\big) + \big(1 - \sum_{j=1}^{J-1} \boldsymbol{\pi}_j\big) I_{[y_i=0]} \tag{2.2}$$

where $I_{[.]}$ is 1 if the specified condition is satisfied and 0 otherwise, and $Pois\big(\boldsymbol{y}; \boldsymbol{\theta}_j\big)$ denotes the Poisson probability mass function for $\boldsymbol{y}$ with mean $\boldsymbol{\theta}_j$.

In order for the mixing proportion and the mean of the Poisson distribution to depend on covariates, we generalized the GZIP mixture model to the GZIP regression

mixture model, where a multinomial regression model is used to link the covariate vector of i^{th} subject, $\boldsymbol{x}_i$, with the mixing proportion, and so it becomes subject specific as $\boldsymbol{\pi}_{ij}$ and a log-linear regression model is used to link $\boldsymbol{x}_i$ with Poisson mean $\boldsymbol{\theta}_j$ and so it becomes subject specific as well. The two link functions are given by:

$$log\left(\theta_{ij}\right) = \boldsymbol{x}_i\boldsymbol{\gamma}_j, \quad i = 1, \ldots, n, \;\; j = 1, \ldots, J-1$$

$$\pi_{ij}\left(\boldsymbol{\beta}_j, \boldsymbol{x}_i\right) = \frac{\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}_j\right)}{1 + \sum_{j=1}^{J-1} e^{\boldsymbol{x}_i\boldsymbol{\beta}_j}}, \quad j = 1, \ldots, J-1 \qquad \pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij}\left(\boldsymbol{\beta}_j, \boldsymbol{x}_i\right)$$

where $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ is $(p+1) \times 1$ row vectors of 1 and covariates and $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ are the corresponding $(p+1) \times 1$ vectors of intercept and regression coefficients for the j^{th} component, respectively. Note that the mixing proportion of the last component $\pi_{iJ}\left(\boldsymbol{\beta}_j, \boldsymbol{x}_i\right)$ is the probability of excess zeros, and it is used as the baseline for the multinomial logit. That is, the logit for the other component relative to π_{iJ} is $log\big(\pi_{ij}/\pi_{iJ}\big) = \boldsymbol{x}_i\boldsymbol{\beta}_j$, $j = 1, \ldots, J-1$. Thus, the generalized zero-inflated Poisson $\big(\text{GZIP}\big)$ regression mixture model can be formulated as below:

$$f\big(\boldsymbol{y}_i; \boldsymbol{\vartheta}, \boldsymbol{x}_i\big) = \sum_{j=1}^{J-1} \boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j, \boldsymbol{x}_i\big) Pois\big(\boldsymbol{y}_i; \boldsymbol{\theta}_{ij}\big(\boldsymbol{\gamma}_j, \boldsymbol{x}_i\big)\big) + \big[1 - \sum_{j=1}^{J-1} \boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j, \boldsymbol{x}_i\big)\big] I_{[y_i=0]} \tag{2.3}$$

where $\boldsymbol{\vartheta} = \big(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \ldots, \boldsymbol{\beta}_{J-1}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_{J-1}\big)$ is the vector of unknown parameters, $I_{[.]}$ is 1 if the specified condition is satisfied and 0 otherwise, and $Pois\big(\boldsymbol{y}; \boldsymbol{\theta}_j\big(\boldsymbol{\gamma}_j, \boldsymbol{x}\big)\big)$ denotes the Poisson probability mass function of y_i with mean $\theta_{ij}\left(\boldsymbol{\gamma}_j, \boldsymbol{x}_i\right)$. Here, for a case-control study, there is only one covariate with two levels, where the case group represents individuals being affected by a disease and control group represents individuals being disease free. In other words, we have $p = 1$ and $x_i = 1$ if the i^{th} subject is affected by a disease, or 0 if the subject is healthy. The link functions

become:

$$log(\theta_{ij}) = \boldsymbol{\gamma}_{j0} + \boldsymbol{\gamma}_{j1}\boldsymbol{x}_i$$
$$\pi_{ij} = \frac{\exp\left(\boldsymbol{\beta}_{j0} + \boldsymbol{\beta}_{j0}\boldsymbol{x}_i\right)}{1 + \sum_{j=1}^{J-1}\exp\left(\boldsymbol{\beta}_{j0} + \boldsymbol{\beta}_{j0}\boldsymbol{x}_i\right)}$$

2.3 Hypothesis Testing in Finite Mixture Models

Suppose for a given bacteria, if the bacteria has no association with disease status, the mixing proportion for each component and the mean parameters should be the same for the two groups of individuals. Equivalently, the mixing proportion and the Poisson mean parameters should no longer depend on $\boldsymbol{x}_i's$ such that all coefficients associated with the disease status, $\boldsymbol{x}_i's$, are 0. To test whether the bacteria of interest is associated with the disease status, we perform an overall test as:

$$H_0 : \boldsymbol{\beta}_{j1} = 0,\ \boldsymbol{\gamma}_{j1} = 0 \qquad \forall j = 1, \ldots, J-1$$
$$H_1 : \exists\ \boldsymbol{\beta}_{j1} \neq 0 \text{ or } \boldsymbol{\gamma}_{j1} \neq 0 \qquad \forall j = 1, \ldots, J-1$$

where $j = 1, \ldots, J-1$ for the $J-1$ different Poisson components and the last J^{th} component containing all zeros. Let the likelihood ratio test statistic be

$$\begin{aligned} \Lambda &= 2\left(L_1(\hat{\boldsymbol{\vartheta}}_1; \boldsymbol{x}; \boldsymbol{y}) - L_0(\hat{\boldsymbol{\vartheta}}_0; \boldsymbol{x}; \boldsymbol{y})\right) \\ &= 2\sum_{i=1}^{n} log\frac{f\left(\boldsymbol{y}_i; \hat{\boldsymbol{\vartheta}}_1; \boldsymbol{x}_i\right)}{f\left(\boldsymbol{y}_i; \hat{\boldsymbol{\vartheta}}_0; \boldsymbol{x}_i\right)} \end{aligned} \tag{2.4}$$

where

$$\hat{\boldsymbol{\vartheta}}_1 = \left(\hat{\beta}_{10}, \hat{\beta}_{11}, \ldots, \hat{\beta}_{J-1,0}, \hat{\beta}_{J-1,1}, \hat{\gamma}_{10}, \hat{\gamma}_{11}, \ldots, \hat{\gamma}_{J-1,0}, \hat{\gamma}_{J-1,1}\right)$$

and

$$\hat{\boldsymbol{\vartheta}}_0 = \left(\hat{\beta}_{10}, \hat{\beta}_{11} = 0, \ldots, \hat{\beta}_{J-1,0}, \hat{\beta}_{J-1,1} = 0, \hat{\gamma}_{10}, \hat{\gamma}_{11} = 0, \ldots, \hat{\gamma}_{J-1,0}, \hat{\gamma}_{J-1,1} = 0\right)$$

are the maximum likelihood estimators of $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_0$ under the alternative and null hypotheses, respectively; $f(.)$ is given in Eqs.(2.3). Under the regularity conditions, the likelihood ratio test statistic is known to follow a chi-square distribution $\left(\chi^2_{df}\right)$ asymptotically under the null hypothesis with the degrees of freedom being the difference of the number of parameters in the models between the alternate hypothesis and the null hypothesis. In this case, the degrees of freedom is $2 \times (J-1)$. That is, $\Lambda \sim \chi^2_{2(J-1)}$ asymptotically.

2.4 Finding Differential Abundance of Bacteria in Case-control Studies

In case-control studies, the various covariates $\boldsymbol{x}_i$ are reduced to one covariate, which is disease status, where $\boldsymbol{x}_i = 1$ denotes an individual with the disease and $\boldsymbol{x}_i = 0$ denotes a healthy individual. The bacterial counts at the genus level are the response variable $\boldsymbol{y}_i$. Since the bacterial counts contain excess zeros and overdispersion, we use the GZIP regression mixture model to model the distribution of bacteria counts. Three combinations of relative bacterial abundance are commonly shown in the gut mucosal microbiome data when using the smallest Bayesian information criterion $\left(\text{BIC}\right)$ to identify the number of components J, where $J = 2, 3, 4$. Therefore, three GZIP regression mixture models are used. The likelihood ratio test can be applied to test the overall association between disease status and the relative abundance.

2.4.1 Two Components ZIP regression mixture model

Suppose for a given bacteria, a GZIP regression mixture model that has two components is adequate to model the abundance count for a given sample. Under the two components GZIP regression mixture model, $J = 2$ and so we have:

$$f\left(\boldsymbol{y}_i; \boldsymbol{\vartheta}, \boldsymbol{x}_i\right) = \boldsymbol{\pi}_i\big(\boldsymbol{\beta}, \boldsymbol{x}_i\big) Pois\big(\boldsymbol{y}_i; \boldsymbol{\theta}_i\big(\boldsymbol{\gamma}, \boldsymbol{x}_i\big)\big) + \big[1 - \boldsymbol{\pi}_i\big(\boldsymbol{\beta}, \boldsymbol{x}_i\big)\big] I_{[y_i=0]}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, and

$$log\big(\boldsymbol{\theta}_i\big) = \gamma_0 + \gamma_1 \boldsymbol{x}_i \tag{2.5}$$

$$log\big(\boldsymbol{\pi}_i / \big(1 - \boldsymbol{\pi}_i\big)\big) = \beta_0 + \beta_1 \boldsymbol{x}_i \tag{2.6}$$

To test whether the given bacteria is associated with the disease status, we test:

$$H_0 : \beta_1 = \gamma_1 = 0$$

$$H_1 : \beta_1 \neq 0 \text{ or } \gamma_1 \neq 0$$

Under the null hypothesis, $\Lambda \sim \chi^2_2$ distribution asymptotically. The null hypothesis corresponds to the situation that the bacteria count in the sample follows a two-component ZIP mixture model given by:

$$f\big(\boldsymbol{y}_i; \boldsymbol{\Omega}\big) = \boldsymbol{\pi} \frac{e^{-\theta} \theta^{y_i}}{y_i!} + \big(1 - \boldsymbol{\pi}\big) I_{[y_i=0]} \qquad y_i = 0, 1, \ldots$$

In this situation, $\boldsymbol{\pi}_i = \boldsymbol{\pi}$ and $\boldsymbol{\theta}_i = \boldsymbol{\theta}$ are constants for all i disregarding the disease status.

