NEGATIVE BINOMIAL MODELLING AND APPLICATIONS FOR MICROBIOME COUNT DATA

by

Chang Chen

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2016

Table of Contents

List of Tables

List of Figures

Abstract

The human microbiome plays an important role in human health and disease. Identification of factors that affect the microbiome composition will eventually allow modulation of the microbiome for therapeutic purposes. The aim of this study is to find a suitable statistics distribution model for the set of microbial operational taxonomic units (OTUs), which are used to categorize bacteria based on sequence similarity, and to use these models to analyze the supra-gingival and sub-gingival plaque microbiome. We model the OTU data with a Negative Binomial (NB) distribution and fit the maximum-likelihood estimates for the NB model parameters. We then develop a gamma-prior distribution to model the underlying composition of each OTU. We use the mean of the calculated posterior distribution as an estimator of the underlying composition of each OTU, analyzing oral cavity microbiome communities based on the posterior means. Likelihood ratio tests identified NB models for some OTUs that differed significantly between sub-gingival plaques and supra-gingival plaques. We also developed a Naïve Bayes Discriminant Analysis (NBDA) approach based on the calculated NB distributions, and performed LASSO regression on the simple proportions and the estimated underlying compositions. The NBDA and LASSO approaches identified OTUs that play a critical role in classification. By replacing simple proportions with distribution models, we explore the underlying composition of OTUs better without losing too much discriminant information.

List of Abbreviations and Symbols Used

Symbols and Abbr.	Description
OTU	Operational Taxonomic Units
HMP	Human Microbiome Project
NB	Negative Binomial
NBDA	Naive Bayes Discriminant Analysis
MLE	Maximum Likelihood Estimate
PCA	Principal component analysis
LR	Likelihood Ratio
FDR	False Discovery Rate
BH-Procedure	Benjamini and Hochberg Procedure
LASSO	Least Absolute Shrinkage and Selection Operator

Acknowledgements

I'd like to thank my supervisors Dr. Hong Gu and Dr. Robert Beiko. I built up a large debt toward them who kept up with the various versions of this thesis. Thanks to their generosity of research and discussion in the weekly meeting, I came back having taken away with inspired ideas of my thesis works. They did offer a great of help on my thesis problems associated with statistics and microbiome.

I would like to thank my thesis readers Dr. Bruce Smith and Dr. Edward Susko. For their comments and assistance in correcting technical and factual mistakes in my thesis. I would like to thank my parents and friends for their support and encouragement.

I derived an unexpected amount of enjoyment writing this thesis, and I hope that all the readers will experience the same.

Chapter 1

Introduction

1.1 Background

The human microbiome plays an important role in providing insights into disease mechanism. Identification of factors that affect the microbiome composition will allow us to modulate the microbiome composition for therapeutical purposes. There are more than 600 prevalent taxa at the species level in human oral cavity [1]. Bacterial communities have significant differences between healthy and diseased oral cavities. As they can cause or prevent infections, this may have a significant impact on fully understanding human general health.The taxonomic composition of a microbial community can provide clues to better understand its structure and ecology [2].

It has been a major concern in environmental microbiology to assess the microbial diversity and distribution [3]. But it's difficult to test the association of microbiome composition with potential environmental factors using OTU abundances. Directly, because OTU data are usually of high dimensionality, non-normality and with phylogenetic structure among the OTUs. Using simple proportions and rarefying of counts for normalization became the most common approaches [4], it's clear that both of these approaches are inappropriate for detection of differentially abundant species. The main problem of current practices in the normalization of microbiome count data is either incorrect or inefficient. Count data arise in numerous biological applications and can often be modelled by a Poisson distribution. However, Poisson distribution has the restriction that the mean and variance are the same which is usually not true in most of the applications. Most often, the observed variation is significantly greater than the mean and an extension to the Poisson model is more appropriate [5]. A popular alternative for modelling count data when the variance is larger than the mean is the Gamma-Poisson model, in which the Poisson rate parameter is a Gamma random variable with fixed coefficient of variation, also known as the Negative Binomial (NB) model [6]. Our research starts from modelling the OTU

count data by NB model.

1.2 OTU Counts Data from Human Oral Cavity

1.2.1 Human Oral Cavity

The human oral cavity plays host to many complex microbial communities. There are more than 600 prevalent taxa at the species level. Those bacterial taxa possess relevant quantitative (microbial richness) and qualitative (microbial community composition) differences between individuals. Bacterial communities have significant differences between healthy and diseased oral cavities [7].

Plaque is composed of bacteria and a matrix that adheres to the outer tooth surface. Supra-gingival plaque is bacteria adherent above the gingiva, whereas bacteria below the gingiva is called sub-gingival plaque. Plaque constantly forms on our teeth when we eat foods or drink beverages with sugars or starches, the bacteria release acids that attack the tooth enamel [24]. The plaque is so sticky that it keeps the acids in contact with our teeth, as a result the acids will break down the enamel and lead to tooth decay. Plaque buildup can also lead to gum disease, the common one is gingivitis. If it progresses, severe periodontal (gum) disease can develop. Bacterial plaque that builds up on teeth and inflamed, allows the bacteria to destroy the underlying bone supporting the teeth.

Bacteria and inflammation in the mouth do more than just threaten the dental health, they are also linked to other problems, including heart attack and dementia, and may well jeopardize our overall health. Scientists have identified several links between poor oral health and other health problems [8]. As the distributions of bacterial can cause or prevent infections, thus it may have a significant impact on fully understanding human general health.

1.2.2 The Operational Taxonomic Units

The original definition of operational taxonomic unit (OTU) is a group of organisms used to classify groups of closely related individuals based on their character states. The term was introduced by Robert R. Sokal and Peter H. A. Sneath in the context of Numerical taxonomy [9]. Nowadays, the term “OTU” generally refers to clusters

of (uncultivable or unknown) microorganisms, grouped by DNA sequence similarity of a specific taxonomic marker gene.

An OTU table is a form of the sequencing results that will finally be really useful to analyze in excel, visualize, etc. It is a table giving the count of the number of sequences in each OTU, for each sample, and the taxonomy of that OTU. We focus on the OTU table of gingival plaques. In this thesis, our data includes 301 samples from sub-gingival plaque which contain 6782 OTUs, and 305 samples from supra-gingival plaque which contain 5277 OTUs.

Sites	Samples	OTUs
Sub-gingival plaque	301	6782
Supra-gingival plaque	305	5277

Table 1.1: Details of human oral cavity gingival plaque samples with associated abbreviations.

1.3 Challenges and Contributions

Microbiome is important for maintaining human health, and when things go wrong it will contribute to disease. Researchers show an increasing interest in human microbiome. But the OTU data from Human Microbiome Project present challenges to ecological and statistical interpretation. In particular, the sequencing depth often vary over several ranges of magnitude, and the data contains many 0's. Also, since the data always consist of hundreds or even thousands of variables but only a few observations, which means $p >> n$ (p is the number of variables and n is the number of observations), therefore we can't apply classical models to this kind of data because of high variance and overfitting. Here we explore several statistics methods to address the OTU data without losing too much important information.

In this thesis, our contributions are: first, model the OTU data with Negative Binomial (NB) model and fit the MLE's for NB model parameters. Then we perform the empirical Bayesian inference for the underlying composition of OTUs in a microbiome sample, try to visualize overall differences in bacterial composition between sample groups through the PCA plots. After that, we also perform the Likelihood Ratio (LR) test for differential distributions analysis, which tests the significant differences

in the distributions of OTUs between sample groups. After the LR test, we find those OTUs with strong predictive power and develop the Naive Bayes Discriminant Analysis (NBDA) based on the NB distributions of those OTUs. Finally we apply LASSO to several transformation of the estimated underlying compositions of OTUs and compare the NBDA results with the prediction accuracy of LASSO.

1.4 Thesis Outline

The remainder of this thesis is organized into 6 chapters. NB model checking and parameter estimation is given in Chapter 2, which includes the MLE of NB model fitting, checking the effects of MLE fitting and empirical distribution for OTU data. Chapter 3 introduces the empirical Bayesian inference for the underlying composition of OTUs. Log-likelihood ratio test are used to find those OTUs differently distributed between sub-gingival plaques and supra-gingival in Chapter 4. Mathematical formulation of the Naïve Bayes Discriminant Analysis Classifiers are given in Chapter 5. Chapter 6 is the application of LASSO, using LASSO to find the important OTUs to do classification; a comparison between the variable selection from LASSO with the significantly differently distributed OTUs by the LR test also given in Chapter 6. The conclusion of this thesis is in Chapter 7, which summarizes the results we achieved so far and gives ideas for future work.

Chapter 2

Modelling OTU Data with Negative Binomial Distribution

The 16S microbial data are in the format of the counts. Generally for each sample, thousands of different OTU counts can result from the preprocessing of the sequence data. However typically these counts are not directly comparable across different samples. A sample with deeper sequencing effort naturally results in more OTU species and more counts for the total number of OTUs. The common practice in this field is to normalize the data by either rarefaction or by taking the proportion of each OTU count out of the total count of the sample. Rarefaction is an ecological approach that standardize the data obtained from samples with different sequencing depth, and compare the OTU richness of the samples using this standardized platform. The approach of rarefaction is to randomly sample the same number of OTUs from each sample, and use this data to compare the communities at a given level of sampling effort. The rarefaction is not a recommended practice because it throws away a lot of data and make the analysis results less accurate [10]. The normalization using proportion is better than rarefaction, but it is quite typical that the underlying composition of OTUs can vary by orders of magnitude, which makes the comparisons between simple proportions not valid, and the resulted proportions are heterogeneous. Paul J. McMurdie and Susan Holmes [10] suggested that a better way to deal with such data is by fitting negative binomial model on these data. In this chapter, we will explore the negative binomial model fitting to the OTU count data.

2.1 MLE for Negative Binomial (NB) Model Parameters

2.1.1 Log-likelihood of NB Model

A counting distribution is a discrete distribution with non-zero probability mass only on the nonnegative integers. Though playing a prominent role in statistical theory, Poisson distribution is not appropriate in many situations, since it requires that the

mean and the variance are equal. Thus Negative Binomial distribution is an excellent alternative to the Poisson distribution, especially in the cases where the observed variance is greater than the observed mean. NB distribution has been widely used in modelling overdispersion in ecological count data [11, 12]. NB distribution can be viewed as a Poisson distribution where the Poisson mean itself is a random variable, distributed according to a Gamma distribution. In other words, NB distribution can be viewed as a generalization of the Poisson distribution, it is also termed as a Gamma-Poisson mixture. Fitting a Gamma-Poisson distribution on OTU counts means that we assume the OTU count follows a Poisson distribution, with its mean given by the sequencing depth multiplied by the underlying unobserved composition of the OTU. The underlying unobserved composition across different individuals for the same OTU follows a Gamma population distribution.

Our data include supra-gingival plaque samples and sub-gingival plaque samples from Human Microbiome Project (HMP) [13], which summarizes the count of each OTU in each sample. It includes 301 samples from sub-gingival plaque which contain 6782 OTUs, and 305 samples from supra-gingival plaque which contain 5277 OTUs. We will demonstrate the NB fitting on each OTU using the sub-gingival plaque samples in this chapter.

Taking a hierarchical model approach with the Gamma-Poisson distribution can provide a satisfactory fit to the underlying composition for many OTUs. Suppose d_i is a linear scaling factor for sample i that accounts for its sequence depth, that is the total read of i^{th} sample, $d_i = \sum_{j=1}^{p} x_{ij}$, hierarchically, x_{ij}, the number of j^{th} OTU in sample i, is Poisson distributed with parameters $d_i\lambda_{ij}$, denoted as $x_{ij} \sim Poisson(d_i\lambda_{ij})$. λ_{ij} is the underlying unobserved parameter of the composition of OTU_j, then λ_{ij} is independently identically Gamma distributed with shape parameter k_j and scale parameter θ_j, denoted as $\lambda_{ij} \overset{i.i.d}{\sim} \Gamma(k_j, \theta_j)$, then $d_i\lambda_{ij} \sim \Gamma(k_j, \theta_j d_i)$. A Gamma mixture of Poisson variable gives the Negative Binomial (NB) distribution, $x_{ij} \sim NB(k_j, \frac{\theta_j d_i}{\theta_j d_i + 1})$, the probability density function is

$$f(x_{ij}; d_i, k_j, \theta_j) = \frac{\Gamma(x_{ij} + k_j)}{x_{ij}!\Gamma(k_j)} \frac{(\theta_j d_i)^{x_{ij}}}{(\theta_j d_i + 1)^{x_{ij}+k_j}}. \tag{2.1}$$

The log-likelihood function for the j^{th} OTU with observed data $x_{ij}, i = 1, ..., n$ is

given by:

$$\begin{aligned}\ell(d_i, k_j, \theta_j; x_{ij}) = & \sum_{i=1}^{n} x_{ij} \log(\theta_j d_i) - \sum_{i=1}^{n} (x_{ij} + k_j) \log(\theta_j d_i + 1) \\ & + \sum_{i=1}^{n} \log(\Gamma(x_{ij} + k_j)) - \sum_{i=1}^{n} \log(\Gamma(k_j)),\end{aligned} \tag{2.2}$$

and this model can be fit by a Newton-Raphson algorithm. We use "nlminb" in R to find the MLEs for parameters k_j and θ_j.