2.4.2 Three And Four Components GZIP regression model

Assume the GZIP regression mixture model with three components is adequate to model the bacterial count of a given sample. With a three-component model, $J = 3$ and the distribution function for the i^{th} observation is:

$$\begin{aligned} f\left(\boldsymbol{y}_i; \boldsymbol{\vartheta}, \boldsymbol{x}_i\right) &= \boldsymbol{\pi}_{i1}\big(\boldsymbol{\beta}_1, \boldsymbol{x}_i\big) Pois\big(\boldsymbol{y}_i; \boldsymbol{\theta}_{i1}\big(\boldsymbol{\gamma}_1, \boldsymbol{x}_i\big)\big) + \boldsymbol{\pi}_{i2}\big(\boldsymbol{\beta}_2, \boldsymbol{x}_i\big) Pois\big(\boldsymbol{y}_i; \boldsymbol{\theta}_{i2}\big(\boldsymbol{\gamma}_2, \boldsymbol{x}_i\big)\big) \\ &\quad + \big[1 - \boldsymbol{\pi}_{i1}\big(\boldsymbol{\beta}_1, \boldsymbol{x}_i\big) - \boldsymbol{\pi}_{i2}\big(\boldsymbol{\beta}_2, \boldsymbol{x}_i\big)\big] I_{[y_i=0]} \end{aligned} \tag{2.7}$$

where $\boldsymbol{\beta}_1 = \big(\beta_{10}, \beta_{11}\big)$, $\boldsymbol{\beta}_2 = \big(\beta_{20}, \beta_{21}\big)$, $\boldsymbol{\gamma}_1 = \big(\gamma_{10}, \gamma_{11}\big)$ and $\boldsymbol{\gamma}_2 = \big(\gamma_{20}, \gamma_{21}\big)$. According to these definitions, the log-link for the Poisson means $\boldsymbol{\theta}_{i1}$, $\boldsymbol{\theta}_{i2}$ and the logit link for the mixing proportion $\boldsymbol{\pi}_{i1}$, $\boldsymbol{\pi}_{i2}$ are given as:

$$log\big(\boldsymbol{\theta}_{i1}\big) = \gamma_{10} + \gamma_{11}\boldsymbol{x}_i \qquad log\big(\boldsymbol{\theta}_{i2}\big) = \gamma_{20} + \gamma_{21}\boldsymbol{x}_i \tag{2.8}$$

$$log\big(\boldsymbol{\pi}_{i1}/\big(1 - \boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2}\big)\big) = \beta_{10} + \beta_{11}\boldsymbol{x}_i, \quad log\big(\boldsymbol{\pi}_{i2}/\big(1 - \boldsymbol{\pi}_{i1} - \boldsymbol{\pi}_{i2}\big)\big) = \beta_{20} + \beta_{21}\boldsymbol{x}_i \tag{2.9}$$

The association between the bacteria and the disease status can be tested by the following overall hypotheses:

$$H_0 : \beta_{11} = \beta_{21} = \gamma_{11} = \gamma_{21} = 0$$

$$H_1 : \beta_{11} \neq 0 \text{ or } \beta_{21} \neq 0 \text{ or } \gamma_{11} \neq 0 \text{ or } \gamma_{21} \neq 0$$

Under the null hypothesis, the likelihood ratio statistic Λ follows a χ^2_4 distribution asymptotically.

Similarly, for the four-component mixture model, $J = 4$ and we have the distri-

bution function as:

$$f\left(\boldsymbol{y}_i;\boldsymbol{\vartheta},\boldsymbol{x}_i\right)=\sum_{j=1}^{3}\boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j,\boldsymbol{x}_i\big)Pois\big(\boldsymbol{y}_i;\boldsymbol{\theta}_{ij}\big(\boldsymbol{\gamma}_j,\boldsymbol{x}_i\big)\big)+\big[1-\sum_{j=1}^{3}\boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j,\boldsymbol{x}_i\big)\big]I_{[y_i=0]}$$

The overall hypothesis test can be formulated as:

$$H_0:\beta_{11}=\beta_{21}=\beta_{31}=\gamma_{11}=\gamma_{21}=\gamma_{31}=0$$

$$H_1:\beta_{11}\neq 0\text{ or }\beta_{21}\neq 0\text{ or }\beta_{31}\neq 0\text{ or }\gamma_{11}\neq 0\text{ or }\gamma_{21}\neq 0\text{ or }\gamma_{31}\neq 0$$

The test statistic Λ follows a χ^2_6 distribution asymptotically under the null hypothesis.

2.5 Parameter Estimation

When the null and alternative hypotheses are defined, a parameter estimation method is needed. Suppose for a J component mixture model, given a sample of n observations based on the mixture model in Eqs.(2.1), the likelihood function is given by:

$$L\left(\boldsymbol{\Omega}\right)=\prod_{i=1}^{n}\left\{\sum_{j=1}^{J}\boldsymbol{\pi}_j\phi_j\left(\boldsymbol{y}_i;\boldsymbol{\theta}_j\right)\right\} \tag{2.10}$$

Then the log-likelihood function becomes:

$$l\left(\boldsymbol{\Omega}\right)=\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{J-1}\boldsymbol{\pi}_j\phi_j\left(\boldsymbol{y}_i;\boldsymbol{\theta}_j\right)+\Big(1-\sum_{j=1}^{J-1}\boldsymbol{\pi}_j\Big)\phi_J\left(\boldsymbol{y}_i;\boldsymbol{\theta}_j\right)\right\} \tag{2.11}$$

Taking the conventional approach for finding the maximum likelihood estimations $\big(\text{MLEs}\big)$ of $\boldsymbol{\pi}'_j s$ and $\boldsymbol{\theta}'_j s$, we differentiate the log-likelihood function in Eqs.(2.11) with respect to $\boldsymbol{\pi}'_j s$ and $\boldsymbol{\theta}'_j s$, and equate them to zero. We obtain the likelihood

equations:

$$\begin{aligned}\frac{\partial l(\boldsymbol{\pi}_j;\boldsymbol{\theta}_j)}{\partial \boldsymbol{\pi}_j} &= \sum_{i=1}^{n} \frac{\phi_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j)-\phi_J(\boldsymbol{y}_i;\boldsymbol{\theta}_j)}{\sum_{j=1}^{J-1}\boldsymbol{\pi}_j\phi_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j)+\Big(1-\sum_{j=1}^{J-1}\boldsymbol{\pi}_j\Big)\phi_J(\boldsymbol{y}_i;\boldsymbol{\theta}_j)} \\ &= 0 \qquad\qquad \forall j = 1,\ldots,J-1\end{aligned}$$

$$\begin{aligned}\frac{\partial l(\boldsymbol{\pi}_j;\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} &= \sum_{i=1}^{n} \frac{\frac{\partial \phi_j\left(\boldsymbol{y}_i;\boldsymbol{\theta}_j\right)}{\partial \boldsymbol{\theta}_j}}{\sum_{j=1}^{J-1}\boldsymbol{\pi}_j\phi_j(\boldsymbol{y}_i;\boldsymbol{\theta}_j)+\Big(1-\sum_{j=1}^{J-1}\boldsymbol{\pi}_j\Big)\phi_J(\boldsymbol{y}_i;\boldsymbol{\theta}_j)} \\ &= 0 \qquad\qquad \forall j = 1,\ldots,J-1\end{aligned}$$

The common denominator of the likelihood equations are subject to sample size and the distribution function and so, there is no explicit solution for the MLE$'s$ of $\boldsymbol{\pi}_j's$ and $\boldsymbol{\theta}_j's$. To solve this problem, the maximum likelihood estimates $\hat{\boldsymbol{\vartheta}}_1$ and $\hat{\boldsymbol{\vartheta}}_0$ can be found using the expectation maximization $\big(\text{EM}\big)$ algorithms discussed by Dempster et al. [1977]. The EM algorithm is an iterative computation algorithm that will converge under certain conditions, so that the parameter estimates are stable and no large change will occur in likelihood values.

The EM algorithm was named by Dempster et al. [1977] because, for each iterative process of the algorithm, it consists of an expectation step followed by a maximization step. In the EM framework, the likelihood function in Eqs.(2.8) is called the incomplete-data likelihood function that only include the observed data of x and y where the complete-data includes both observed and unobserved data. In the mixture model, unobserved data refers to the underlying unobserved membership that an observation belongs to which mixing component. The unobserved data is referred to as the missing data. Suppose unobserved data is denoted by $\boldsymbol{z}_i = \big(\boldsymbol{z}_{i1},\ldots,\boldsymbol{z}_{iJ}\big)$,

where J is the number of components. It can be labelled as:

$$\boldsymbol{z}_{ij} = \begin{cases} 1, & \text{if the } i^{th}\text{subject is from the } j^{th}\text{component of the mixture model} \\ 0, & \text{otherwise} \end{cases}$$

and $\sum_{j=1}^{J} z_{ij} = 1$. So, $\boldsymbol{z}_i$ can be reduced to $J-1$ elements and z_{iJ} can be determined by $1 - \sum_{j=1}^{J-1} z_{ij}$. For each subject $\boldsymbol{z}_i$ is ungrounded and is assumed to be one of the j components using the identity vector. Using the combination of the observed data $\big(\boldsymbol{y},\ \boldsymbol{x}\big)$ and unobserved data $\boldsymbol{z}_i$ for $i = 1, \ldots, n$, a complete-data likelihood can be written as:

$$L_c(\boldsymbol{\vartheta}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J-1} \left[\boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) Pois\big(\boldsymbol{y}_i, \boldsymbol{\gamma}_j, \boldsymbol{x}_i\big)\right]^{\boldsymbol{z}_{ij}} \left[1 - \sum_{j=1}^{J-1} \boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x})\right]^{\left(1-\sum_{j=1}^{J-1} \boldsymbol{z}_{ij}\right)} \right\} \tag{2.12}$$

A complete-data log-likelihood can be written as:

$$l_c(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} \boldsymbol{z}_{ij} \log\left[\boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i) Pois\big(\boldsymbol{y}_i, \boldsymbol{\gamma}_j, \boldsymbol{x}_i\big)\right] + \big(1 - \sum_{j=1}^{J-1} \boldsymbol{z}_{ij}\big) \log\left[1 - \sum_{j=1}^{J-1} \boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i)\right] \right\} \tag{2.13}$$

The expectation and maximization steps in the EM algorithm are outlined as follows. Suppose at the $(r+1)^{th}$ iteration, given the observed data $\big(\boldsymbol{y},\ \boldsymbol{x}\big)$ and the currently updated parameter estimated $\boldsymbol{\vartheta}^{(r)} = \Big(\boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}\Big)$.

E-step: Taking the expectation of the complete-data log-likelihood for the i^{th} subject. We have:

$$\begin{aligned} E\Big[l_c\big(\boldsymbol{\vartheta}^{(r)}; \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_i\big)\Big] &= E\Bigg[\textstyle\sum_{i=1}^{n} \Big\{ \sum_{j=1}^{J-1} \boldsymbol{z}_{ij} \log\left[\boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j^{(r)}, \boldsymbol{x}_i) Pois\big(\boldsymbol{y}_i, \boldsymbol{\gamma}_j^{(r)}, \boldsymbol{x}_i\big)\right] \Big\} \\ &\qquad + \textstyle\sum_{i=1}^{n} \Big\{ \big(1 - \sum_{j=1}^{J-1} \boldsymbol{z}_{ij}\big) \log\left[1 - \sum_{j=1}^{J-1} \boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i)\right] \Big\} \Bigg] \\ &= \textstyle\sum_{i=1}^{n} \Big\{ \sum_{j=1}^{J-1} E\left(\boldsymbol{z}_{ij}\right) \log\left[\boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j^{(r)}, \boldsymbol{x}_i) Pois\big(\boldsymbol{y}_i, \boldsymbol{\gamma}_j^{(r)}, \boldsymbol{x}_i\big)\right] \Big\} \\ &\qquad + \textstyle\sum_{i=1}^{n} \Big\{ \big(1 - \sum_{j=1}^{J-1} E\left(\boldsymbol{z}_{ij}\right)\big) \log\left[1 - \sum_{j=1}^{J-1} \boldsymbol{\pi}_{ij}(\boldsymbol{\beta}_j, \boldsymbol{x}_i)\right] \Big\} \end{aligned}$$

It is clear that given $\big(\boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\vartheta}^{(r)}\big)$, $\boldsymbol{z}_i \sim Multinomial\big(1, \boldsymbol{\varsigma}_i\big)$ where $\boldsymbol{\varsigma}_i = \big(\varsigma_{i1}, ..., \varsigma_{i,J-1}\big)^{\intercal}$ and $\varsigma_{ij} = P\left(z_{ij} = 1 \mid \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\vartheta}^{(r)}\right)$.

$$\begin{aligned} \boldsymbol{z}_{ij}^{(r+1)} &= E\big(z_{ij} \mid \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\vartheta}^{(r)}\big) = P\left(z_{ij} = 1 \mid \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{\vartheta}^{(r)}\right) \\ &= \frac{\boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j^{(r)}, \boldsymbol{x}_i\big) Pois\big(\boldsymbol{y}_i, \boldsymbol{\gamma}_j^{(r)}, \boldsymbol{x}_i\big)}{\sum_1^j \boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j^{(r)}, \boldsymbol{x}_i\big) Pois\big(\boldsymbol{y}_i, \boldsymbol{\gamma}_j^{(r)}, \boldsymbol{x}_i\big) + \big(1 - \sum_1^j \boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j^{(r)}, \boldsymbol{x}_i\big)\big)} \end{aligned} \tag{2.14}$$

The maximization step is applied by maximizing the expectation of $l_c\big(\boldsymbol{\beta}, \boldsymbol{\gamma}\big)$ given the observed data with $\hat{\boldsymbol{z}}_{ij}$. We let

$$E\big[l_c\big(\boldsymbol{\vartheta} \mid \boldsymbol{y}_i, \boldsymbol{x}_i, \boldsymbol{z}_{ij}^{(r+1)}\big)\big] = Q_1 + Q_2 \tag{2.15}$$

$$\text{where } Q_1 = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} \hat{\boldsymbol{z}}_{ij}^{(r+1)} \log\big[Pois\big(\boldsymbol{y}_i, \boldsymbol{\gamma}_j, \boldsymbol{x}_i\big)\big] \right\} \tag{2.16}$$

$$Q_2 = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} \hat{\boldsymbol{z}}_{ij}^{(r+1)} \log\big[\boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j, \boldsymbol{x}_i\big)\big] + \big(1 - \sum_{j=1}^{J-1} \hat{\boldsymbol{z}}_{ij}^{(r+1)}\big) \log\big[1 - \sum_{j=1}^{J-1} \boldsymbol{\pi}_{ij}\big(\boldsymbol{\beta}_j, \boldsymbol{x}_i\big) I_{[y_i=0]} \right\} \tag{2.17}$$

The MLE of $\gamma_j's$ and $\beta_j's$ taken from the $\big(r+1\big)^{th}$ iteration are calculated as follows:

$$\frac{\partial Q_1}{\partial \boldsymbol{\gamma}_j} = \left(\frac{\partial Q_1}{\partial \gamma_{0j}}, \frac{\partial Q_1}{\partial \gamma_{1j}}\right)^{\intercal} = 0 \tag{2.18}$$

$$\frac{\partial Q_2}{\partial \boldsymbol{\beta}_j} = \left(\frac{\partial Q_2}{\partial \beta_{0j}}, \frac{\partial Q_1}{\partial \beta_{1j}}\right)^{\intercal} = 0 \tag{2.19}$$

for $j = 1, \ldots, J-1$. Since there are no analytical solutions for the equation system in Eq.(2.18) and (2.19), we use the Newton Raphson algorithm with iteratively re-weighted least squares $\big(\text{IRWLS}\big)$ instead [Lim et al., 2014] within each $(r+1)^{th}$ maximization step. The updated estimates of $\boldsymbol{\beta}$ at the $(t+1)^{th}$ IRWLS iteration are given by:

$$\beta_j^{(t+1)} = \begin{pmatrix} \beta_{j0}^{(t+1)} \\ \\ \beta_{j1}^{(t+1)} \end{pmatrix} = \big(X^{\intercal} V_j^{(t)} X\big)^{-1} X^{\intercal} V_j^{(t)} \zeta_j^{(t)} \tag{2.20}$$

where $V_j^{(t)} = diag\big(v_{1j}^{(t)}, \ldots, v_{nj}^{(t)}\big)$ with $v_{ij}^{(t)} = \pi_{ij}^{(t)}\big(1 - \pi_{ij}^{(t)}\big)$, and $\zeta_j^{(t)} = X\boldsymbol{\beta}_j^{(t)} + V_j^{(t)}\boldsymbol{z}_j^{(r+1)} - I_y {V_j^{(t)}}^{-1}\boldsymbol{z}_J^{(r+1)} = \big(\zeta_{1j}^{(t)}, \ldots, \zeta_{nj}^{(t)}\big)^\intercal$, with $I_y = diag\big(I_{[y_1=0]}, \ldots, I_{[y_n=0]}\big)$. Here $V_j^{(t)}$ is the $\big(n \times n\big)$ weight matrix and $\zeta_j^{(t)}$ is the working adjusted response vector. They are updated at each IRWLS iteration based on the updated multinomial probability parameters $\pi_{ij}^{(t)} = \frac{\exp\big(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j^{(t)}\big)}{1+\exp\big(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j^{(t)}\big)}$, for $i = 1, \ldots, n$, $j = 1, \ldots, J-1$. For the log-linear model of Eq.(2.12), the updated estimates are given by:

$$\gamma_j^{(t+1)} = \begin{pmatrix} \gamma_{j0}^{(t+1)} \\ \\ \gamma_{j1}^{(t+1)} \end{pmatrix} = \big(X^\intercal W_j^{(t)} X\big)^{-1} X^\intercal W_j^{(t)} \xi_j^{(t)} \tag{2.21}$$

where $W_j^{(t)} = diag\big(w_{1j}^{(t)}, \ldots, w_{nj}^{(t)}\big)$ with $w_{ij}^{(t)} = z_{ij}^{(t)}\theta_{ij}^{(t)}$, and $\xi_j^{(t)} = X\boldsymbol{\gamma}_j^{(t)} + \big[diag\big(\theta_j^{(t)}\big)\big]^{-1}\big(\boldsymbol{y} - \boldsymbol{\theta}_j^{(t)}\big) = \big(\xi_{ij}^{(t)}, \ldots, \xi_{nj}^{(t)}\big)^\intercal$ with $\theta_{ij}^{(t)} = \exp\big(\boldsymbol{x}_i^\intercal \boldsymbol{\gamma}_j^{(t)}\big)$. Similarly, $W_j^{(t)}$ and $\xi_j^{(t)}$ are the weighted matrix and the working adjusted response vector that are updated at each IRWLS iteration based on the updated Poisson mean parameter $\theta_{ij}^{(t)}$, for $i = 1, \ldots, n$, $j = 1, \ldots, J-1$. In R [Team, 2014], the IRWLS estimate of $\boldsymbol{\beta}$ can be obtained using function `rmultinom` from the `nnet` package [Venables and Ripley, 2013], and $\boldsymbol{\gamma}$ is estimated using function `glm`, and so, the MLEs $\hat{\boldsymbol{\beta}}^{(r+1)}$ and $\hat{\boldsymbol{\gamma}}^{(r+1)}$ are obtained in the $\big(r+1\big)^{th}$ maximization-step.

The Aitken acceleration-based stopping criterion $\big(\text{ACC}\big)$ determines the stationary point in an EM algorithm and is used to stop the EM algorithm. This criterion is applied on a sequence of log-likelihood and suggest to stop the EM algorithm when the absolute difference of Aitken accelerated estimates l_A between the k^{th} and $k+1^{th}$ iteration is less than a desired tolerance [Lindsay, 1995] as:

$$\mid l_A^{(k+1)} - l_A^{(k)} \mid < tol \tag{2.22}$$

where the Aitken accelerated estimate of the log-likelihood is given by [Böhning et al., 1994]:

$$l_A^{(k+1)} = l^{(k+1)} + \frac{l^{(k+1)} - l^{(k)}}{1 - c^{(k)}} \tag{2.23}$$

and

$$c^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}} \tag{2.24}$$

Here, $l^{(k)}$ denotes the value of the incomplete likelihood computed using $\boldsymbol{\vartheta}^{(k)}$ at the k^{th} iteration. Thus, the ACC uses the parameter estimate of at least four iterations to determine the convergence of the incomplete log-likeihood, which would be a more preferable choice. We set $tol = 1 \times 10^{-4}$ in this thesis.

In the EM algorithm, the way in which the initial values are specified may influence the process of the EM algorithm [Ng et al., 2012]. If a poor starting point is selected, the EM algorithm may converge very slowly. On the other hand, the EM algorithm may converge quickly if the initial value is already very close to the true value. In order to make simulation studies precise, we set a value that is very different from the true value of the parameters as the initial value.

Chapter 3

Simulation Study

In this chapter, a simulation study is used to evaluate whether or not the null distribution of the likelihood ratio statistic converges to the assumed χ^2 reference distribution with the appropriate degrees of freedom. We select three components of the GZIP regression mixture model to model the relative abundance of bacterial counts as an example, in order to demonstrate the process used to generate a sample. Simulations of the three scenarios : two-component ZIP regression model, three-component and four-component GZIP regression mixture model, are conducted and both the type I error rate and statistical power of each scenario are assessed to evaluate the proposed test for the association between a given bacteria and the disease outcome of interest.