We use the sub-gingival plaque data to estimate k_j and θ_j for the j^{th} OTU. In Figure 2.1, the graph in the left is the result of maximum likelihood estimates for k's, the values of those $\hat{k}$'s above red dashed line are greater than 112482, corresponding estimates for $\hat{\theta}$'s are smaller than $1e-09$, which are not in the reasonable interval for the NB parameters. These typically mean that the resulted estimates for both NB mean $(k_j\theta_j d_i)$ and variance$(k_j\theta_j d_i + k_j\theta_j^2 d_i^2)$ are almost 0. Due to the high sparsity of the data, the program doesn't really converge on these OTUs. From the estimate results, we can see that this fitting procedure doesn't work perfectly. There are 1266 cases in the program failed to converge, 4435 cases give excessively large estimates for k's and correspondingly small estimates for θ's. This means it's possible that NB model is not a good model for some OTUs and even if NB model is the correct model, there are many cases we can't get the reasonable MLEs for the parameters. When looking into those OTUs failing to give reasonable parameter estimations, from Figure 2.2, we can see that those OTUs are very sparse with mean percentage of zeros in each OTU as 97.81%, the mean of the total sum count of each OTU is 19.6774 for 301 observations. There is not enough information in these sparse OTUs with only several small non-zero observations (0,1 or 2) to estimate the distribution. We come to the conclusion that those are sparse and rare OTUs, as a result, we remove those 5701 cases which can not provide reasonable parameter estimates for k's and θ's.

2.1.2 Check the Effects of MLE Fitting based on Parametric Bootstrap of NB Model

The parameters of NB model are unknown and estimated from the original sub-gingival plaque count data. In this section, we check the fitting effects of our procedure supposing that the NB model is correct based on the remaining 1081 OTUs, which are

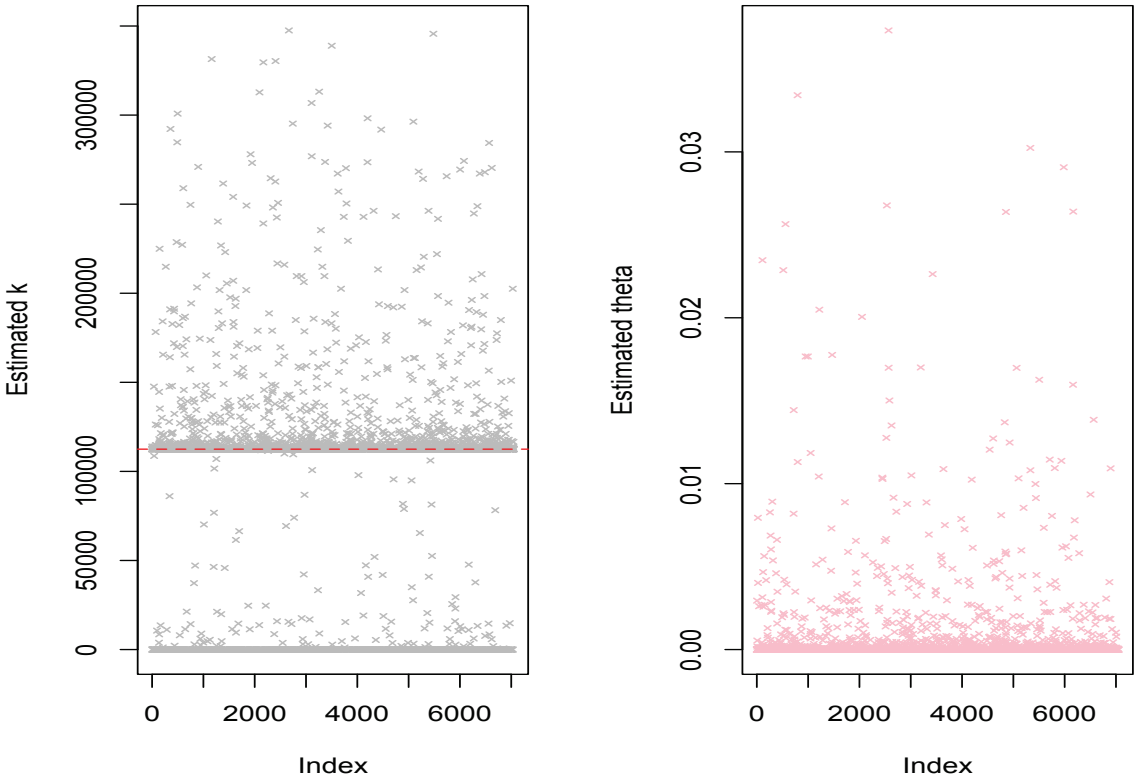

Figure 2.1: MLE estimates for k and θ.

able to provide reasonable parameters in last section. We parametrically bootstrap the NB data based on the MLEs of the parameters estimated from the original data and the sequencing depths also from the original data.

We generate new count data x'_{ij} from NB($\hat{k}_j, \frac{\hat{\theta}_j d_i}{\hat{\theta}_j d_i + 1}$). In this way, we simulate count number for each OTU of each sample. The number of each sample for each OTU is Negative Binomial distributed. We then estimate the Gamma-Poisson parameters $\hat{k}'_j$'s, $\hat{\theta}'_j$'s from the simulated data.

This simulation is illustrated by a flowchart as shown in Figure 2.3.

Figure 2.4 shows the comparison between parameters the simulation is based on and parameters estimated from simulated data. For most of the simulated OTUs, the estimated k's and θ's are quite close to its true values (those black crosses in the plot). There is a small portion of OTUs when the true k is close to 0, the estimated k' is too large and the corresponding estimated θ' is too small (the grey crosses in the

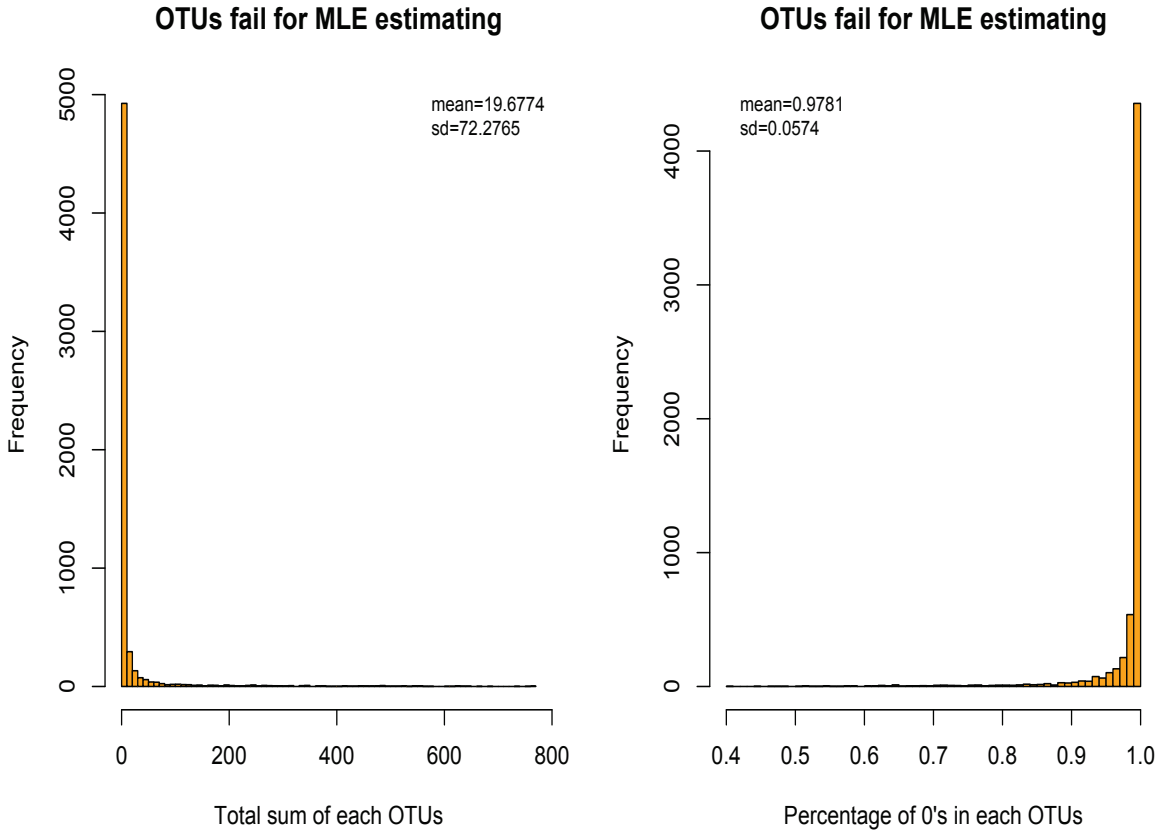

Figure 2.2: Summary of those OTUs which can not provide reasonable parameters for NB model.

plot). This can be seen from Figure 2.5, that the mean $\frac{k}{\frac{1}{\theta}}$'s estimated from simulated data are equal to the true mean the data are simulated from. It is again the sparsity in the simulated data for these OTUs that has resulted in the bias in the estimated parameters. In future, with an improved program for estimating NB parameters, some of these issues might be able to be resolved.

2.2 Evaluation of NB Model Fit to OTU Data

The parametric bootstrap of NB Model shows that some k's and θ's are not reliable, but their corresponding underlying composition means are reliable, these estimated NB models may or may not fit the OTU data very well. We need to assess the goodness-of-fit of NB model to the OTU data. The straightforward approach is employing a simple graphical method in which an overlay of the theoretical distribution

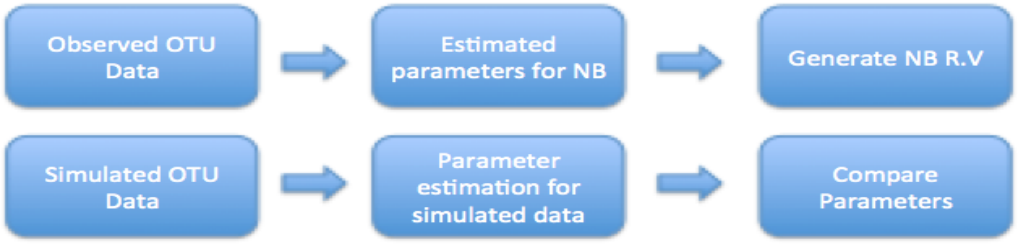

Figure 2.3: Simulation procedure for the parametric bootstrap of NB Model.

is displayed on a histogram of the data and a visual assessment is made to determine the quality of the fit.

2.2.1 Empirical Distribution for OTU Data

An intuitional way of evaluating how well the NB model fits the data is to compare the estimated gamma distribution with the empirical data. As we know, the underlying composition of j^{th} OTU for the i^{th} sample, $\lambda_{ij} \sim \Gamma(k_j, \theta_j)$. The observed count is Poisson distributed with mean $d_i\lambda_{ij}$, thus we can simply compare the observed proportion $\frac{x_{ij}}{d_i}$ with the estimated Gamma distribuiton $\Gamma(\hat{k}_j, \hat{\theta}_j)$.

From the top graphs of Figure 2.6, we can see that, for the abundant OTUs, the gamma$(\hat{k}, \hat{\theta})$ can fit the empirical data quite well; but for the rare OTUs, the fitted gamma distribution doesn't match the histogram of the OTU proportions at all. That means for some OTUs the estimated $\hat{k}$'s and $\hat{\theta}$'s are not reliable, even though their values are in the reasonable range. From the histograms we know that it is exactly again the sparsity in these data has prevented the accurate estimates of the parameters. It is not possible however to visualize all these more than one thousand fittings to decide which one is reliable, thus we will conduct a goodness-of-fit test to evaluate the Negative Binomial fitting to these OTUs. This can be done by using formal χ^2 statistical tests.

2.2.2 Check the Model Fit by Likelihood Ratio Tests

In this section, for those 1081 OTUs, we do a likelihood ratio test by comparing saturated Poisson model log-likelihood with the NB (Gamma-Poisson) model log-likelihood to check the goodness-of-fit of the NB (Gamma-Poisson) model.