3.1 Simulation Model

A case-control study design is carried out in the simulation study. Suppose we generate a sample of size n individuals in which about 50% of individuals are in the case $\big($disease$\big)$ group and 50% of individuals are in the control $\big($healthy$\big)$ group. Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\intercal$ be a vector representing the disease status that $x_i = 1$ if individual i is in the disease group or $x_i = 0$ otherwise. So we generate x_i from Bernoulli$\big(0.5\big)$. Given the disease status for each individual, we generate the bacterial counts according to the three-component GZIP regression mixture model that:

$$\pi_{i1} = \frac{\exp\big(\beta_{10} + \beta_{11}x_i\big)}{1 + \sum_{j=1}^{2}\exp\big(\beta_{j0} + \beta_{j1}x_i\big)} \qquad \theta_{i1} = \exp\big(\gamma_{10} + \gamma_{11}x_i\big)$$

$$\pi_{i2} = \frac{\exp\big(\beta_{20} + \beta_{21}x_i\big)}{1 + \sum_{j=1}^{2}\exp\big(\beta_{j0} + \beta_{j1}x_i\big)} \qquad \theta_{i1} = \exp\big(\gamma_{20} + \gamma_{21}x_i\big)$$

$$\pi_{i3} = 1 - \pi_{i1} - \pi_{i2}$$

We set $\boldsymbol{\beta}'s$ and $\boldsymbol{\gamma}'s$ at different values such that we obtain different sets of expected marginal mixing proportions of $(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3)^\intercal$ and expected marginal Poisson mean parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Note that when we set $\beta_{j1} = 0$ and $\gamma_{j1} = 0$, for all $j = 1, 2$, the distribution of the bacterial counts no longer depends on the disease status. That is, we generate the bacterial count data for each individual under the null hypothesis that there is no association between the bacteria under investigation and the disease status. For example, we set $\boldsymbol{\beta}_j = \big(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}\big)^\intercal = \big(-1.83, 0, -3.24, 0\big)^\intercal$. We have $\big(\pi_{i1}, \pi_{i2}, \pi_{i3}\big)^\intercal = \big(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3\big)^\intercal = \big(0.14, 0.03, 0.83\big)^\intercal$. We set $\boldsymbol{\gamma}_j = \big(\gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}\big)^\intercal = \big(1.4, 0, 3.53, 0\big)^\intercal$. We have $\big(\theta_{i1}, \theta_{i2}\big)^\intercal = \big(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\big)^\intercal = \big(4, 34\big)^\intercal$.

On the other hand, when the bacteria under investigation is associated with the disease status, we set $\beta_{j1} \neq 0$ and $\gamma_{j1} \neq 0$. For example, we set $\boldsymbol{\beta}_j = \big(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}\big)^\intercal = \big(-1.83, 0.51, -3.24, 0.26\big)^\intercal$. We have $\boldsymbol{\pi}_i = \big(\pi_{i1}, \pi_{i2}, \pi_{i3}\big)^\intercal = \big(0.2028, 0.0382, 0.759\big)^\intercal$ when $x_i = 1$ and $\boldsymbol{\pi}_i = \big(\pi_{i1}, \pi_{i2}, \pi_{i3}\big)^\intercal = \big(0.14, 0.03, 0.83\big)^\intercal$ when $x_i = 0$, and such that we have the marginal expected $\boldsymbol{\pi}_i = \big(\pi_{i1}, \pi_{i2}, \pi_{i3}\big)^\intercal = \big(0.1714, 0.0341, 0.7945\big)^\intercal$. If we set $\boldsymbol{\gamma}_j = \big(\gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}\big)^\intercal = \big(1.4, -0.29, 3.53, -0.1\big)^\intercal$. We have $\boldsymbol{\theta}_i = \big(\theta_{i1}, \theta_{i2}\big)^\intercal = \big(3, 30\big)^\intercal$ when $x_i = 1$ and $\boldsymbol{\theta}_i = \big(\theta_{i1}, \theta_{i2}\big)^\intercal = \big(4, 34\big)^\intercal$ when $x_i = 0$, and such that we have the marginal expected $\boldsymbol{\theta}_i = \big(\theta_{i1}, \theta_{i2}\big)^\intercal = \big(3.5, 32\big)^\intercal$.

Suppose given a set of $\boldsymbol{\vartheta} = \big(\boldsymbol{\beta}^\intercal, \boldsymbol{\gamma}^\intercal\big)^\intercal$ and x_i, we obtain $\boldsymbol{\pi}_i$ and $\boldsymbol{\theta}_i$. Given $\boldsymbol{\pi}_i$, we generate $\boldsymbol{z}_i = \big(z_{i1}, z_{i2}\big)^\intercal$ from Multinomial $\big(1, \boldsymbol{\pi}_i\big)$, where $\boldsymbol{\pi}_i = \big(\pi_{i1}, \pi_{i2}\big)^\intercal$ for the membership of which component the individual i belongs to. For example, if $\boldsymbol{z}_i = \big(1, 0\big)^\intercal$, individual i belongs to the 1^{st} component and then, the bacterial count y_i is generated according to Poisson $\big(\theta_{i1}\big)$. If $\boldsymbol{z}_i = \big(0, 0\big)^\intercal$, individual i belongs to the 3^{rd} component that the bacterial count $y_i = 0$. So our procedure generates paired data $\big(x_i, y_i\big)^\intercal$ for each individual i based on the given parameter values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Each simulated data set $\big(\boldsymbol{x}, \boldsymbol{y}\big)$ generated under the three-component mixture model will be fitted by the three-component mixture model via the EM algorithm under the null hypothesis and under the alternative hypothesis. The likelihood ratio test statistic, Λ, is calculated using the incomplete likelihoods. The null hypothesis will be rejected if the LRT statistic Λ is greater than the $\big(1 - \alpha\big)$ quantile of the $\chi^2_{2(J-1)} \equiv \chi^2_4$ distribution where α will be specify.

For type I error assessment, we set n=100 and 300, and consider different sets

of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}'s$ with β_{j1} and γ_{j1} are set to 0, $j = 1, 2$. For each combination, we simulate 10,000 data sets to obtain the precision of significant level at 10^{-4} under the null hypothesis. The empirical null rejection rates based on the 10,000 replicates are recorded for each combination of settings.

A similar simulation procedure is used to generate data for the power assessment. We set the sample size $n = 100$ and 300. Different sets of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ values are considered for assessing the detection power at different challenging levels. For each combination, 1,000 data sets are simulated. The empirical powers based on the 1,000 replications are recorded for each combination of settings.

As mentioned before, we also generate data set from a two-component and a four-component mixture model by using similar simulation procedure. For two-component model, we only need to specify 4 parameters $\boldsymbol{\beta} = \big(\beta_0, \beta_1\big)^{\intercal}$ and $\boldsymbol{\gamma} = \big(\gamma_0, \gamma_1\big)^{\intercal}$ for computing $\boldsymbol{\pi}_i$ and $\boldsymbol{\theta}_i$, and $\boldsymbol{z}_i \sim Multinomial\big(1, \boldsymbol{\pi}_i\big)$. We reject the null hypothesis at the α-level of significance if the LRT statistic Λ is greater than the $\big(1-\alpha\big)^{th}$ upper quantile of the χ^2_2 distribution. For the four component mixture model, we need to specify 12 parameters: $\boldsymbol{\beta} = \big(\beta_{10}, \beta_{11}, \ldots, \beta_{30}, \beta_{31}\big)^{\intercal}$, $\boldsymbol{\gamma} = \big(\gamma_{10}, \gamma_{11}, \ldots, \gamma_{30}, \gamma_{31}\big)^{\intercal}$, and $\boldsymbol{z}_i = \big(z_{i1}, z_{i2}, z_{i3}\big) \sim Multinomial\big(1, \boldsymbol{\pi}_i\big)$. At the α-level, the null hypothesis is rejected according to the threshold of $\chi^2_{6,1-\alpha}$ value.

3.2 Simulation Results

3.2.1 Type I Error Assessment

The results of the empirical null rejection rates for assessing the type I error rate based on the 10,000 replicates for each combination of parameter and sample size settings are presented in Tables 3.1, 3.2 and 3.3. Each table contains information regarding the values of parameters, sample sizes and null rejection rates. Note that the choices of parameter values are based on the model fitting on the real bacterial count data, so that the distributions of the simulated count data are closed to distribution of the real observed bacteria count data. The result for the test with a two-component model show that the empirical type I error rates are close to the nominal level when $\alpha = 0.05$. The mean and standard deviation of the empirical type I error for sample size of 100 $\big(0.0531$ and 3.37×10^{-3} respectively$\big)$ are slightly greater than the mean and the standard deviation for sample size of 300 $\big(0.0514$, and $1.78 \times 10^{-3}\big)$.

Among the three tables, Table 3.1 displays the steadiest and most accurate empirical type I error rates for the tests within the two-component model. The difference between the largest and the smallest error rate on Table 3.1 is smaller $\big(0.0107\big)$ than on either of the other two tables. Also, the mean of the type I error rate is 0.052. Sample size dose not affect the accuracy of rejecting the null hypothesis as demonstrated by the means $\big($and the standard deviations$\big)$ of type I error rates with different sample sizes $\big($n=100, 300$\big)$ being 0.053 $\big(3.37 \times 10^{-3}\big)$ and 0.051 $\big(1.78 \times 10^{-3}\big)$. In two-component scenario, the results suggest that the type I error rate is close to

the nominal level and the χ^2_2 distribution approximated the distribution of the LRT statistic under the null hypothesis well.

Tables 3.2 and 3.3 are results based on the three-component and four-component models. As more components are included in a sample with the fixed sample size, particularly when there are more than one Poisson component in the sample, the model fittings are expected to be more challenging. In studies 4, 6 and 7 in Table 3.2 and study 5 in Table 3.3, parameter values are set to represent the special case. In these studies, all the values for the type I error rate with a sample size of 100 are about 0.035, which is lower than nominal level. Furthermore, the type I error rate increases when the sample size is increased. This move can be explained because individuals from component one that their count data are generated from a Poisson distribution with mean of 5.3, and 7% of individuals from components two that their count data are generated from Poisson with mean of 32.1, and the reminder of 83% individuals all have counts of zero. Because 83% of the samples do not need to be identified, the rest of the sample can be matched to its corresponding components easily, making it is easier to reject the null hypothesis correctly. Overall, the χ^2 distribution is generally appropriate to be used to approximate the LRT statistics under the null.

3.2.2 Statistical Power Assessment

The results of the null rejection rates for assessing the statistical power of the proposed method are presented in Tables 3.4, 3.5 and 3.6 for the hypothesis tests within the two-component, three-component, and four-component models. The re-

sults are based on 1000 replications with a sample size of 100 and 300 for each study that represents a different combination of parameter values.

Based on the results in all three tables, as expected, the power for a sample size of 100 is less than the power with sample size 300 in each study. Similarly, the the number of components in a given sample with fixed sample size will influence the statistical power of a test. If the sample size is fixed, as the number of components increases, less sample size and thus less information will be allocated to each component. Among the three tables, the test within the four-component model (Table 3.6) has less power. In the study 7 of Table 3.6 our simulation model mimics a scenario where there is a component that has an extremely low mixing proportion in the given population, says 3%. If we set the sample size to 100, only 3 individuals are sampled from this component, which would pose challenges to detect the different distributions of count data between the case group and control group. The resulting power is 0.471 for sample size of 100. Increasing the sample size provides more information that can be used for statistical inference and thus a high power of the test is expected. As a result, the power is 0.991 for sample size of 300.

Table 3.1: Empirical Type I error rate assessment based on 10000 simulations in each study for a two-component ZIP regression model.

Study	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	**Sample Size**	**Null Rejection Rate**			
	β_{10}	γ_{10}		$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Study 1	-2.22	2.04	n=100	0.0014	0.0118	0.0598	0.1145
			n=300	0.0007	0.0106	0.0525	0.1054
Study 2	-1.89	2.72	n=100	0.0008	0.0112	0.0515	0.1045
			n=300	0.0011	0.0083	0.0509	0.1002
Study 3	-1.63	2.31	n=100	0.0008	0.0113	0.0544	0.1082
			n=300	0.0012	0.0101	0.0491	0.102
Study 4	-1.5	1.56	n=100	0.0015	0.0125	0.0567	0.1061
			n=300	0.0013	0.0104	0.0495	0.0964
Study 5	-1.41	2.22	n=100	0.0014	0.0093	0.0511	0.1031
			n=300	0.0008	0.0112	0.0531	0.1065
Study 6	-0.94	1.58	n=100	0.0012	0.0119	0.0535	0.1068
			n=300	0.0017	0.0098	0.0516	0.1039
Study 7	-0.79	2.93	n=100	0.001	0.0094	0.0494	0.1023
			n=300	0.001	0.0102	0.0503	0.0992
Study 8	0.84	2.85	n=100	0.0004	0.0107	0.0503	0.1023
			n=300	0.0012	0.0091	0.0513	0.1008
Study 9	1.42	3.87	n=100	0.0011	0.0102	0.0513	0.1018
			n=300	0.0021	0.0113	0.0547	0.1026

Table 3.2: Empirical Type I error rate assessment based on 10000 simulations in each study for a three-component GZIP regression model.

Study	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	**Sample Size**	**Null Rejection Rate**			
	(β_{10}, β_{20})	$(\gamma_{10}, \gamma_{20})$		$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Study 1	(-2.13, -2.55)	(1.67, 3.47)	n=100	0.0014	0.0072	0.0383	0.0762
			n=300	0.001	0.0084	0.0402	0.0833
Study 2	(-1.83, -3.24)	(1.4, 3.53)	n=100	0.0007	0.0096	0.044	0.0916
			n=300	0.0007	0.0095	0.0494	0.0933
Study 3	(-1.65, -2.45)	(0.33, 2.3)	n=100	0.0004	0.008	0.0523	0.1093
			n=300	0.0015	0.0127	0.0573	0.1069
Study 4	(-1, -3.1)	(1.19, 4.63)	n=100	0.0005	0.0075	0.0432	0.0936
			n=300	0.0009	0.0118	0.0524	0.1021
Study 5	(-1, -2.1)	(2.1, 4.5)	n=100	0.0006	0.0067	0.0357	0.0731
			n=300	0.0008	0.0091	0.0472	0.0932
Study 6	(-0.54, -2.12)	(-0.34, 1.73)	n=100	0.0003	0.0043	0.0357	0.0808
			n=300	0.0009	0.0094	0.0483	0.1009
Study 7	(0.3, -0.36)	(1.2, 3.79)	n=100	0.0006	0.0113	0.0586	0.11
			n=300	0.0011	0.0107	0.0516	0.0994
Study 8	(0.54, -0.49)	(2.05, 3.77)	n=100	0.0014	0.0084	0.0403	0.082
			n=300	0.0007	0.0097	0.0507	0.1051
Study 9	(0.55, -0.37)	(2.14, 4.17)	n=100	0.0008	0.0082	0.0466	0.0868
			n=300	0.0012	0.0093	0.0493	0.098

Table 3.3: Empirical Type I error rate assessment based on 10000 simulations in each study for a four-component GZIP regression model.

Study	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	**Sample Size**	**Null Rejection Rate**			
	$(\beta_{10}, \beta_{20}, \beta_{30})$	$(\gamma_{10}, \gamma_{20}, \gamma_{30})$		$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Study 1	(-1.89, -1.32, -1.87)	(1.22, 2.72, 4.2)	n=100	0.0017	0.0115	0.051	0.1021
			n=300	0.0019	0.0105	0.051	0.1076
Study 2	(-1.63, -1.36, -1.67)	(0.8, 2.53, 3.71)	n=100	0.0011	0.013	0.0568	0.1103
			n=300	0.0015	0.0115	0.0532	0.1071
Study 3	(-1.21, -1.5, -2.78)	(1.17, 2.62, 3.36)	n=100	0.0007	0.0085	0.0446	0.0883
			n=300	0.0008	0.0111	0.0496	0.0942
Study 4	(-1.15, -1.89, -2.57)	(1.15, 2.61, 3.93)	n=100	0.001	0.0113	0.0537	0.1069
			n=300	0.0016	0.0108	0.0532	0.1082
Study 5	(-1, -2, -1.82)	(1.82, 3.45, 4.68)	n=100	0.0174	0.024	0.0433	0.0598
			n=300	0.0011	0.007	0.037	0.0752
Study 6	(-0.13, -1.11, -2.16)	(0.6, 2, 3.1)	n=100	0.0003	0.0069	0.0461	0.0968
			n=300	0.0014	0.0123	0.0542	0.1124
Study 7	(-0.12, -0.44, -0.67)	(1.7, 3.2, 5.2)	n=100	0.0012	0.0098	0.0487	0.098
			n=300	0.0016	0.011	0.0531	0.1081
Study 8	(-0.1, -0.11, -0.9)	(1.9, 3.3, 4.9)	n=100	0.0006	0.0093	0.0476	0.091
			n=300	0.0007	0.0092	0.0525	0.102

Table 3.4: The power comparisons for different sample size based on 1000 replications in different studies within a two-components ZIP regression model.

Study	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	**Sample Size**	**Power**			
	(β_{10}, β_{11})	$(\gamma_{10}, \gamma_{11})$		$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Study 1	(-3.38, 1.82)	(2.86, 0.55)	n=100	0.563	0.828	0.956	0.981
			n=300	0.999	1	1	1
Study 2	(-2.42, 0.98)	(2.84, 0.54)	n=100	0.799	0.952	0.982	0.99
			n=300	1	1	1	1
Study 3	(-2.42, 0.38)	(3.3, -0.45)	n=100	0.355	0.607	0.807	0.882
			n=300	0.97	0.994	1	1
Study 4	(-2.22, -0.58)	(3.28, -0.74)	n=100	0.539	0.699	0.802	0.837
			n=300	1	1	1	1
Study 5	(-2.22, 0.18)	(4.35, -0.31)	n=100	0.603	0.789	0.907	0.94
			n=300	1	1	1	1
Study 6	(-2.21, 0.77)	(1.57, 1.24)	n=100	0.985	0.997	0.999	0.999
			n=300	1	1	1	1
Study 7	(-1.75, 0.19)	(3.17, 0.38)	n=100	0.643	0.847	0.945	0.969
			n=300	0.997	1	1	1
Study 8	(-1.12, 0.65)	(1.78, 0.36)	n=100	0.278	0.555	0.776	0.865
			n=300	0.959	0.988	0.998	0.998
Study 9	(-0.57, -0.98)	(2.58, -1.03)	n=100	0.998	1	1	1
			n=300	1	1	1	1
Study 10	(0.57, 0.53)	(3.19, 0.2)	n=100	0.802	0.934	0.987	0.994
			n=300	1	1	1	1

Table 3.5: The power comparisons for different sample size based on 1000 replications in different studies within a three-components GZIP regression model.

Study	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	**Sample Size**	**Power**			
	$(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21})$	$(\gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21})$		$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Study 1	(-1.53, 1.24, -2.77, 1.47)	(0.71, -2.96, 3.4, -1.34)	n=100	0.754	0.898	0.952	0.977
			n=300	1	1	1	1
Study 2	(-1.42, -0.59, -3.88, 1.52)	(-2.46, 3.2, 3.5, 0.65)	n=100	0.465	0.673	0.847	0.909
			n=300	1	1	1	1
Study 3	(-1, 0.52, -1.44, -0.32)	(3.13, 0.37, 5.39, 0.14)	n=100	0.721	0.755	0.794	0.823
			n=300	0.771	0.794	0.815	0.828
Study 4	(-1, 1.1, -3.1, -0.74)	(1.19, -0.33, 4.63, -0.34)	n=100	0.209	0.426	0.646	0.742
			n=300	0.929	0.993	0.998	0.999
Study 5	(-0.86, -0.43, -1.69, -0.86)	(1, -1.17, 3.56, -1.14)	n=100	0.92	0.976	0.994	0.999
			n=300	1	1	1	1
Study 6	(-0.64, -0.32, -2.23, -0.24)	(0.89, -1.3, 3.62, 0.14)	n=100	0.287	0.601	0.834	0.914
			n=300	0.986	0.999	0.999	0.999
Study 7	(-0.54, 0.12, -2.12, -0.12)	(-0.34, 1.1, 1.73, 0.7)	n=100	0.073	0.242	0.537	0.657
			n=300	0.797	0.926	0.981	0.994
Study 8	(0.21, -1.1, -1.52, -1.35)	(0.69, -0.75, 2.9, 0.15)	n=100	0.47	0.742	0.893	0.947
			n=300	0.997	1	1	1
Study 9	(0.24, 0.39, -0.69, -1.04)	(3.26, -0.12, 5.13, 0.43)	n=100	0.81	0.837	0.861	0.876
			n=300	0.906	0.908	0.912	0.916
Study 10	(0.47, -1.82, -1.1, -1.45)	(1, -0.18, 3, 0.24)	n=100	0.678	0.88	0.951	0.974
			n=300	1	1	1	1

Table 3.6: The power comparisons for different sample size based on 1000 replications in different studies within a four-components GZIP regression model.