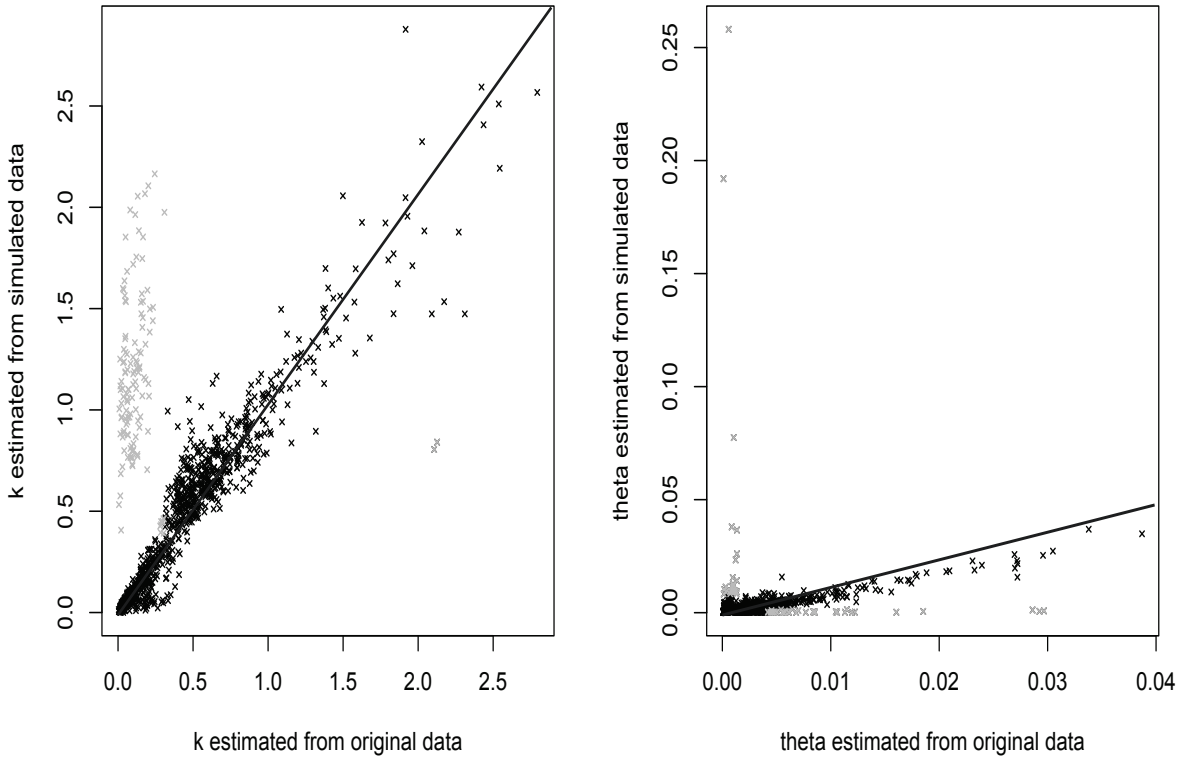

Figure 2.4: Parameters estimated from original data v.s. parameters estimated from simulated data.

For j^{th} OTU, the saturate model $x_{ij} \sim poisson(d_i\lambda_{ij})$,

$$f_{poisson}(x_{ij}, d_i; \lambda_{ij}) = \frac{(d_i\lambda_{ij})^{x_{ij}} e^{-d_i\lambda_{ij}}}{x_{ij}!},$$

where $d_i\lambda_{ij} = x_{ij}$. For NB (Gamma-Poisson) model,

$$f_{NB}(x_{ij}, d_i; k_j, \theta_j) = \frac{\Gamma(x_{ij}+k_j)}{x_{ij}!\Gamma(k_j)} \frac{(\theta_j d_i)^{x_{ij}}}{(\theta_j d_i + 1)^{x_{ij}+k_j}}.$$

Their log-likelihood functions are as follows:

$$\ell(\lambda_{ij}; x_{ij}, d_i) = x_{ij} log(x_{ij}) - x_{ij}, \tag{2.3}$$

$$\begin{aligned} \ell(k_j, \theta_j; x_{ij}, d_i) =& x_{ij} log(\theta_j d_i) - (x_{ij}+k_j) log(\theta_j d_i + 1) \\ &+ log(\Gamma(x_{ij}+k_j)) - log(\Gamma(k_j)). \end{aligned} \tag{2.4}$$

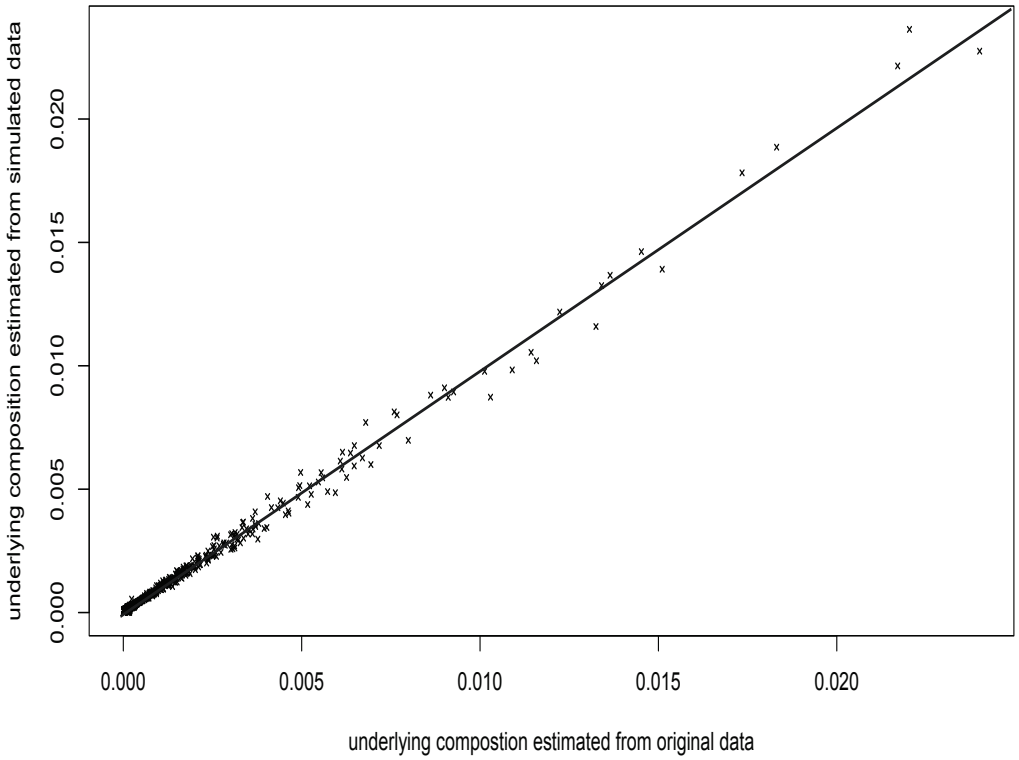

Figure 2.5: Underlying composition mean estimated from original data v.s. underlying composition mean estimated from simulated data.

The null hypothesis is H_0: The NB model is true. The likelihood ratio statistic follows a χ^2 distribution with $df = (n-2)$, where $n = 301$ is sample size for each OTU in the sub-gingival plaque data. The likelihood ratio statistic is:

$$\begin{aligned} T_{\chi^2} =& 2 \times log \frac{\prod_{i=1}^{n} f_{poisson}(x_{ij}; \lambda_{ij})}{\prod_{i=1}^{n} f_{NB}(x_{ij}, d_i; k_j, \theta_j)} \\ =& 2 \times (\sum_{i=1}^{n} \ell(\lambda_{ij}; x_{ij}, d_i) - \sum_{i=1}^{n} \ell(k_j, \theta_j; x_{ij}, d_i)) \\ =& 2 \times (\sum_{i=1}^{n} x_{ij} log(x_{ij}) - \sum_{i=1}^{n} x_{ij} - \sum_{i=1}^{n} x_{ij} log(\theta_j d_i) + \sum_{i=1}^{n} (x_{ij} + k_j) log(\theta_j d_i + 1) \\ & - \sum_{i=1}^{n} log(\Gamma(x_{ij} + k_j)) + \sum_{i=1}^{n} log(\Gamma(k_j))) \end{aligned} \tag{2.5}$$

Summaries of both χ^2 statistics and p-values are provided in Table 2.1. Larger

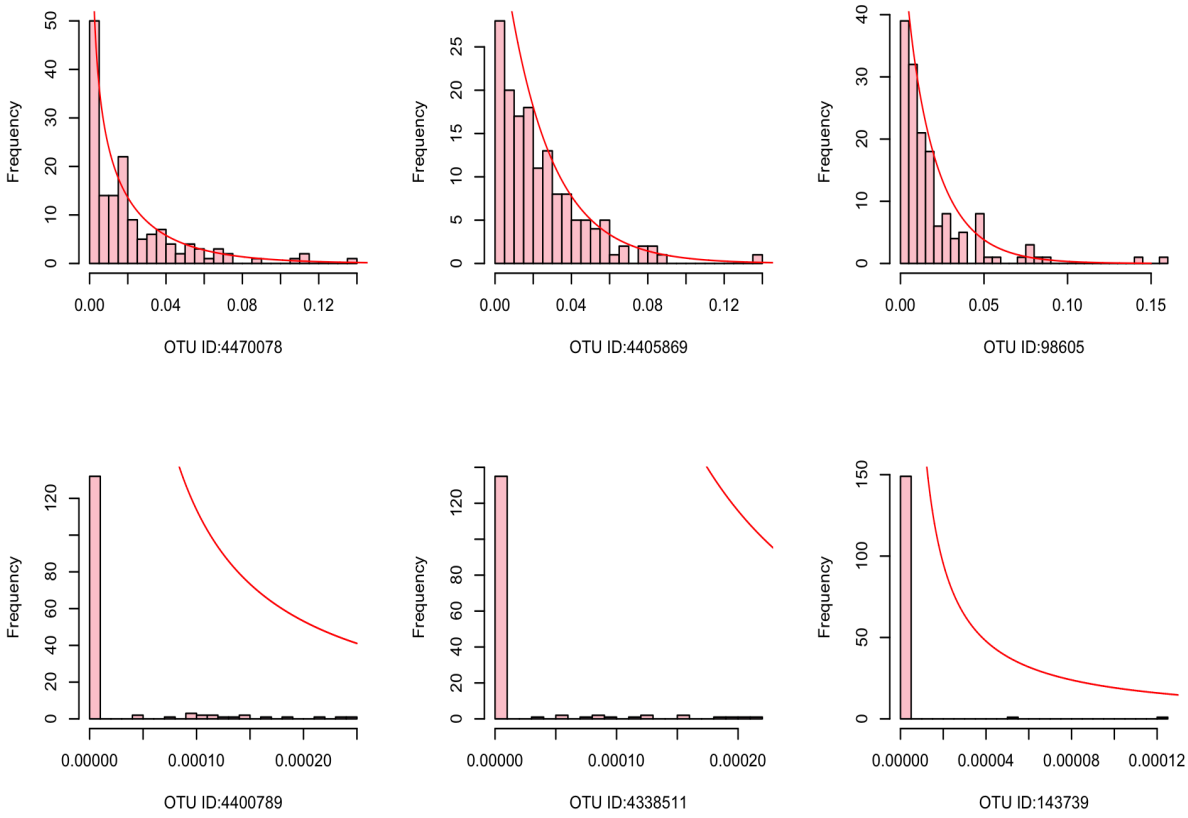

Figure 2.6: Histograms of OTU proportions

values of χ^2 statistics lead to small p-values, which provide evidence against the NB model. Under significant level $\alpha = 0.05$, 776 OTUs have no evidence against NB model, that means we can fit NB model to these 776 OTUs' count data. In the rest of the thesis we will only focus on these 776 OTUs for sub-gingival plaque data.

Similarly, we perform NB model parameter estimates and χ^2 statistics test based on supra-gingival plaque samples. Parameter estimates from 798 OTUs are in reasonable range and through χ^2 statistics test, 492 of them fit well with an NB model. Among those 776 OTUs resulted from sub-gingival plaques and 492 OTUs resulted from supra-gingival plaques, there are 361 common OTUs. Our inference of the comparisons on the underlying composition of OTUs will be focused on these 361 OTUs which are successful in NB model fitting in both sub-gingival plaques and supra-gingival plaques.

Sub		$df = 299$
χ^2	$\chi^2 > 340$	$\chi^2 \leq 340$
P value	$p \leq 0.05$	$p > 0.05$
OTU No.	305	776

Supra		$df = 303$
χ^2	$\chi^2 > 344$	$\chi^2 \leq 344$
P value	$p \leq 0.05$	$p > 0.05$
OTU No.	306	492

Table 2.1: Log-likelihood ratio test to check the NB model fitting.

Chapter 3

Empirical Bayesian Inference for the Underlying Composition of OTUs in a Microbiome Sample

We are interested in understanding the underlying composition of different OTUs in a microbiome sample and their relationships to the corresponding population. The observed data are Poisson distributed with the underlying composition multiplied by a sequencing depth parameter as the mean. In order to better estimate the underlying composition of OTUs, we explore to use the Bayesian posterior mean instead of the commonly used simple proportion normalization.

A common method for normalization of OTU count data is using the simple proportion [14], which normalizes count data by dividing OTU read counts by the total number of reads in each sample. The simple proportion is not a biased estimate, however they are associated with different variance due to the different sequencing depths for different samples. We develop a new method to normalize count data through the posterior distribution means and examine what difference this will make for the OTU composition estimates.

3.1 Posterior — Mean as a Compromise between Data and Prior Information

The process of Bayesian inference involves the prior distribution, $p(\Theta)$ and a posterior distribution, $p(\Theta|x)$[15]. For example, in the Poisson example with the Gamma prior distribution, Bayesian statistics involve the following steps:

- Define the prior distribution that incorporates the subjective beliefs about a parameter, in this example it is Gamma distribution, with prior mean $k\theta$ and prior variance $k\theta^2$.

- Collect OTU count data x and sequencing depth d.

- Update the prior distribution with the data using Bayes' theorem to obtain a posterior distribution. The posterior distribution is a probability distribution that represents the updated beliefs about the parameter after having seen the data.

- Analyze the posterior distribution and summarize it (mean, i.e., posterior mean $\frac{x+k}{d+\frac{1}{\theta}}$, median, standard deviation, quantiles,...).

The posterior mean, $\frac{x+k}{d+\frac{1}{\theta}}$, is a compromise between the prior mean and the sample proportion, $\frac{x}{d}$. As the data sample increases, the prior mean plays an increasingly smaller role. This is a general feature of Bayesian inference: the posterior distribution is centered at a point that represents a compromise between the prior information and the data, and the compromise is controlled to a greater extent by the data as the sample size increases.