Study	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$	**Sample Size**	**Power**			
	$(\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}, \beta_{31})$	$(\gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21}, \gamma_{30}, \gamma_{31})$		$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Study 1	(-2.53, 1.77, -2.52, 0.77, -2.81, 0.82)	(1.91, -0.85, 3.22, -0.24, 5.15, -0.74)	n=100	0.244	0.553	0.795	0.869
			n=300	0.992	0.998	1	1
Study 2	(-2.12, 0.43, -2.51, 0.69, -3.2, 0.28)	(0.63, -2.57, 2.71, -0.55, 5.38, -1)	n=100	0.723	0.774	0.853	0.901
			n=300	1	1	1	1
Study 3	(-1.63, 1.37, -1.36, 0.27, -1.67, 0.8)	(0.8, -0.1, 2.53, 0.42, 3.71, 0.77)	n=100	0.864	0.901	0.918	0.927
			n=300	0.992	0.992	0.992	0.992
Study 4	(-1.21, 1.33, -1.5, 1, -2.78, 1.52)	(1.17, -0.7, 2.62, -0.33, 3.36, -0.33)	n=100	0.146	0.397	0.649	0.791
			n=300	0.903	0.963	0.987	0.994
Study 5	(-1.15, 0.35, -1.89, 1.57, -2.57, 0.87)	(1.15, 0.77, 2.61, 0.76, 3.93, 0.98)	n=100	0.809	0.818	0.822	0.826
			n=300	0.961	0.961	0.961	0.961
Study 6	(-1, 0.78, -2, 1.62, -1.82, 0.86)	(1.82, -1.53, 3.45, -1.1, 4.68, -0.4)	n=100	0.697	0.7332	0.75	0.756
			n=300	0.866	0.868	0.87	0.871
Study 7	(-0.95, 0.96, -1, 0.92, -1.85, 1.6)	(1.53, -1.9, 3.05, -0.9, 4.27, -0.71)	n=100	0.815	0.833	0.848	0.851
			n=300	0.972	0.973	0.974	0.974
Study 8	(-0.34, 0.7, -2.1, 1, -2.77, -0.14)	(0.3, -1.93, 2.84, -1.47, 3.73, 1.1)	n=100	0.434	0.589	0.732	0.797
			n=300	0.702	0.761	0.791	0.802
Study 9	(-0.13, 2.52, -1.11, 1.66, -2.16, -2.89)	(0.6, -0.91, 2, 0.16, 3.1, -0.91)	n=100	0.053	0.235	0.471	0.647
			n=300	0.846	0.962	0.991	0.995
Study 10	(0.32, -2.1, -1.1, -1.48, -2.38, -1.33)	(-1.24, 0.13, 1.75, -0.53, 4.63, -0.34)	n=100	0.117	0.293	0.519	0.644
			n=300	0.9	0.976	0.998	0.998

Chapter 4

Application

In this chapter, we will apply our proposed method to analyze the gut microbiome data from the colorectal cancer study conducted by Nakatsu et al. [2015]. We present the background of the data set and the study and then briefly describe the model selection procedure for the choice of using different number of mixture components to different bacteria. At the end, we present the result of using the likelihood ratio test for detecting differentially abundant bacteria between the case and control groups.

4.1 Data Description

Colorectal cancer has been shown to be associated with gut microbial dysbiosis [Nakatsu et al., 2015]. The study collected 160 samples of gut mucosal microbiome; of those, 61 were collected from independent healthy subjects, 47 were collected from independent adenoma subjects, and 52 were collected from independent

carcinoma subjects to investigate the gut microbiome communities at different stages of colorectal tumorigenesis. Nakatsu et al. [2015] performed 16s ribosomal RNA gene sequencing on the samples. The raw read sequence of these samples is publicly available in the Sequence Read Archive (SRA), National Center for Biotechnology Information (NCBI) database. The raw read sequence data of each individual sample is preprocessed by a pipeline using the software "mothur" [Schloss et al., 2009] and is prepared by Stephen [2017] in order to quantify the relative abundance with a typical taxonomic identity.

The main preprocessing steps can be summarized as: aligning sequences to reference genome, clustering and assembling sequences to operational taxonomic units (OTUs), assigning taxonomies to the OTUs. In the output file, each bacteria species will be assigned to a genus, a lower level of a hierarchy of taxonomy rank, then in turn to family, order, class, phylum, kingdom and domain. Because the precision of bacteria identification at the species level is known to be less reliable and unsatisfactory, our analysis will be performed at a more reliable level, the genus level. Note, a "sub-sample" function in "mothur" software is used to normalize the total bacterial counts across all samples. In order to eliminate the bacteria that has very low abundance and such that has no information to distinguish itself between the case group and control group, bacteria with a relative abundance (relative to full total bacteria count) of less than 0.001 are eliminated, resulting in a total of 85 genera remaining in the analysis.

In this thesis, we consider the group of 52 subjects with carcinoma as the case group and the group of 61 healthy subjects as the control group and leaving a total

of 113 subjects for analysis. Over 85 genera and 113 subjects, there are about 39.7% of entries with the count being equal to zero, which reflects the typical zero-inflated problem in microbiome data. In order to confirm the existence of overdispersion in the bacterial counts, we compare the sample variance and mean. The resulting overdispersion exists because the sample variance is much larger than the sample mean and the sample variance of the total counts of bacteria $\left(79253.77\right)$ is much larger than the sample mean $\left(90.7798\right)$.

The Figure 4.1 shows the overall distribution of bacterial counts of *Parabacteroides* and the distributions under the case group and control group. In the plot with overall bacterial counts, it is obvious to see that there are more than twice observations when count is zero than others. This characteristic explains that the distribution contains more zeros than expected under a Poisson distribution. In the same plot, it is also clear to find that there is more than one Poisson component in the non-zero counts because there are gaps around count when count is 100. When the count is greater than 120, the counts are distributed in a dispersed pattern. Similarly, the plots with bacterial counts under the case group and the control group also have these characteristics but they have different patterns. Thus, considering these characteristics of bacterial counts, the GZIP regression mixture model is commonly used.

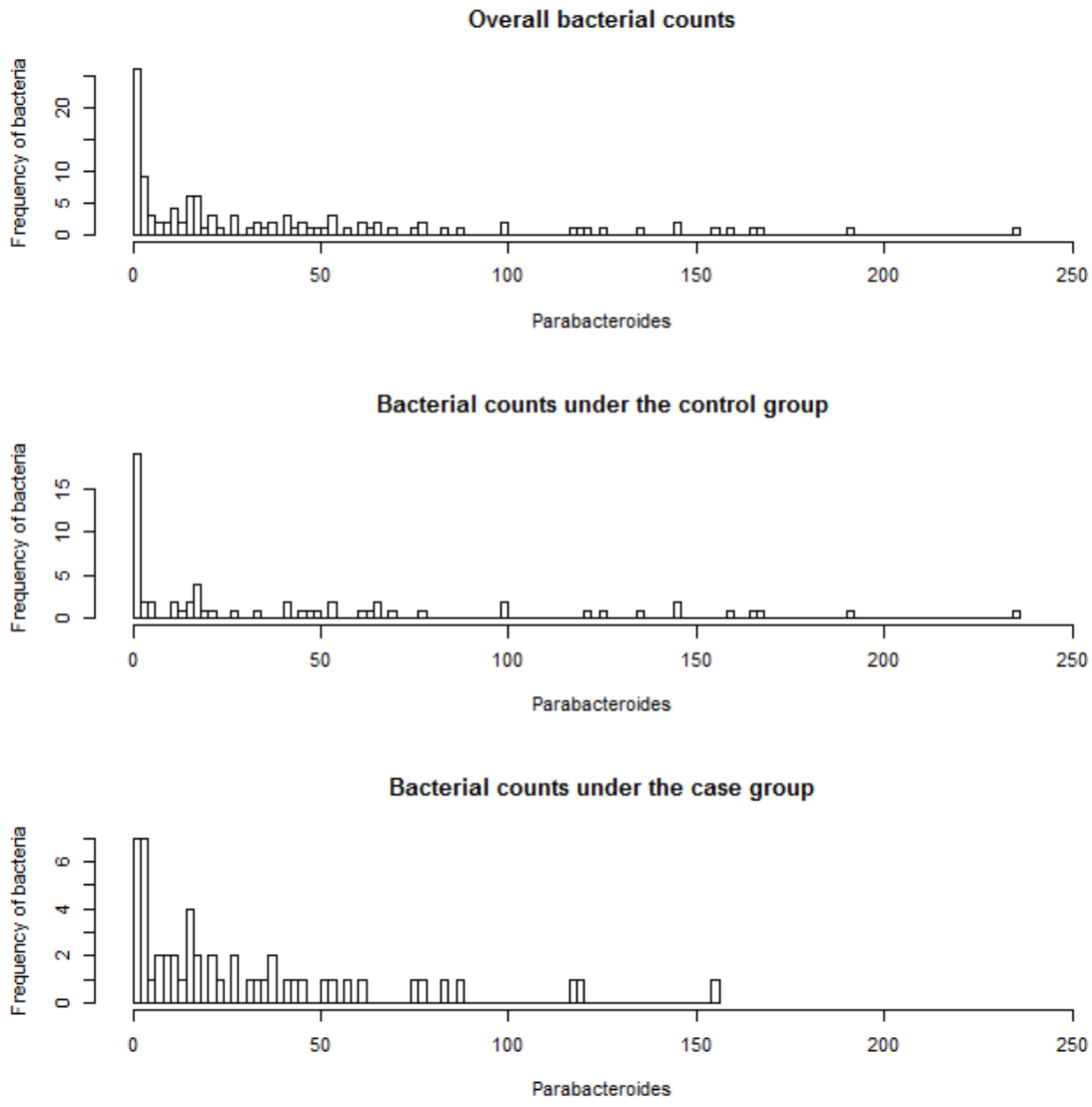

Figure 4.1: Histogram plot of *Parabacteroides* bacterial counts

4.2 Determination of the Number of Components

For each bacteria at genus level, in order to determine the suitable number of components in the sample, we fit a two-component, three-component and four-component GZIP regression mixture models using the EM algorithm. The Bayesian information

criterion (BIC) is used for model selection. The BIC is given as:

$$BIC = -2lnL_m + m \ln n$$

where L_m is the maximized log-likelihood of the model, m is the number of parameters in the model and n is the sample size. Based on the result of model selection for fitting these 39 bacteria [Siyu, 2017], there are three bacteria samples best fitted using the two-component ZIP regression model, five bacteria samples best fitted using the three-component GZIP regression mixture model, and 31 bacteria samples best fitted using the four-component GZIP regression mixture model. The rest of the 23 samples cannot be fitted using these GZIP regression mixture model and thus will not be analyzed using our method.

4.3 Results of Real Analysis

Because there are a total of 39 hypothesis tests, the Bonferroni controlling procedure, significance level of $0.05/39 = 0.00128$, is used for each individual test. The three bacteria that are fitted and tested under the two-component ZIP regression model are: *Odoribacter*, *Pedobacter*, and *Gemella*. Among these three bacteria, *Gemella* is found to be extremely significantly associated (p-value ≈ 0) with the carcinoma. There are two outlier observed OTU counts data: 789 and 817 in the *Gemella*, and the log-likelihood cannot be calculated when we fit the data under the null hypothesis with a Poisson mean of 115. The R function `dpois` (789, 115) and function `dpois` (789, 115) are equal to 0, so that the log-likelihood is equal to -Inf. However, The log-likelihood can be calculated when we fit the data under the alter-

native hypothesis with a Poisson mean 161.54. Depending on the difference between log-likelihood, we can find that the distributions of bacterial counts under the null and under the alternative are different. After removing those two extreme counts data (789, 817), we obtain the log-likelihood under the alternative (-5786.004) and under the null (-6305.844). The p-value is 0. So, in both the cases of including and excluding the two outlier, the null hypothesis is rejected. The *Pedobacter* is also found to be associated (p-value= 1.32×10^{-12}) with the carcinoma as well.

Among five bacteria that are fitted and tested under the three-component GZIP regression mixture model, four bacteria are found to be significantly associated with the carcinoma disease. For example, *Aeromonas* (p-value= 6.53×10^{-10}), *Christensenellaceae_R-7_group* (p-value= 1.55×10^{-15}), *Enterococcus* (p-value= 1.08×10^{-8}) and *Staphylococcusthe* (p-value= 7.65×10^{-11}) are found to be strongly associated with the carcinoma.

Among 31 bacteria that are fitted and tested under the four-component GZIP regression mixture model, 22 bacteria are found to be strongly associated with the carcinoma disease with their p-value ranging from 1.28×10^{-6} to 0. There are two bacteria: *Tyzzerella* (2.8×10^{-2}) and *Parasutterella* (1.27×10^{-2}) are found to be moderately significant as the p-values are less than 0.05 but greater than the adjusted significance level of 0.00128. The bacteria *Campylobacter* has undefined p-value (`NAN`). An extremely large outlier count of 1198 is observed in the sample when comparing to the others that have a range of 0 to 326 counts. This sample is failed to be grouped into any one of the three Poisson components under the null and the alternative. Because there is such an extreme outlier in the sample, the

model selection procedure does not suggest an additional component in the model for this sample only. However, when calculating the likelihood value, the probability of having such observation is nearly 0 for all Poisson components and such that the log-likelihood has -infinity value by using the R. After removing those the extreme count data $\left(1198\right)$, we obtain the log-likelihood under the alternative $\left(\text{-}325.8731\right)$ and under the null $\left(\text{-}322.1846\right)$ and the p-value is 1. This suggests that the bacteria *Campylobacter* is not associated with the carcinoma.

Table 4.1: Bacteria found to be associated with carcinoma

Two-component model		
Bacteria		Significance level
Genus	Phylum	(P-value)
Gemella[1]	Firmicutes	0
Odoribacter	Bacteroidetes	1.36×10^{-2}
Pedobacter	Bacteroidetes	1.32×10^{-12}
Three-component model		
Aeromonas[3]	Proteobacteria	6.53×10^{-10}
Christensenellaceae_R-7_group	Firmicutes	1.55×10^{-15}
Enterococcus[2]	Firmicutes	1.08×10^{-8}
Eggerthella[2]	Actinobacteria	2.05×10^{-3}
Staphylococcus	Firmicutes	7.65×10^{-11}
Four-component model		
Anaerococcus	Firmicutes	1.52×10^{-8}
Collinsella	Actinobacteria	1.78×10^{-15}
Coprococcus_1	Firmicutes	6.3×10^{-7}
Coprococcus_3	Firmicutes	1.15×10^{-11}
Campylobacter	Proteobacteria	NAN
Dialister[3]	Firmicutes	9.99×10^{-10}
[Eubacterium]_coprostanoligenes_group	Firmicutes	1.62×10^{-7}
[Eubacterium]_hallii_group	Firmicutes	1
Fusicatenibacter	Firmicutes	2.44×10^{-15}
Intestinibacter	Firmicutes	3.02×10^{-14}
Lachnoclostridium	Firmicutes	0
Lachnospiraceae_NK4A136_group	Firmicutes	1
Lactobacillus[3]	Firmicutes	6.86×10^{-14}
Mogibacterium[3]	Firmicutes	0
Neisseria	Proteobacteria	9.61×10^{-11}
Parabacteroides	Bacteroidetes	6.7×10^{-10}
Paraprevotella	Bacteroidetes	4.13×10^{-10}
Peptostreptococcus[1]	Firmicutes	0
Pseudobutyrivibrio[3]	Firmicutes	0
Peptoniphilus	Firmicutes	2.35×10^{-14}
Parasutterella	Proteobacteria	1.27×10^{-2}
Pseudomonas	Proteobacteria	1.28×10^{-6}
Pseudoflavonifractor	Firmicutes	1
Romboutsia	Firmicutes	1
Roseburia	Firmicutes	0
Ruminococcaceae_UCG-002	Firmicutes	9.22×10^{-11}
Ruminococcaceae_UCG-005	Firmicutes	1
Ruminococcaceae_UCG-008	Firmicutes	1
Ruminococcaceae_UCG-014	Firmicutes	2.27×10^{-7}
Subdoligranulum	Firmicutes	9.99×10^{-15}
Tyzzerella	Firmicutes	2.8×10^{-2}

1. *Agreement with* Nakatsu et al. [2015], 2.*Agreement with* Wang et al. [2012]

3. *Agreement with* Chen et al. [2012]

Chapter 5

Conclusion and Future work

The purpose of this thesis is to develop a general framework for testing associations between the abundance of a given bacteria in microbial communities and the outcome of interest (e.g. disease status). This framework enables us to differentiate the distributions of the abundance of bacteria with different conditional groups. In this thesis, the most commonly found three different mixture models based on the GZIP regression were covered to address the issues of concerning zero-inflated heterogeneous count data observed in the sample. When fitting a mixture model, an EM algorithm is typically used when the component membership for each individual is unknown. The parameter estimates were provided by the EM algorithm for all three models.

Our proposed method is validated and its performance is evaluated by simulation study. The simulation study results show that, under the null hypothesis, the χ^2 distribution generally approximates the distribution of likelihood ratio test statis-

tic Λ developed under the GZIP regression mixture model very well. However, a large sample size is desired to provide sufficient information for model fitting when the number of components increases. When the sample size is small, the number of components of the mixture model is prohibited. To better account for the over dispersion pattern within each Poisson component, it is worth pursuing a generalized zero-inflated negative binomial (GZINB) regression mixture model for model fitting and a consequently an association test under this framework.

We applied our method to a study of gut mucosal microbiome communities for different groups of people with respect to their cancer status. One of the objective of the study is to identify bacteria at genus rank that have differentiated the distribution abundance between individuals in the healthy group and the carcinoma group. In the association analysis of the microbiome data of Nakatsu et al. [2015], *Gemella* and *Peptostreptococcus* are found to be associated with the carcinoma, which is also found by Nakatsu et al. [2015]. Wang et al. [2012] mentioned that *Enterococcus*, *Peptostreptococcus*, *Eggerthella* and *Gemella* exhibited a relatively higher abundance in the gut microbiota of carcinoma patients. *Pseudobutyrivibrio*, *Lactobacillus*, *Peptostreptococcus*, *Gemella*, *Mogibacterium*, *Dialister* and *Aeromonas* are enriched in carcinoma-associated candidates [Chen et al., 2012]. The result from our association analysis also suggested these findings too.

Chapter 6

Appendix

```
library(LaplacesDemon)
library(nnet)

dslnex=function(x,b1,b2,b3,b4,r1,r2,r3,r4,n){

  x=rbern(n,0.5)
  dx=1+exp(b1+b2*x)+exp(b3+b4*x)
  p1=exp(b1+b2*x)/dx
  p2=exp(b3+b4*x)/dx
  p3=1-p1-p2
  pi=cbind(p1,p2,p3)

  t1=exp(r1+r2*x)
  t2=exp(r3+r4*x)
  t3=rep(0,n)
  t=cbind(t1,t2,t3)

```

```
18
19    z=t(apply(pi,1,function(w) rmultinom(1,1,w)))
20    T=cbind(rpois(n, t[,1]),rpois(n, t[,2]),0)
21
22
23    y=mapply(function(x, y) t(x)%*%y, split(T,row(T)),split(z,row(z)))
24
25
26    count=cbind(x,y)
27
28    return(count)
29  }
30
31
32
33
34  ntestm=function(x,new_y,index,b01,b11,b02,b12,r01,r11,r02,r12,n){
35
36    p1=exp(b01+b11*x)/(1+exp(b01+b11*x)+exp(b02+b12*x))
37    p2=exp(b02+b12*x)/(1+exp(b01+b11*x)+exp(b02+b12*x))
38    p3=1-p1-p2
39
40    t1=exp(r01+r11*x)
41    t2=exp(r02+r12*x)
42    id=c(rep(0,index), rep(1,(n-index)))
43    mat=cbind(p1,p2,p3,t1,t2,id)
44    #print(mat)
45    result <- lapply(by(mat,mat[,6],identity),as.matrix)
```

```
  #print(result)
  a= result[[1]][,1]
  b= result[[1]][,2]
  c= result[[1]][,3]
  g= result[[1]][,4]
  h= result[[1]][,5]
  #print(a)
  d=result[[2]][,1]
  e=result[[2]][,2]
  m=result[[2]][,4]
  n=result[[2]][,5]
  #print(m)
  z1=(a*exp(-g))/(a*exp(-g)+b*exp(-h)+c)
  z2=(b*exp(-h))/(a*exp(-g)+b*exp(-h)+c)
  ztest1=cbind(z1,z2,1-z1-z2)

  z11=(d*dpois(new_y,m))/(d*dpois(new_y,m)+e*dpois(new_y,n))
  ztest2=cbind(z11,1-z11,0)

  zhat=rbind(ztest1,ztest2)
  return(zhat)
}

licn=function(x,new_y,index,b01,b11,b02,b12,r01,r11,r02,r12,n){
  l=c()
  dx=1+exp(b01+b11*x)+exp(b02+b12*x)
  p1=exp(b01+b11*x)/dx
  p2=exp(b02+b12*x)/dx
```

```
  p3=1-p1-p2

  t1=exp(r01+r11*x)
  t2=exp(r02+r12*x)

  id=c(rep(0,index), rep(1,(n-index)))
  mat=cbind(p1,p2,p3,t1,t2,id)
  #print(mat)
  result <- lapply(by(mat,mat[,6],identity),as.matrix)
  #print(result)
  a= result[[1]][,1]
  b= result[[1]][,2]
  c= result[[1]][,3]
  g= result[[1]][,4]
  h= result[[1]][,5]
  #print(a)
  d=result[[2]][,1]
  e=result[[2]][,2]
  m=result[[2]][,4]
  n=result[[2]][,5]

  l1=log(a*exp(-g)+b*exp(-h)+c )
  l2=log(d*dpois(new_y,m)+e*dpois(new_y,n))
  L=sum(l1,l2)
  return(L)
}

AAC=function(L){
```

```
  if (length(L)<4){
    val=1
    return(val)
  } else {
    # b is the length of L
    b=length(L)
    Ck_1=(L[b]-L[b-1])/(L[b-1]-L[b-2])
    la_k1=L[b-1]+(L[b]-L[b-1])/(1-Ck_1)

    Ck=(L[b-1]-L[b-2])/(L[b-2]-L[b-3])
    la_k=L[b-2]+(L[b-1]-L[b-2])/(1-Ck)

    val=abs(la_k1-la_k)

    if(is.nan(val)) val=0
    return(val)
  }
}

EM_func <- function(x,new_y,y,index,b01,b11,b02,b12,r01,r11,r02,r12,n) {
  while (val>1e-04){
    #k=k+1
    #print(k)
    z_star=ntestm(x,new_y,index,b01,b11,b02,b12,r01,r11,r02,r12,n)

    data=data.frame(cbind(z_star,x,y))
    colnames(data) <- c("z1","z2","z3","x","y")