3.2 Empirical Bayesian Posterior Mean for OTU Data

We calculate the posterior mean of each OTU in sub-gingival plaque base on $\hat{k}$'s and $\hat{\theta}$'s we estimated from our data in Chapter 2. In fact, this is not really a Bayesian inference, because the priors are from the data as well. Such procedure is commonly called empirical Bayes estimation. By using such a prior, we are estimating the OTU composition of each sample by borrowing information from the estimated population composition, i.e. from all other samples. In our case, for each column, the Poisson model for data $x_{ij} \sim Poisson(d_i\lambda_{ij})$,where x_{ij} is the observed count number of OTU_j in sample i, $i = 1, ..., 301, j = 1, ..., 361$, where d_i is a linear scaling factor for sample i that accounts for its sequence depth. λ_{ij} is the underlying composition of j^{th} OTU in i^{th} observation that is being inferred.

In Bayes statistics, the parameterization with α and β is more common for Gamma distribution, in this notation, with prior distribution $\lambda_{ij} \sim \Gamma(\alpha_j, \beta_j)$, now $\alpha_j = k_j, \beta_j = \frac{1}{\theta_j}$, where the resulting posterior distribution for i^{th} sample is $\lambda_{ij}|x_{ij} \sim$

$\Gamma(\alpha_j + x_{ij}, \beta_j + d_i)$. Now we can get the posterior mean estimate for λ_{ij},

$$\begin{aligned}\hat{\lambda}_{ij} = E(\lambda_{ij}|x_{ij}) = \frac{\alpha_j + x_{ij}}{\beta_j + d_i} = \frac{k_j + x_{ij}}{\frac{1}{\theta_j} + d_i} \\ = \frac{\frac{1}{\theta_j}}{\frac{1}{\theta_j} + d_i} \cdot \frac{k_j}{\frac{1}{\theta_j}} + \frac{d_i}{\frac{1}{\theta_j} + d_i} \cdot \frac{x_{ij}}{d_i}.\end{aligned} \tag{3.1}$$

These estimates are readily provided by the MLEs of the NB fitting in the last chapter. This provides another way of normalization for the microbiome count data. The posterior mean for the i^{th} sample, $\frac{k_j + x_{ij}}{\frac{1}{\theta_j} + d_i}$, is a compromise between the prior mean $\frac{k_j}{\frac{1}{\theta_j}}$, which is estimated from all samples and weighted by $\frac{k_j}{\frac{1}{\theta_j} + d_i}$, and the simple proportion $\frac{x_{ij}}{d_i}$, which is weighted by $\frac{d_i}{\frac{1}{\theta_j} + d_i}$. Since d_i is the sequencing depth, which is often much larger than $\frac{1}{\theta_j}$, thus, the prior mean $\frac{k_j}{\frac{1}{\theta_j}}$ has very little effects for most samples except the ones with very low sequencing depths. The posterior means are mostly approximately equal to the simple proportions, but the simple proportions which are 0's will be estimated by some small non-zero numbers. This change makes it much easier to perform the analysis on the log transformed composition.

Figure 3.1 left panel is the PCA plot based on the covariance matrix of simple proportion of sub-gingival plaques and Figure 3.1 right panel is the PCA plot based on the covariance matrix of Bayesian posterior mean we estimated for sub-gingival plaques. We can see that the right panel is quite the same as left panel with 180° rotation. The simple proportions contain so many 0's that we are unable to perform a log-transformation on the simple proportions. The Bayesian posterior means change the simple proportions which are 0's into slightly non-zero's. This enable us to perform a log-transformation on the posterior means. The bottom plot in Figure 3.1 is the PCA plot based on the covariance matrix of log-transformation of the posterior mean. We can see that after log-transformation of the posterior mean, the PCA plot is quite different from the original one. The log-transformation of the posterior mean estimates for the OTU compositions have greatly smoothed the data and reduced the extreme influence from the spurious simple proportion estimates. The data after the log-transformation now look much more "normal distributed".

Similarly, based on the 361 common OTUs which are successful in NB model fitting in both sub-gingival plaques and supra-gingival plaques, we separately estimate their

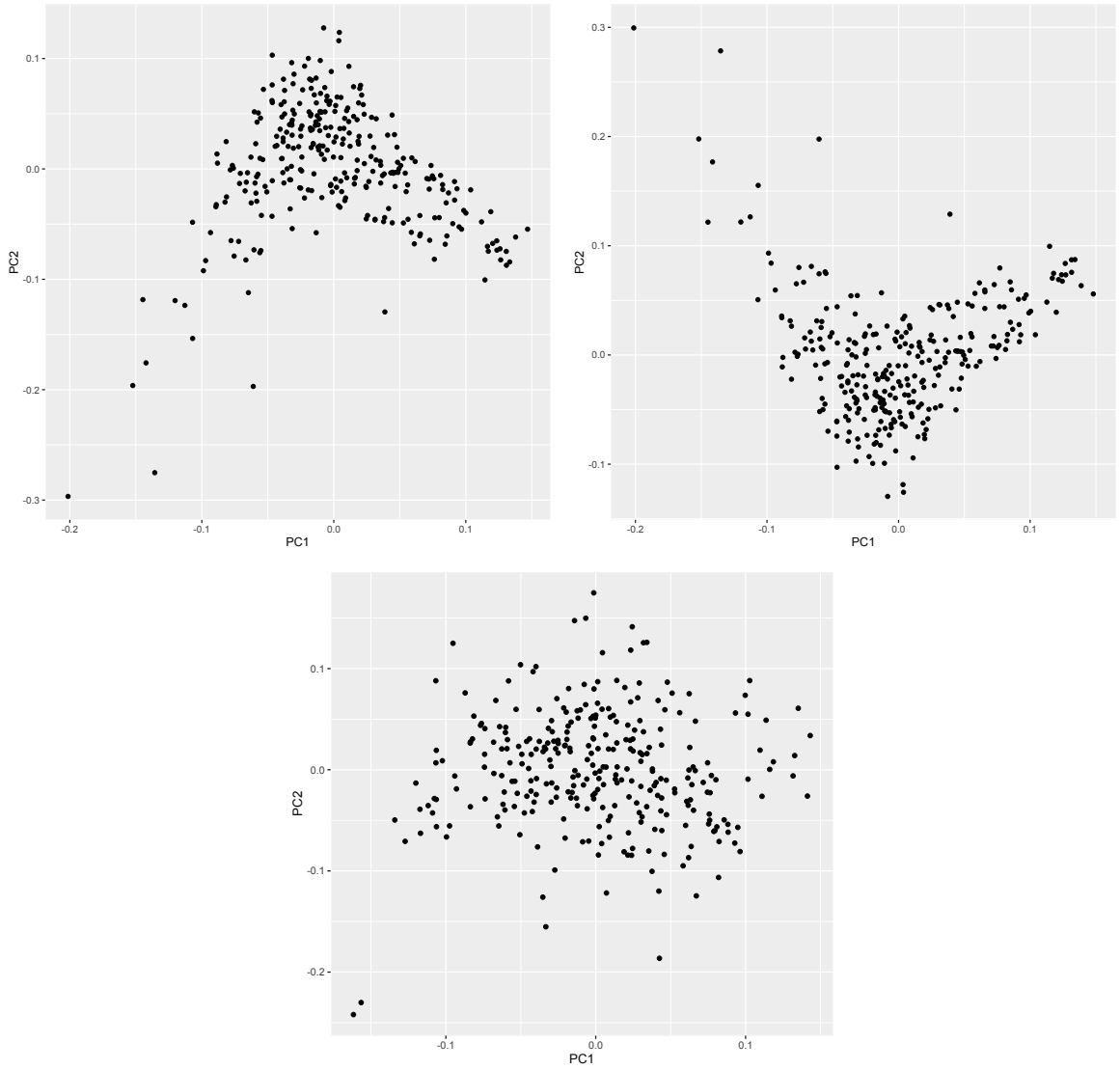

Figure 3.1: PCA plots based on two analysis methods.

posterior means $\hat{\lambda}^1_{ij}$ $(i = 1, ..., 301)$ and $\hat{\lambda}^2_{ij}$ $(i = 1, ..., 305)$, perform PCA on the log-transformation of $\hat{\lambda}^1_{ij}$ and $\hat{\lambda}^2_{ij}$. In Figure 3.2, the top graph is the PCA plot based on the covariance matrix of simple proportions of sub-gingival plaques and supra-gingival plaques. We also combine posterior means estimated from sub-gingival plaque with posterior means estimated from supra-gingival plaque together, then performed the log-transformation on the posterior means and calculate the covariance matrix over pooled posterior means. The bottom graph is the PCA plot based on the covariance matrix of the log-transformed data. The separation is quite clear in the bottom plot, although the change from simple proportion to the posterior mean is very small for

each data point.

Comparing the coefficients of the first eigenvector and the second eigenvector from two principal component analysis methods, from Figure 3.3, we can see that the coefficients of the log-transformation of the posterior means are quite different from the coefficients of the simple proportions, which means the PCA eigenvector directions from these two procedures are very different. The directions from log-transformation of posterior means are more decided by majority of the OTUs while as the PCA direction from the simple proportions are mainly related to several OTUs with larger observations.

3.3 Posterior Mean Estimate for OTU Composition based on a Mixture Gamma Prior

The data of human oral cavity from HMP is about two populations (301 sub-gingival plaque samples and 305 supra-gingival plaque samples) for each OTU. To fully use all the information, we explore the posterior mean estimates based on a mixture Gamma prior.

For i^{th} sample j^{th} OTU, the mixture Gamma prior can be written as $\pi\Gamma(k_{1j}, \theta_{1j}) + (1-\pi)\Gamma(k_{2j}, \theta_{2j})$, where π is the prior probability that an observation from the sub-gingival plaque population. The posterior probability under such a mixture Gamma prior is

$$p(\lambda_{ij}|x_{ij}) = \frac{(\pi\Gamma(k_{1j}, \theta_{1j}) + (1-\pi)\Gamma(k_{2j}, \theta_{2j}))L(x_{ij}|\lambda_{ij})}{\int(\pi\Gamma(k_{1j}, \theta_{1j}) + (1-\pi)\Gamma(k_{2j}, \theta_{2j}))L(x_{ij}|\lambda_{ij})d\lambda_{ij}} \tag{3.2}$$

The posterior mean based on the above posterior distribution is

$$
\begin{aligned}
&E(\lambda_{ij}|x_{ij}) = \int \lambda_{ij} p(\lambda_{ij}|x_{ij}) d\lambda_{ij} \\
=&\pi \int \lambda_{ij} \frac{\Gamma(k_{1j},\theta_{1j})L(x_{ij}|\lambda_{ij})}{\int(\pi\Gamma(k_{1j},\theta_{1j})+(1-\pi)\Gamma(k_{2j},\theta_{2j}))L(x_{ij}|\lambda_{ij})d\lambda_{ij}} d\lambda_{ij} \\
&+(1-\pi)\int \lambda_{ij} \frac{\Gamma(k_{2j},\theta_{2j})L(x_{ij}|\lambda_{ij})}{\int(\pi\Gamma(k_{1j},\theta_{1j})+(1-\pi)\Gamma(k_{2j},\theta_{2j}))L(x_{ij}|\lambda_{ij})d\lambda_{ij}} d\lambda_{ij} \\
=&\frac{\pi\int\Gamma(k_{1j},\theta_{1j})L(x_{ij}|\lambda_{ij})d\lambda_{ij}}{\int(\pi\Gamma(k_{1j},\theta_{1j})+(1-\pi)\Gamma(k_{1j},\theta_{1j}))L(x_{ij}|\lambda_{ij})d\lambda_{ij}} \int \lambda_{ij} \frac{\Gamma(k_{1j},\theta_{1j})L(x_{ij}|\lambda_{ij})}{\int\Gamma(k_{1j},\theta_{1j})L(x_{ij}|\lambda_{ij})d\lambda_{ij}} d\lambda_{ij} \\
+&\frac{(1-\pi)\int\Gamma(k_{2j},\theta_{2j})L(x_{ij}|\lambda_{ij})d\lambda_{ij}}{\int(\pi\Gamma(k_{1j},\theta_{1j})+(1-\pi)\Gamma(k_{1j},\theta_{1j}))L(x_{ij}|\lambda_{ij})d\lambda_{ij}} \int \lambda_{ij} \frac{\Gamma(k_{2j},\theta_{2j})L(x_{ij}|\lambda_{ij})}{\int\Gamma(k_{2j},\theta_{2j})L(x_{ij}|\lambda_{ij})d\lambda_{ij}} d\lambda_{ij} \\
=&\frac{\pi f_1(x_{ij})}{\pi f_1(x_{ij})+(1-\pi)f_2(x_{ij})}\hat{\lambda}^1_{ij} + \frac{(1-\pi)f_2(x_{ij})}{\pi f_1(x_{ij})+(1-\pi)f_2(x_{ij})}\hat{\lambda}^2_{ij}
\end{aligned}
\tag{3.3}
$$

The mixture posterior mean is:

$$
\hat{\lambda}^{mix}_{ij} = \frac{\pi f_1(x_{ij})}{\pi f_1(x_{ij})+(1-\pi)f_2(x_{ij})}\hat{\lambda}^1_{ij} + \frac{(1-\pi)f_2(x_{ij})}{\pi f_1(x_{ij})+(1-\pi)f_2(x_{ij})}\hat{\lambda}^2_{ij}, \tag{3.4}
$$

where $\pi = \frac{n_1}{n}$ is the prior probability that samples come from sub-gingival plaque, $f_1(x_{ij})$ is the NB probability estimated from sub-gingival plaque; $f_2(x_{ij})$ is the NB probability estimated from supra-gingival plaque. Thus the whole term (in front of $\hat{\lambda}^1_{ij}$) is the posterior probability that x_{ij} is from population 1. $\hat{\lambda}^1_{ij}$ is the posterior mean of λ_{ij} calculated using the prior of the 1^{st} population, similarly for the second term in (3.4). Thus the mixture prior posterior mean estimate of λ_{ij} is the weighted average of the posterior means from two different priors with the weights given by the posterior probability of the observation from two different populations.