```

```
130     model1<-multinom(cbind(z3,z1,z2)~x, data)
131
132     model2<-glm(y~x, family=poisson, weights=z1, data)
133
134     model3<-glm(y~x, family=poisson, weights=z2, data)
135
136     b01=summary(model1)$coefficients[1,1]
137     b11=summary(model1)$coefficients[1,2]
138     b02=summary(model1)$coefficients[2,1]
139     b12=summary(model1)$coefficients[2,2]
140     r01=summary(model2)$coefficients[1,1]
141     r11= summary(model2)$coefficients[2,1]
142     r02=summary(model3)$coefficients[1,1]
143     r12= summary(model3)$coefficients[2,1]
144
145     likelihood=licn(x,new_y,index,b01,b11,b02,b12,r01,r11,r02,r12,n)
146
147     L=c(L,likelihood)
148     plot(L)
149     val=AAC(L)
150   }
151   #a=c(b01,b11,b02,b12,r01,r11,r02,r12)
152   Lihat1=licn(x,new_y,index,b01,b11,b02,b12,r01,r11,r02,r12,n)
153   #b=c(a,Lihat1)
154   return(Lihat1)
155 }
156
157
```

```
ntest3=function(bhat1,bhat2,rhat1,rhat2,index,new_y){

  phat1=exp(bhat1)/(1+exp(bhat1)+exp(bhat2))


  phat2=exp(bhat2)/(1+exp(bhat1)+exp(bhat2))

  phat3=1-phat1-phat2

  that1=exp(rhat1)

  that2=exp(rhat2)

 #a=sort(y)
 #index=length(a[y==0])


  dx=phat1*exp(-that1)+phat2*exp(-that2)+phat3
  z1=(phat1*exp(-that1))/dx
  z2=(phat2*exp(-that2))/dx
  v=c(z1,z2,1-z1-z2)
  zhat1=matrix(rep(v,each=index),nrow=index  )


 #remove.value=0
 #new_y=y[!y == remove.value]

```

```
186   z11=(phat1*dpois(new_y,that1))/(phat1*dpois(new_y,that1)+phat2*dpois(
        new_y,that2))
187
188   z12=1-z11
189
190   zhat2=cbind(z11,z12,0)
191
192
193   zhat=rbind(zhat1,zhat2)
194
195   return(zhat)
196
197 }
198
199
200 lic3=function(bhat1,bhat2,rhat1,rhat2,index,new_y){
201
202   phat1=exp(bhat1)/(1+exp(bhat1)+exp(bhat2))
203
204   phat2=exp(bhat2)/(1+exp(bhat1)+exp(bhat2))
205
206   phat3=1-phat1-phat2
207
208   that1=exp(rhat1)
209
210   that2=exp(rhat2)
211
212
```

```
  l1=log(phat1*exp(-that1)+phat2*exp(-that2)+phat3 )*index
  l2=log(phat1*dpois(new_y,that1)+phat2*dpois(new_y,that2))
  L=sum(l1,l2)
  return(L)
}


AAC2 <- function(L1) {
  if (length(L1) <4) {
    val1=1
    return(val1)
  }else {
    k= length(L1)
    cstar_1=(L1[k]-L1[k-1])/(L1[k-1]-L1[k-2])
    la=L1[k-1]+(L1[k]-L1[k-1])/(1-cstar_1)
    #print(la)
    cstar_2=(L1[k-1]-L1[k-2])/(L1[k-2]-L1[k-3])
    old_la=L1[k-2]+(L1[k-1]-L1[k-2])/(1-cstar_2)
    #print(old_la)
    val1 = abs(la-old_la)
    #print(val1)
    if (is.nan(val1)) val1=0
    return( val1 )

  }

}

```

```
241
242
243 EM_func2=function(y,bhat1,bhat2,rhat1,rhat2,index,new_y){
244
245   while (val1>1e-04) {
246     #k=k+1
247     #print(k)
248
249
250     z_star2=ntest3(bhat1,bhat2,rhat1,rhat2,index,new_y)
251
252
253     data=data.frame(cbind(z_star2,y))
254     colnames(data)=c("z1","z2","z3","y")
255     model1<-multinom(cbind(z3,z1,z2)~1, data)
256     model2<-glm(y~1, family=poisson, weights=z1, data)
257
258     model3<-glm(y~1, family=poisson, weights=z2, data)
259
260     bhat1=summary(model1)$coefficients[1,1]
261
262     bhat2=summary(model1)$coefficients[2,1]
263
264     rhat1=summary(model2)$coefficients[1,1]
265
266     rhat2=summary(model3)$coefficients[1,1]
267
268
```

```
269     likelihood3=lic3(bhat1,bhat2,rhat1,rhat2,index,new_y)
270
271     L1=c(L1,likelihood3)
272     #print(L1)
273     plot(L1)
274     val1=AAC2(L1)
275
276   }
277   #c=c(bhat1,bhat2,rhat1,rhat2)
278   #print(c)
279   Lihat0=lic3(bhat1,bhat2,rhat1,rhat2,index,new_y)
280   #d=c(Lihat0,c)
281   return(Lihat0)
282
283   #c=c(bhat1,bhat2,rhat1,rhat2)
284   #return(c)
285
286 }
287
288
289
290 for (j in 11:20){
291
292 outcome=c()
293 for (i in 1:1000){
294   b1 = -0.72
295   b2 = 0
296   b3 =-1.82
```

```
b4 = 0


r1 = 2.59
r2 = 0
r3 =  4.68
r4 = 0
n=300

simu=dslnex(x,b1,b2,b3,b4,r1,r2,r3,r4,n)
x=simu[,1]
y=sort(simu[,2])

index=length(y[y==0])



remove.value=0
new_y=y[!y == remove.value]

b01=0.1
b11=0.2
b02=-0.2
b12=0.3
r01=0.5
r11=-0.8
r02=3
r12=-0.7

L=c()
```

```
  val=1

  #k=0

  para=EM_func(x,new_y,y,index,b01,b11,b02,b12,r01,r11,r02,r12,n)



  bhat1=-2.2
  bhat2= -0.85
  rhat1=1.1
  rhat2=2.7


  L1=c()
  val1=1
  #k=0


  para2=EM_func2(y,bhat1,bhat2,rhat1,rhat2,index,new_y)
  outcome = rbind(outcome, c(para,para2))

}
LRT=2*(outcome[,1]-outcome[,2])
#LRT=2*(outcome[,9]-outcome[,10])
#print(LRT)
pvalue=1-pchisq(LRT,4,lower.tail=TRUE )
#sum(pvalue<0.05)
```

```
353 number=sum(pvalue<0.05)
354 outcome=rbind(outcome,number)
355 write.table(outcome, file = paste(c("C:/Users/Nina/Desktop/3component_
    outcome/under the null/use intercept from Ha/8/outcome",j,".txt"),
    collapse=""), row.names=FALSE, col.names=FALSE)
356 }
```

Bibliography

Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics 46*(2), 373–388.

Carroll, R. J. and D. Ruppert (1988). *Transformation and weighting in regression*, Volume 30. CRC Press.

Chen, W., F. Liu, Z. Ling, X. Tong, and C. Xiang (2012). Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PloS one 7*(6), e39743.

Collins, M. (1997). The em algorithm. *fulfillment of Written Preliminary Exam II requirement*.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Jamshidian, M. and R. I. Jennrich (1993). Conjugate gradient acceleration of the em algorithm. *Journal of the American Statistical Association 88*(421), 221–228.

Lim, H. K., W. K. Li, and L. Philip (2014). Zero-inflated poisson regression mixture model. *Computational Statistics & Data Analysis 71*, 151–158.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–163. JSTOR.

Manichanh, C., N. Borruel, F. Casellas, and F. Guarner (2012). The gut microbiota in ibd. *Nature Reviews Gastroenterology and Hepatology 9*(10), 599–608.

McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. *Computational Statistics & Data Analysis 54*(3), 711–723.

Nakatsu, G., X. Li, H. Zhou, J. Sheng, S. H. Wong, W. K. K. Wu, S. C. Ng, H. Tsoi, Y. Dong, N. Zhang, et al. (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nature communications 6*, 8727.

Ng, S. K., T. Krishnan, and G. J. McLachlan (2012). The em algorithm. In *Handbook of computational statistics*, pp. 139–172. Springer.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al. (2010). A human gut microbial gene catalog established by metagenomic sequencing. *nature 464*(7285), 59.

Qin, J., Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen,

et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature 490*(7418), 55–60.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology 75*(23), 7537–7541.

Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods 9*(8), 811–814.

Siyu, C. (2017). *Msc theis in Statistics*.

Stephen, B. (2017). Generalized linear regression model with lasso and group lasso regularization methods for predicting disease status using the microbiome data. *Msc theis in Bioinformatics*.

Team, R. C. (2014). R: A language and environment for statistical computing. vienna, austria: R foundation for statistical computing; 2014.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. I. Gordon (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature 449*(7164), 804.

Turnbaugh, P. J., R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *nature 444*(7122), 1027–131.

Van den Broek, J. (1995). A score test for zero inflation in a poisson distribution. *Biometrics*, 738–743.

Venables, W. N. and B. D. Ripley (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Wang, T., G. Cai, Y. Qiu, N. Fei, M. Zhang, X. Pang, W. Jia, S. Cai, and L. Zhao (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME journal 6*(2), 320–329.