In summary, we have the following procedure:

- For j^{th} OTU, $j = 1, ..., 361$, estimate NB parameters $\hat{k}_{1j}, \hat{\theta}_{1j}$ from sub-gingival plaque samples; at the same time, estimate NB parameters $\hat{k}_{2j}, \hat{\theta}_{2j}$ from supra-gingival plaque samples;
- For i^{th} sample j^{th} OTU, $i = 1, ..., 606$, calculate the posterior mean $\frac{\hat{k}_{1j}+x_{ij}}{\frac{1}{\hat{\theta}_{1j}}+d_i}$ base on parameters from sub-gingival plaque group, denote it as $\hat{\lambda}^1_{ij}$;
- For i^{th} sample j^{th} OTU, $i = 1, ..., 606$, calculate the posterior mean $\frac{\hat{k}_{2j}+x_{ij}}{\frac{1}{\hat{\theta}_{2j}}+d_i}$ base on parameters from supra-gingival plaque group, denote it as $\hat{\lambda}^2_{ij}$;

- For i^{th} sample j^{th} OTU, $i = 1, ..., 606$, calculate the mixture posterior mean $\hat{\lambda}_{ij}^{mix} = \frac{\pi f_1(x_{ij})}{\pi f_1(x_{ij})+(1-\pi)f_2(x_{ij})}\hat{\lambda}_{ij}^{1} + \frac{(1-\pi)f_2(x_{ij})}{\pi f_1(x_{ij})+(1-\pi)f_2(x_{ij})}\hat{\lambda}_{ij}^{2}$.

We apply the PCA on the derived mixture prior posterior mean estimates for the OTU composition on both super gingival and sub gingival plaque data, in comparison with the PCA analysis based on the log-transformation of the mixture prior posterior means. Figure 3.4 shows the projected data on the 1^{st} and 2^{nd} principal components.

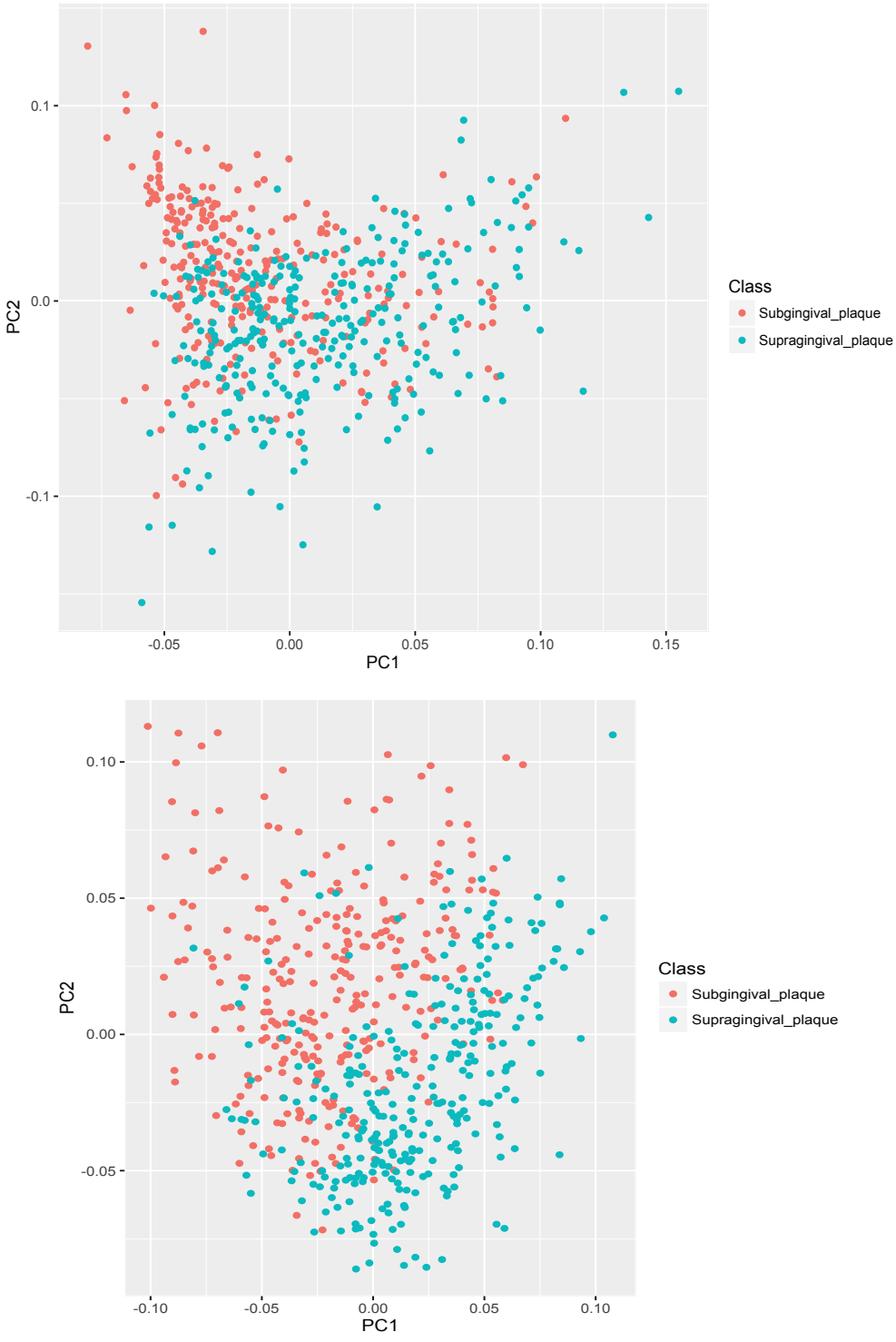

Figure 3.2: Top: PCA plot based on simple proportion of sub-gingival plaques and supra-gingival plaques; the bottom: PCA plot of log-transformation of posterior means separately estimated from sub-gingival plaque and supra-gingival plaque.

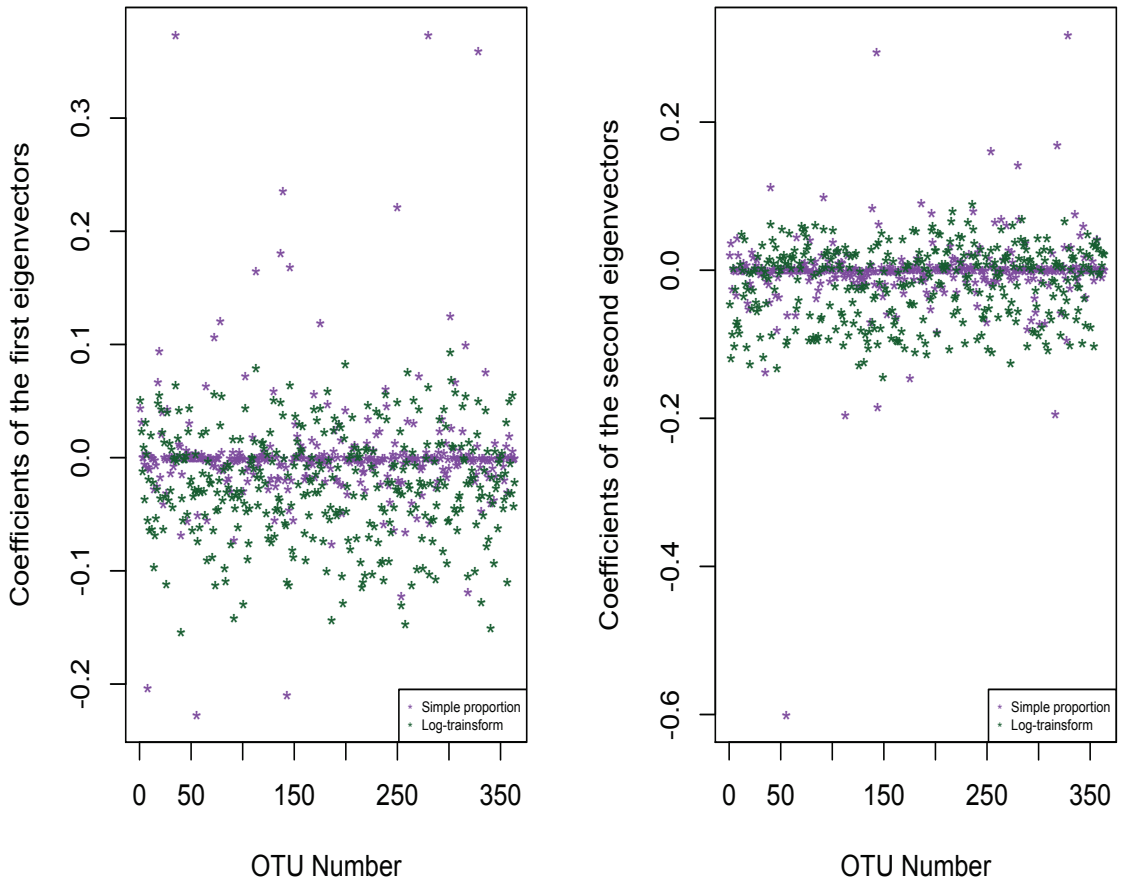

Figure 3.3: Coefficients of the first eigenvector and the second eigenvector from two principal component analysis.

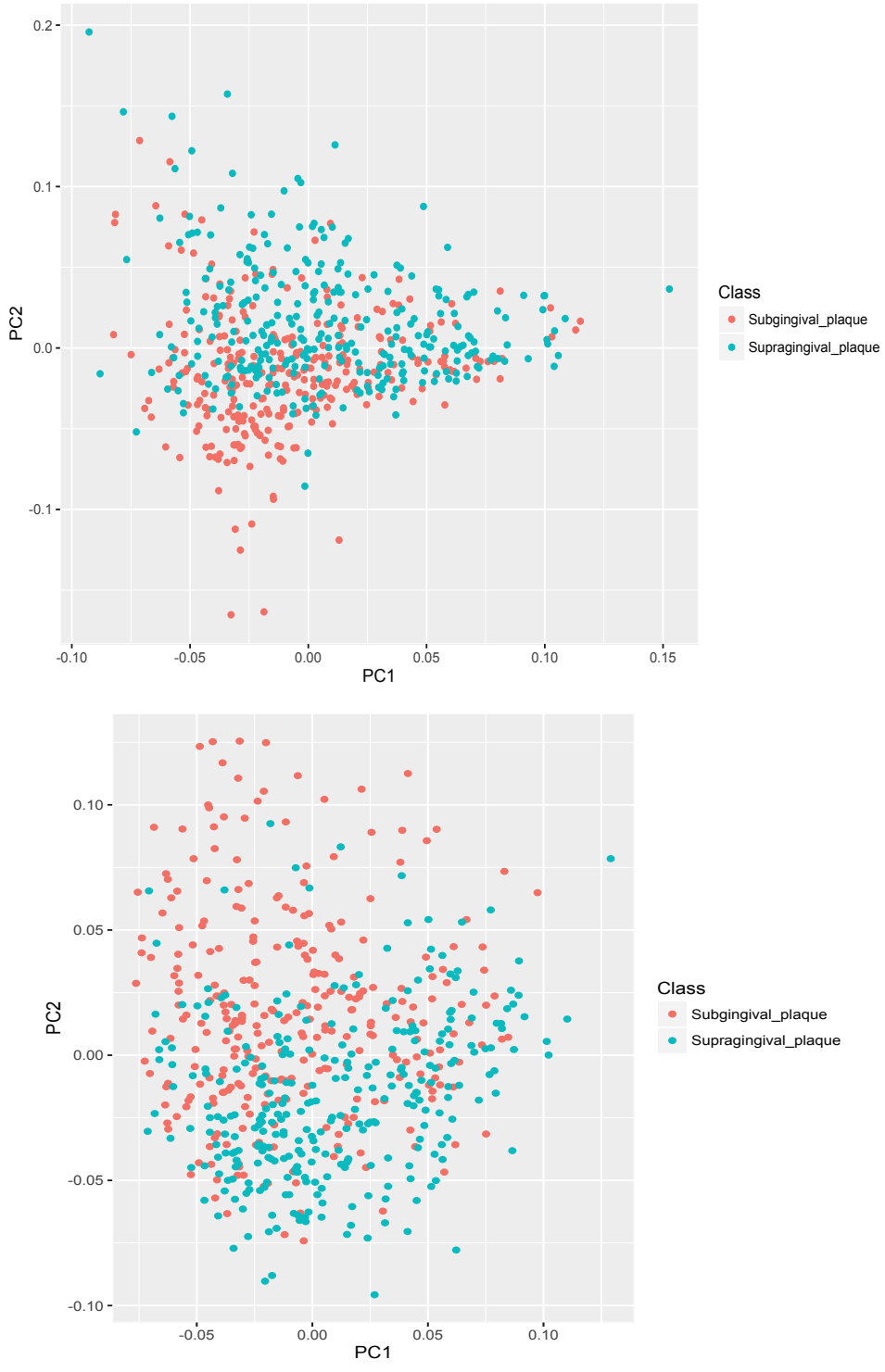

Figure 3.4: Top: PCA plot based on mixture posterior mean from sub-gingival plaques and supra-gingival plaques; bottom: PCA plot based on log-transformation of mixture posterior means.

Chapter 4

Differentially Distributed OTUs in Two Environments

It is quite natural that such questions would arise: are the NB models estimated from sub-gingival plaques the same with NB model estimated from supra-gingival plaques for the same OTUs? Which OTUs will have different underlying compositions between these two populations? Do they play a critical role in classification? We try to find out those OTUs which have different underlying compositions between sub-gingival plaques and supra-gingival plaques. In this chapter, the log-likelihood ratio test based on NB distribution will be employed to explore these issues.

4.1 Likelihood Ratio Test

The likelihood ratio test is conducted as following: in the null model, we use the whole dataset to estimate the NB model, let $\hat{k}_0$, $\hat{\theta}_0$ be the values of the parameters that maximize the likelihood function. Let the maximum likelihood function be written as $L_0(\hat{k}_0, \hat{\theta}_0)$. In the alternative model, we separate the OTUs into two groups – sub-gingival plaques and supra-gingival plaques. An NB model is fitted on the sub-gingival plaques samples and the supra gingival plaques samples separately. Let $\hat{k}_1$, $\hat{\theta}_1$ be the values of the parameters that maximize the sub-gingival plaques likelihood function; let $\hat{k}_2$, $\hat{\theta}_2$ be the values of parameters that maximize the supra-gingival plaques likelihood function. Then the maximum likelihood function of the alternative model can be written as $L_a(\hat{k}_1, \hat{\theta}_1, \hat{k}_2, \hat{\theta}_2)$.

H_0: The OTU in sub-gingival plaque and supra-gingival plaque follows the same NB distribution;

H_a: The OTU in sub-gingival plaque and supra-gingival plaque follows different NB distributions.

The test statistic is $T = 2log(L_a/L_0)$. The null hypothesis is rejected if $T > c$, where c is the critical value at significant level $\alpha = 0.05$. T has a chi-square distribution with degree of freedom equal to 2. Based on the likelihood ratio test results,

we can decide the significantly different OTUs between the sub-gingival plaques and supra-gingival plaques.

4.2 False Discovery Rate Control (FDR) and BH Method

When we conduct a single hypothesis test, we choose a rejection threshold to control Type I error rate. In the Log-likelihood Ratio Test mentioned above, we perform 361 simultaneous hypothesis tests. With multiple tests, choosing a rejection threshold becomes more complicated. Each of the tests has possible Type I and Type II errors, and there are many ways to combine them. The probability of Type I error increases with the number of tests. FDR control offers a way to choose a threshold, by increasing power while maintaining some principled bound on error.

Benjamini and Hochberg [16] introduced the FDR (False Discovery Rate) and show a procedure independently from Simes [17, 18]. The procedure - which is called the BH procedure - is simple to calculate.

Consider testing m hypotheses, $H_1, H_2, \ldots, H_m$ based on their respective p values, $p_1, p_2, \ldots, p_m$. Consider that a fraction q^* of discoveries are allowed (tolerated) to be false. Sort the p values in ascending order, $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ and denote $H_{(i)}$ the hypothesis corresponding to $p_{(i)}$. Let k be the largest i for which $p_{(i)} \leq \frac{i}{m}q^*$. Then reject all $H_{(i)}, i = 1, 2, \ldots, k$. The BH procedure has found many applications across different fields, including neuroimaging, as introduced by Genovese et al. [19].

4.3 OTUs Differently Distributed between Sub-gingival Plaque and Supra-gingival Plaque

We apply the Log-likelihood Ratio Test to find which OTUs are significantly differently distributed between sub-gingival plaques and supra-gingival plaques. The selection process is defined by the following steps:

- For sub-gingival plaque microbime samples, use parameters estimated from sub-gingival plaque group ($\hat{k}_1$,$\hat{\theta}_1$) to calculate NB probability for each OTU, denoted as $f_1(x_{ij})$;

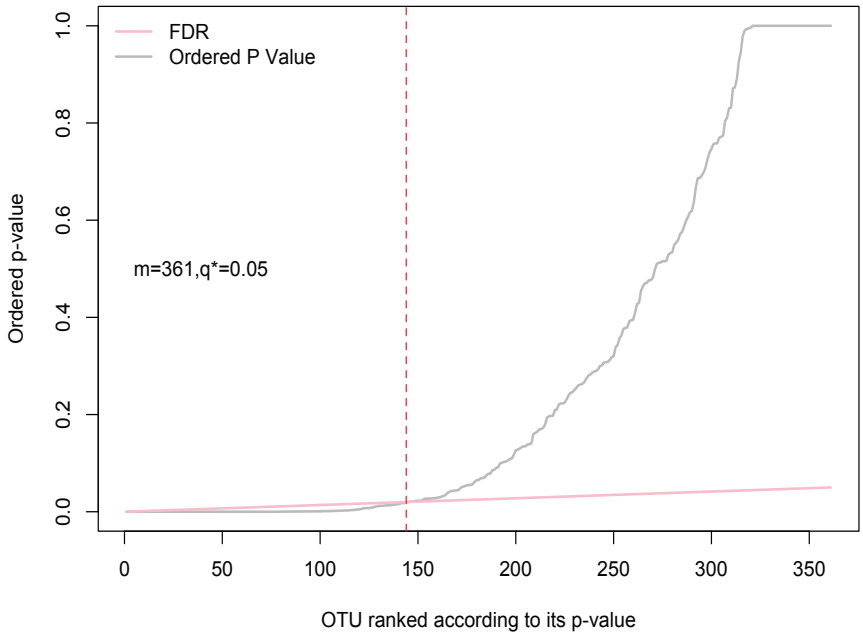

Figure 4.1: Use FDR (False Discovery Rate) to determine the number of differentially distributed OTUs between sub-gingival plaque and supra-gingival plaque.

- For supra-gingival plaque microbime samples, use parameters estimated from supra-gingival plaque group ($\hat{k}_2$,$\hat{\theta}_2$) to calculate NB probability, denoted as $f_2(x_{ij})$;

- For each OTU, calculate NB probability based on the parameters ($\hat{k}_0$,$\hat{\theta}_0$) estimated from all the samples, denoted as $f_0(x_{ij})$;

- Calculate $T = 2 \times log\frac{\prod_{j=1}^{n_1} f_1(x_{ij}) \prod_{j=(n_1+1)}^{n} f_2(x_{ij})}{\prod_{j=1}^{n} f_0(x_{ij})}$, n_1 is the number of sub-gingival plaque samples and n is the total number of samples;

- Use FDR (False Discovery Rate) to determine the number of OTUs, which have significantly different distributions between sub-gingival plaque and supra-gingival plaque.

According to the results shown in Figure 4.1, the first 144 p values are smaller than the FDR, that means the first 144 OTUs are significantly different between sub-gingival plaques and supra-gingival plaques. The order of the p values also provides

an order of the significance of the test, thus the OTUs with the smallest p values are the most differentially distributed between two populations. And interestingly, most of those OTUs which have quite different coefficients between PCA of simple proportions and PCA of log-transformation of posterior means are in these 144 significantly differently distributed OTUs. Are these OTUs really good at classification? We'll verify them with the OTUs ordered according to their predictive power in the next chapter.

Chapter 5

Naïve Bayes Discriminant Analysis based on NB Likelihood

The Linear Discriminant Analysis (LDA) for classification was first developed by R.A. Fisher in 1936 [20]. Fisher's LDA searches for a linear combination of variables to best separate two classes.

5.1 Review of Linear Discriminant Analysis

LDA can be derived from simple probabilistic models which model the class conditional distribution of the data $P(X|y = k)$ for each class k, where X is a vector of the measurement and y is the class membership. Predictions can then be obtained by using Bayes' rule:

$$P(y = k|X) = \frac{P(X|y = k)P(y = k)}{P(X)} = \frac{P(X|y = k)P(y = k)}{\sum_l P(X|y = l) \cdot P(y = l)}, \tag{5.1}$$

and we select the class k which maximizes this conditional probability. More specifically, for linear and quadratic discriminant analysis, $P(X|y)$ is modelled as a multivariate Gaussian distribution with density:

$$p(X|y = k) = \frac{1}{(2\pi)^n |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(X - \mu_k)^t \Sigma_k^{-1}(X - \mu_k)\right). \tag{5.2}$$

To use this model as a classifier, we need to estimate from the training data the class priors $P(y = k) = \frac{n_k}{N}$, that is the proportion of instances of class k; the class means μ_k (by the empirical sample class means) and the covariance matrices (either by the empirical sample class covariance matrices, or by a regularized estimator). In LDA, the Gaussians for each class are assumed to share the same covariance matrix: $\Sigma_k = \Sigma$ for all k. This leads to linear decision boundary, as can be seen by the log-probability ratios $\log[P(y = k|X)/P(y = l|X)]$:

$$\log\left(\frac{P(y = k|X)}{P(y = l|X)}\right) = 0 \Leftrightarrow (\mu_k - \mu_l)\Sigma^{-1}X = \frac{1}{2}(\mu_k^t \Sigma^{-1}\mu_k - \mu_l^t \Sigma^{-1}\mu_l) \tag{5.3}$$

LDA assigns y to Class k if $\log\left(\frac{P(y=k|X)}{P(y=l|X)}\right) > 0$ for all $l(l \neq k)$. We generalize LDA from Normal distribution to NB distribution here, developing a Naïve Bayes Discriminant Analysis (NBDA) based on NB model. We try to find whether our NBDA method produces a classifier whose accuracy is as good as other more complex methods. In addition, by variable selection, there are a number of ways to select key discriminating OTUs, the NBDA may select the most important OTUs that discriminate the populations which can provide the information that the black box type of classification methods incapable of.

5.2 Mathematical Formulation of the Naïve Bayes Discriminant Analysis Classifiers

Naïve Bayes here means that we assume OTUs to be independent. Some notations come as follows:

- The prior probability of class m is π_m, $\sum_{m=1}^{M} \pi_m = 1$,π_m is estimated simply by empirical frequencies of the training set:

$$\hat{\pi}_m = \frac{number \quad of \quad samples \quad in \quad class \quad m}{total \quad number \quad of \quad samples} \tag{5.4}$$

- The class-conditional density of $x_i = (x_{i1}, \cdots, x_{ip})$ in class $G = m$ is $f_m(x_i)$,

$$f_m(x_i) \sim \prod_{j=1}^{p} NB(k_{mj}, \frac{\theta_{mj} d_i}{1 + \theta_{mj} d_i}). \tag{5.5}$$

- Compute the posterior probability

$$Pr(G = m|X = x_i) = \frac{f_m(x_i)\pi_m}{\sum_{\ell=1}^{M} f_\ell(x_i)\pi_\ell}. \tag{5.6}$$

For Class m, Sample i,

$$\begin{aligned} f_m(x_i) &= \prod_{j=1}^{p} f(x_{ij}; d_i, k_{mj}, \theta_{mj}) \\ &= \prod_{j=1}^{p} \frac{\Gamma(x_{ij} + k_{mj})}{x_{ij}!\Gamma(k_{mj})} (\frac{\theta_{mj} d_i}{\theta_{mj} d_i + 1})^{x_{ij}} (1 - \frac{\theta_{mj} d_i}{\theta_{mj} d_i + 1})^{k_{mj}} \end{aligned} \tag{5.7}$$

Consider the ratio: $\frac{Pr(G=m|X=x_i)}{Pr(G=\ell|X=x_i)} = \frac{\pi_m f_m(x_i)}{\pi_\ell f_\ell(x_i)}$, thus we assign x_i to class m if $\pi_m f_m(x_i)$ is the maximum for $\pi_l f_l(x_i)$ for $l = 1, ..., M$.

In our case, $M = 2$ (sub-gingival plaques and supra-gingival plaques). Denote sub-gingival plaques as Class 1, supra-gingival plaques as Class 2, note that the decision boundary is:

$$\begin{aligned} & log\frac{p_1}{1-p_1} = log\frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} = log\frac{\pi_1}{\pi_2} + log\frac{f_1(x)}{f_2(x)} \\ & = log\frac{\pi_1}{\pi_2} + \sum_{j=1}^{p}[log\Gamma(x_{ij} + \hat{k}_{1j}) - log(\Gamma(\hat{k}_{1j}) + x_{ij}log(\frac{\hat{\theta}_{1j}d_i}{\hat{\theta}_{1j}d_i + 1}) \\ & + \hat{k}_{1j}log(1 - \frac{\hat{\theta}_{1j}d_i}{\hat{\theta}_{1j}d_i + 1})] - \sum_{j=1}^{p}[log\Gamma(x_{ij} + \hat{k}_{2j}) - log\Gamma(\hat{k}_{2j}) \\ & + x_{ij}log(\frac{\hat{\theta}_{2j}d_i}{\hat{\theta}_{2j}d_i + 1}) + \hat{k}_{2j}log(1 - \frac{\hat{\theta}_{2j}d_i}{\hat{\theta}_{2j}d_i + 1})] = 0 \end{aligned} \tag{5.8}$$

5.3 Results of Naïve Bayes Discriminant Analysis (NBDA)

Note that the order of the OTUs in Chapter 4 was calculated based on the whole data set. Strictly speaking we should have used the training data only to rank the variables. We separate the whole dataset (301 sub-gingival plaques and 305 supra-gingival plaques with 361 common OTUs) into training data and test data in about 2:1 ratio. The training data contain 200 sub-gingival plaque samples and 203 supra-gingival plaque samples, while the test data contain 101 sub-gingival plaque samples and 102 supra-gingival plaque samples. We use the training data to build a predictive model and use the test data to see how well the model performs on new samples.

We divide the training data into 10 equal sized subsets to do a 10-fold cross-validation. For each cross-validation process, the i^{th} $(i = 1, 2, ..., 10)$ set is used for testing while the other 9 sets are for training. For each cross-validation, the training data includes 180 sub-gingival plaque samples and 183 supra-gingival plaque samples while the test data has 21 sub-gingival plaque samples and 20 supra-gingival plaque samples. We estimate parameters $\hat{k}_{1j}, \hat{\theta}_{1j}$ of NB models from sub-gingival plaque samples and $\hat{k}_{2j}, \hat{\theta}_{2j}$ from supra-gingival plaque samples in training data. Based on parameters estimated in training data, we perform NBDA based on j^{th} $(j = 1, ..., 361)$

OTU individually and calculate each OTU's predictive error. That is, in test data, for i^{th} $(i = 1, ..., 41)$ sample j^{th} OTU, we calculate $log\frac{f_1(x_{ij})}{f_2(x_{ij})}$.

For each sample, if the probability from sub-gingival plaque is greater than the probability from supra-gingival plaque, we assign it into sub-gingival plaque and vice versa. Then we calculate this OTU's predictive error. Repeating the process for each fold of the cross-validation, a test error can be obtained for each sample based on each OTU. We arrange the test errors of each sample for each OTU from the 10-fold cross-validation in a predictive error matrix of m by p ($m = 10$,$p = 361$), the column means gives the final predictive error estimation for each OTU. We rank these 361 OTUs according to their final predictive errors, the smaller the predictive error is, the stronger predictive power the OTU has.

Now these 361 OTUs are ordered according to their predictive power, re-order the columns of the predictive error matrix according to the OTU predictive power order, add the first k terms for each row ($k = 1, ..., 361$), plus $log(\frac{\pi_1}{\pi_2})$ (here use training ratio $\frac{180}{183}$ to estimate $\frac{\pi_1}{\pi_2}$), this provides us the *log*-odds for using first k OTUs to predict. Then we can easily get a cross-validated error for using the first k OTUs to predict.

Recall that the one standard error rule is a way of choosing k from the CV-error curve, in which we choose the simplest model whose error is within one standard error of the minimal error. This would indicate that this much simpler model is not worse, at least not in a statistically significant way. Using the one standard error rule, choose $k = 3$ (OTU ID: $4459671, 4452538, 4338372$) as our best choice (see Figure 5.1). These three OTUs are all highly ranked in the LR test in Chapter 4 (top 5, top 6 and top 14), which verifies our guess that OTUs which have significantly different distributions between sub-gingival plaque samples and supra-gingival plaques play an important role in classification.

The classification error e depends on the number of samples incorrectly classified (false positives plus false negatives) and is evaluated by the formula:

$$e = \frac{f}{n} \tag{5.9}$$

where f is the number of sample cases incorrectly classified, and n is the total number of sample cases. For reference purpose, we perform NBDA sequentially by including the first k $(k = 1, ..., 361)$ of these 361 ordered OTUs and train the classifier on all of

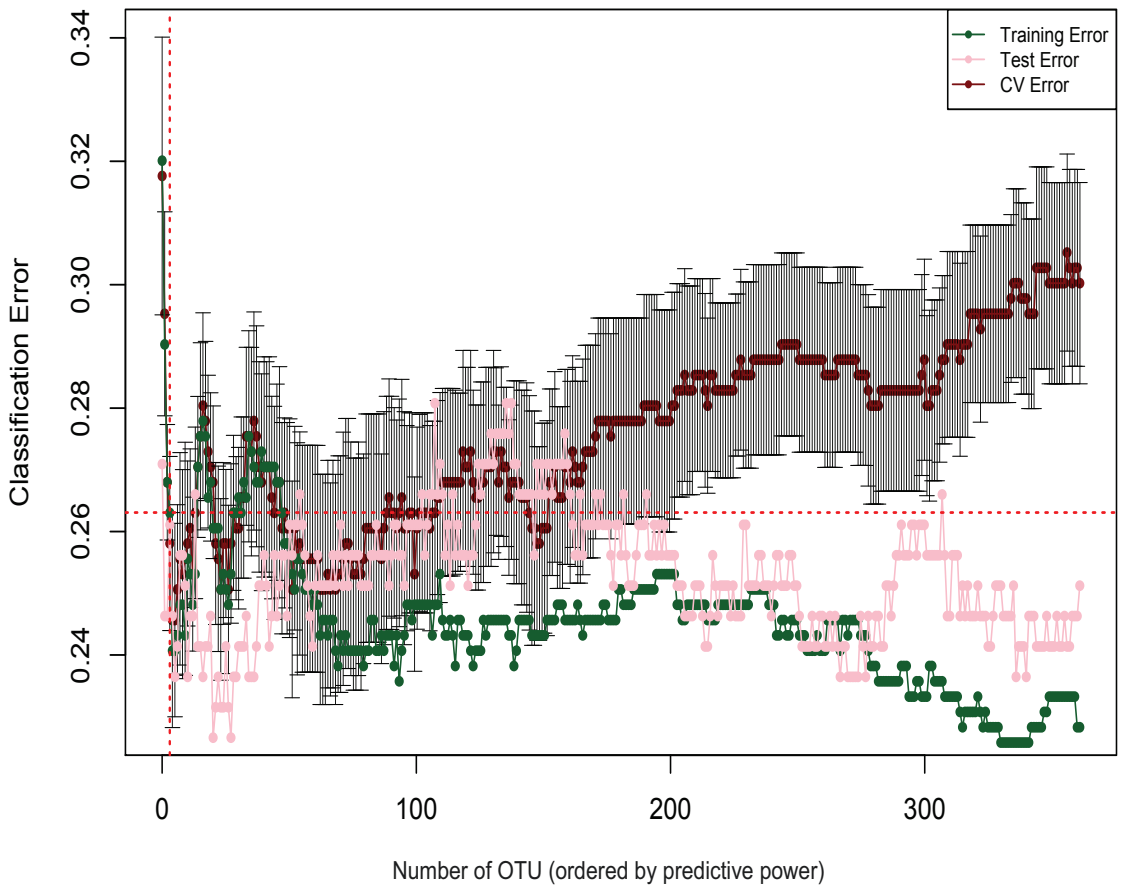

Figure 5.1: Classification error of Naïve Bayes Discriminant Analysis on sub-gingival plaque and supra-gingival plaque samples.

the training set to obtain the test errors for each k. The test errors are also presented in Figure 5.1. With k=3, we get the following training/test errors:

training error: 0.2679

test error: 0.2463

Interestingly, Figure 5.1 shows that using only the first significantly differently distributed OTU in the model, the training error is 0.32 and test error is 0.27. A binary classification of the sub-gingival and supra-gingival samples with features selection was performed in [21], Random Forest feature permutation showing the best performance of a trained model with an accuracy of 79.8% on 20 feature. After including

the first three significantly differently distributed OTUs in our NBDA model, the training error fluctuates within a narrow range and so does the test error. Note that by fixing the order of OTUs according to the predictive power, this simplifies the variable selection procedure, but this may not be an optimal choice, in which case it also confines the possibility of choosing the optimum OTUs for the discriminant analysis. Another element that can obviously improve the performance is to jointly model the OTUs instead of treating them as independent. However it is not obvious how to model the OTUs jointly. The work in this thesis indeed is the starting effort for the eventual goal of jointly modelling the OTUs with the phylogenetic tree relationship considered in the model.

The fact that only three OTUs can predict so accurately means that these three OTUs indeed are worth of being looked more carefully. After all, classification for this data is not the purpose of the analysis, we are not really interested in predicting a future observation is actually a sub-gingival plaque or supra-gingival plaque data point. The accuracy of the classification merely is used to demonstrate that the OTUs selected are the most important elements for two different communities. From this point of view, the method proposed in this thesis is far better than a black-box classification methods that can achieve slightly higher predictive accuracy.

Chapter 6

Application of LASSO

One assumption of NBDA is the independence of the variables, this could result in the reduced performance of NBDA. The resulted procedure of NBDA is to add the log probability difference for the selected k OTUs. One way to possibly improve this is to use a logistic regression on the log probability differences to select better linear coefficients to help partially addressing the problem of inaccurate assumption of independence. We choose to apply LASSO on these log probability difference scores to test if LASSO can also help to choose the best variables, if so, this could also be used to replace the ranking step in the NBDA procedure. As we know, LASSO (Least Absolute Shrinkage and Selection Operator) is a method that not only performs variable selection but also ranks as one of the top prediction methods as well [22]. Thus in addition to comparing the predictive accuracy, we can compare OTUs picked up by LASSO with those OTUs significantly differently distributed between sub-gingival plaques and supra-gingival plaques by the LR test in Chapter 4.

6.1 Review of LASSO

LASSO is a regression analysis method with ability to perform subset selection. Given a linear regression with standardized predictors $Z = (z_{ij})$, for $i = 1, 2, ..., N$ and $j = 1, 2, ..., P$, and response values $y_1, ..., y_i, ..., y_N$, the LASSO solves the ℓ_1-penalized regression problem of finding $\beta = (\beta_1, ..., \beta_j, ...\beta_P)$ to minimize

$$\sum_{i=1}^{N}(y_i - \sum_{j=1}^{P} z_{ij}\beta_j)^2 + c\sum_{j=1}^{P}|\beta_j|. \tag{6.1}$$

This is equivalent to minimizing the sum of squares with a constraint of the form $\sum_{j=1}^{P}|\beta_j| \leq \lambda$. λ is a tuning parameter, the larger λ, the smaller effect from the constraint. It is similar to ridge regression (ℓ_2 norm), which has constraint $\sum_{j=1}^{P}\beta_j^2 \leq$

t, t is a constant, it's a tuning parameter. Because of the form of the ℓ_1-penalty, the lasso is able to perform variable selection and shrinkage, while ridge regression only can do shrinkage. With λ value increasing, LASSO will output more non-zero coefficients, which corresponds to more variables selected.

The R package "glmnet" is used to solve the LASSO fitting. By setting the family="binomial", y should be a factor with two levels, in our case, with '1' indicating sub-gingival plaque, and '0' indicating supra-gingival plaque. By default in the package, the selection of OTUs is controlled by the regularization parameters λ, which is chosen by a cross-validation procedure on training data. The algorithm gives us a sequence of λ's with the associated cross-validated errors, there's one λ with the minimum mean cross-validated error, is usually chosen as the tuning parameter value and the variables with non-zero coefficients in the model at this tuning parameter value are the selected variables.

6.2 Application of LASSO on Four Different Types of Input variables

We apply LASSO on the oral cavity data using the same training dataset and test dataset as the NBDA, comparing with the predictive accuracy from NBDA and comparing the OTUs selected from LASSO with those OTUs significantly differently distributed between sub-gingival plaques and supra-gingival plaques in Chapter 4. We apply LASSO on four different types of input variables using the 361 OTUs in Chapter 2, and the same training data and test data as in Chapter 5.

In the first procedure, we perform LASSO regression analysis on the simple proportions of OTUs in sub-gingival plaques and supra-gingival plaques; in the second procedure, different from Chapter 2, where only one population (sub-gingival plaques was used) to estimate the posterior means, here we use data from both populations to estimate the Bayesian posterior means of OTUs, then applying LASSO on the Bayesian posterior means of OTUs estimated from sub-gingival plaques and supra-gingival plaques. The third procedure is performing a log-probability-difference transformation on the OTU data. That is, for j^{th} OTU of i^{th} observation, we calculate the NB probability of sub-gingival plaques $f_1(x_{ij}, k_{1j}; \theta_{1j})$, k_{1j} and θ_{1j} are estimated from the counts of j^{th} OTU in sub-gingival plaque training samples; meanwhile, we calculate the NB probability of supra-gingival plaques $f_2(x_{ij}; k_{2j}, \theta_{2j})$, k_{2j} and θ_{2j} are

estimated from the counts of j^{th} OTU in supra-gingival plaque training samples; then calculate the log difference of these two probabilities $log\frac{f_1(x_{ij};k_{1j},\theta_{1j})}{f_2(x_{ij};k_{2j},\theta_{2j})}$. We then apply LASSO on the log-probability-difference. This enables significant OTUs' ratio probability to be weighted by regression coefficients. The fourth LASSO procedure is based the mixture posterior means of OTUs as introduced in Section 3.3. When we use the posterior means or log probability difference as LASSO input, we estimate the prior information or the negative binomial parameters from the training data. We fix these parameters and pre-process the test data based on the training data parameters to get the test errors.

6.3 Results and Discussion

Figure 6.1 shows us the classification errors of LASSO based on simple proportions, and Bayesian posterior means estimated from sub-gingival plaques and supra-gingival plaques pooled together. The "glmnet" algorithm gives us a sequence of λ's (tuning parameter) with the associated cross-validated errors the λ with the minimum mean cross-validated error, is usually chosen as the tuning parameter value and the variables with non-zero coefficients in the model at this tuning parameter value are the selected variables. When applying LASSO on the simple proportion of OTUs, LASSO selects 38 variables into the model, corresponding test error is 0.2167. Among these 38 OTUs, 22 of them are significantly differently distributed between sub-gingival plaques and supra-gingival plaques by the LR test in Chapter 4. For the Bayesian posterior means estimated from sub-gingival plaques and supra-gingival plaques, LASSO selects 46 OTUs (33 of them have significantly different distributions between sub-gingival plaques and supra-gingival plaques) and the corresponding test error is 0.2266. In Figure 6.2, top graph is the classification error when applying LASSO on the log probability difference, the number of non-zero coefficients from LASSO is 45 (23 have significantly different distributions between sub-gingival plaques and supra-gingival plaques), and the test error is 0.2512; the bottom graph is the classification error when applying LASSO on the mixture posterior mean of OTUs, 36 OTUs are selected by LASSO, 19 of them are ranked as significantly differently distributed between sub-gingival plaques and supra-gingival plaques by LR test, and the test error is 0.2266.

Recall Naive Bayes Discriminant Analysis in Chapter 5, we only need three OTUs

Procedure	*Test.err*	*Training.err*	*Std.err*	*Selected.OTU*
NBDA	0.2463	0.2679	0.0302	3
		(0.2161,0.2765)		

Table 6.1: Classification error from NBDA

and achieve the test error as 0.2463, in Table 6.1, we use binomial distribution to estimate the standard error for the test error,

$$Std.err = \sqrt{\frac{p(1-p)}{n}}, \tag{6.2}$$

where $p = 0.2463, n = 203$. We can see that all the four test errors from the LASSO fall into $(0.2161, 0.2765)$, that is to say all this five test errors are consistent, this can demonstrate that the OTUs selected in NBDA are the most important elements for two different communities.

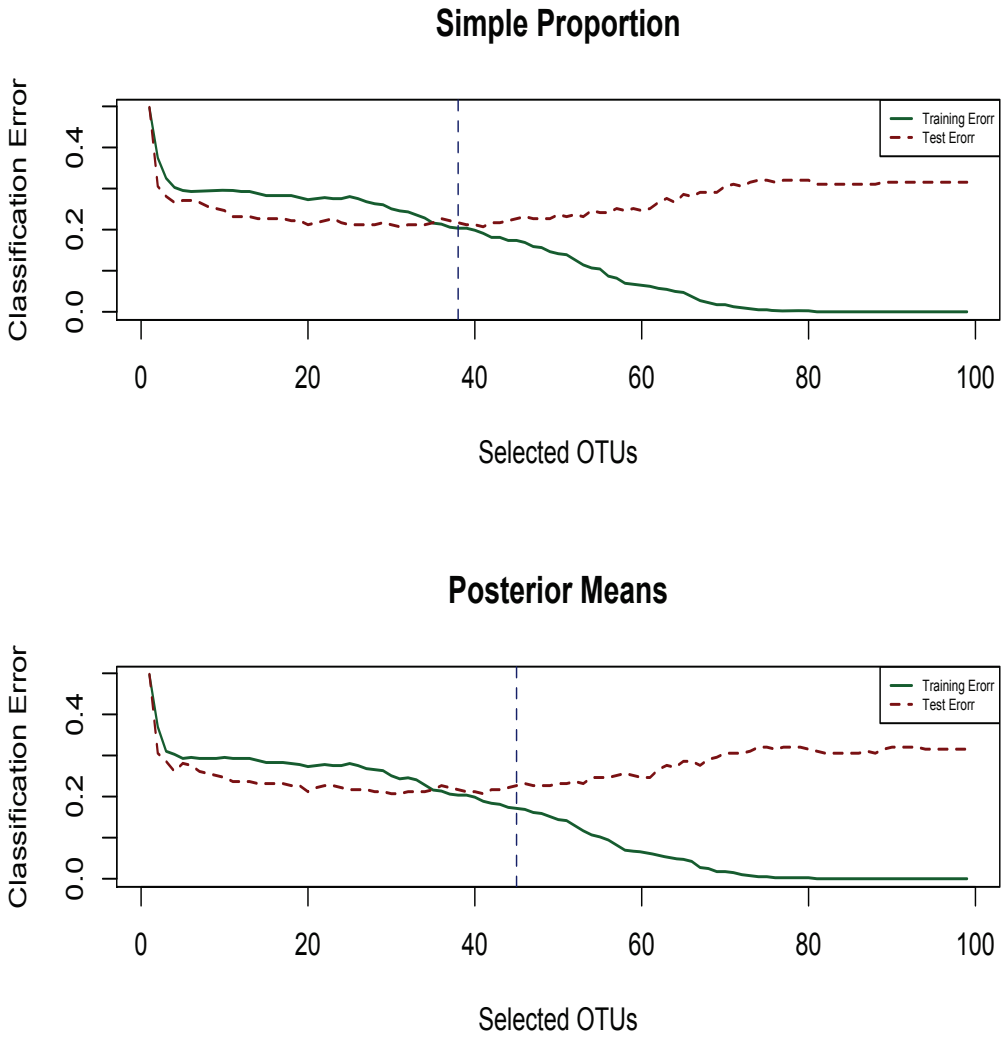

Figure 6.1: Top: classification error from simple proportions; bottom: classification error from Bayesian posterior means estimated from sub-gingival plaques and supra-gingival plaques.

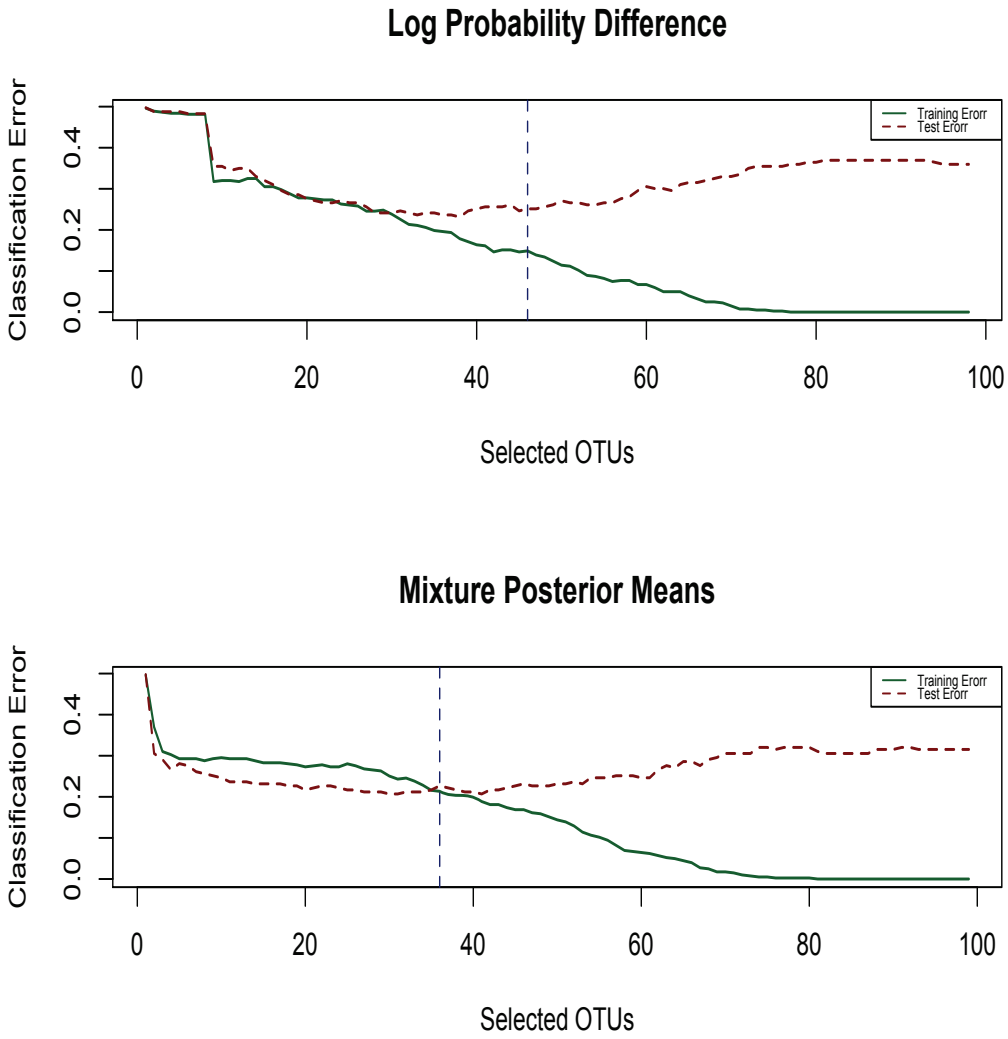

Figure 6.2: Top: classification error from log-probability-difference; bottom: classification error from Bayesian mixture posterior means.

Chapter 7

Conclusion

In this thesis, we first model the OTU data with Negative Binomial (NB) model and fit the MLE's for NB model parameters. The results of parametric bootstrap of NB model and the empirical data distributions show that NB model could model some part of the OTU count data well, this enables us to better estimate the underlying composition for these OTUs. Then we perform the empirical Bayesian inference for the underlying composition of these OTUs, try to visualize overall differences in bacterial composition between sample groups through the PCA plots. The simple proportions contain so many 0's that we are unable to perform a log-transformation on the simple proportions. The Bayesian posterior means change the simple proportions which are 0's into slightly non-zero's. Using the Bayesian posterior mean instead of the commonly used simple proportion normalization enables us to perform a log-transformation on the posterior means. The log-transformation on the posterior means provides a better PCA plot, which helps us explore the data in a better way.

The LR test can be used to determine that some NB models estimated from subgingival plaques are significantly different from the NB model estimated from supragingival plaques for the same OTUs. We find those OTUs with strong predictive power and develop the Naive Bayes Discriminant Analysis (NBDA) based on the NB distributions of those OTUs. By fixing the order of OTUs according to the predictive power, NBDA can effectively reduce the dimensionality of the data, choosing the optimum OTUs for discriminant analysis. After that, we apply LASSO to several transformation of the estimated underlying compositions of OTUs. And our developed NBDA and LASSO verify that those OTUs play a critical role in classification.

In our thesis, we assume OTUs to be independent, one element that can obviously improve the performance is to jointly model the OTUs instead of treating them as independent. However it is not obvious how to model the OTUs jointly. The future work of this thesis is to jointly modelling the OTUs with the phylogenetic tree

relationship considered in the model.

Bibliography

[1] Floyd E Dewhirst, Tuste Chen, Jacques Izard, Bruce J Paster, Anne CR Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G Wade. The human oral microbiome. *Journal of bacteriology*, 192(19):5002–5017, 2010.

[2] Hiroshi Mori, Fumito Maruyama, and Ken Kurokawa. Vitcomic: visualization tool for taxonomic compositions of microbial communities based on 16s rrna gene sequences. *BMC bioinformatics*, 11(1):1, 2010.

[3] Catherine A Lozupone, Micah Hamady, Scott T Kelley, and Rob Knight. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and environmental microbiology*, 73(5):1576–1585, 2007.

[4] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, 2013.

[5] Jun Lu, John K Tomfohr, and Thomas B Kepler. Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC bioinformatics*, 6(1):1, 2005.

[6] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

[7] Elisabeth M Bik, Clara Davis Long, Gary C Armitage, Peter Loomer, Joanne Emerson, Emmanuel F Mongodin, Karen E Nelson, Steven R Gill, Claire M Fraser-Liggett, and David A Relman. Bacterial diversity in the oral cavity of 10 healthy individuals. *The ISME journal*, 4(8):962–974, 2010.

[8] Angelo Mariotti. Dental plaque-induced gingival diseases. *Annals of periodontology*, 4(1):7–17, 1999.

[9] Rodney Needham. Polythetic classification: convergence and consequences. *Man*, pages 349–369, 1975.

[10] Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531, 2014.

[11] Andreas Lindén and Samu Mäntyniemi. Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7):1414–1421, 2011.

[12] Gary C White and Robert E Bennetts. Analysis of frequency count data using the negative binomial distribution. *Ecology*, 77(8):2549–2557, 1996.

[13] Peter J Turnbaugh, Ruth E Ley, Micah Hamady, Claire Fraser-Liggett, Rob Knight, and Jeffrey I Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449(7164):804, 2007.

[14] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–1202, 2013.

[15] Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

[16] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[17] R John Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[18] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[19] Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

[20] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[21] Robert G Beiko. Microbial malaise: How can we classify the microbiome? *Trends in microbiology*, 23(11):671–679, 2015.

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

[23] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq:an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[24] Bruce J Paster, Susan K Boches, Jamie L Galvin, Rebecca E Ericson, Carol N Lau, Valerie A Levanos, Ashish Sahasrabudhe, and Floyd E Dewhirst. Bacterial diversity in human subgingival plaque. *Journal of bacteriology*, 183(12):3770–3783, 2001.