Predictive Analysis for Microbiome Data Related to Parkinson Disease

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Man Chen

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
142 Mcclean Hall, 106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5E6
Canada
OR
Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
1 16 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

There are a lot of studies indicating that human gut microbiota is likely to have connections with Parkinson disease (PD). Based on these indications, this thesis explores the association between PD and human gut microbiota, from a statistical machine learning perspective. With the purpose of identifying the association between PD and gut microbiota, we assess the predictivity of microbial operational taxonomy units (OTUs) that are extracted from participants' gut. The raw dataset is 16s rRNA sequencing data from 327 participants' gut. Then, we pick OTU feature. OTU dataset contains high dimensional counts with excessive zeros.

In analysis part, we use linear support vector machine (SVM) and logistic regression combined with L1 penalty and elastic-net penalty, to identify informative OTUs for PD. L1 penalty is able to do shrinkage for features, which effectively implements feature selection by enabling coefficients of non-significant variables to be zero. Elastic-net penalty is capable of doing grouping effect, in case of ignorance of some correlated variables. Under these two penalties, SVM and logistic regression can achieve sufficiently good predictive results as well as feature selection.

In order to make full use of dataset and to avoid overfitting, cross-validation (CV) is implemented. We compare internal cross-validation (ICV) with external cross-validation (ECV). ICV chooses a subset of OTUs based on all samples first. These selected OTUs are plugged into predictive models as predictors. The test accuracy of predictive model with ICV is overestimated. The cause of biased accuracy is that the training samples are not external to test samples. Instead of selecting features based on all samples, ECV does feature selection only based on training samples in each fold, which is much more corrective. We apply leave one out cross-validation (LOOCV) to choose tuning parameters.

We discuss several evaluation metrics, of which error rate is employed to choose optimal values of tuning parameters. Logistic regression with L1, elastic-net, and SVM with L1, elastic-net demonstrate good predictive results. Error rates of these models with optimal λ are 0.223, 0.238, 0.278 and 0.275, respectively. We select significant OTUs based on coefficients. Finally, we find overlaps between our selections and others' results.

Acknowledgements

Contents

List of Tables

List of Figures

List of Abbreviations

PD	Parkinson Disease
OTU	operational Taxonomy Unit
SVM	Support Vector Machine
CV	Cross Validation
ICV	Internal Cross Validation
ECV	External Cross Validation
LOOCV	Leave One Out Cross Validation
AMLP	Average Minus Probability
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
PRC	Precision Recall Curve
AUPR	Area Under the Precision Recall Curve
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
KKT	Karush Kuhn Tucker

Chapter 1

Introduction

1.1 Background and Motivation

Parkinson disease (PD) was first proposed by Dr. James Parkinson in 1817. At that time, he called this disease "shaking palsy". Nowadays, PD becomes a hot topic, which has caused huge damages to patients' normal life. PD is a long-term progressive disorder of central nervous system. Among all neurodegenerative disorders, PD is relatively common. According to the information from Parkinson Disease Foundation, there will be more than 1 million American PD patients by 2020, and more than 10 million people worldwide are living with PD. Generally, the incidence of PD increases with age. Most PD patients are aged people, but an estimated 4% of people with PD are diagnosed before age 50.

With PD, patients tend to suffer a series of movement disorder. In the early stage of PD, the most obvious symptoms are tremor, muscle stiffness, rigidity in facial expressions, slow movement and difficulty with balance. Gradually, with the progression of PD, some sever symptoms occur. PD patients in advanced stage tend to suffer from a series of emotional and thinking behaviors, such as depression, anxiety, difficulty with attention and memory[1][2]. Besides, when the condition becomes severer and severer, some other problems also happen, like sleep disturbance and dementia [2].

So far, the exact cause of Parkinson's is still unclear. However, both genetic factors and environmental factors are believed to be main causes. In some researchers' opinion, virus is also one of main triggers for PD [3]. Scientists also find some signs linked to PD. The low-level dopamine and the accumulation of an abnormal protein, Lewy bodies, have been found as signs related to PD. Usually, the death of cells in specified area of the midbrain will consequently lead to motor symptoms of the disease and inadequate dopamine in this

area [1]. Even though we know it can also accumulate Lewy bodies [3], the exact reason why these cells die is hardly known.

Despite of inexact causes of PD, scientists recognize a series of factors which let people be more likely to get PD. The probability that men get PD is one and a half times higher than that of women. Compared with Whites, African Americans or Asians are less likely to get PD. PD usually happens between the ages of 50 and 60. Only 5-10 % patients are before 40. Heredity is also a potential factor in causing PD, which means people having close families with PD are more likely to suffer from PD. Head injury can also increase the risk of getting PD [3].

Recently, relationship between PD and human gut has drawn researchers' attention. Some gastrointestinal symptoms, such as drooling, dysphagia, constipation and defecatory dysfunction, have often been observed before motor signs of PD [4]. Gastrointestinal inflammation also occurs in PD patients [5]. These phenomena result in assumptions that there is a link between PD and gut, and that the human gut microbiota can be a potential source of novel therapeutics. These assumptions inspire researchers' more studies on relationship between gut and PD, more specifically, between PD and gut microbiome.

The gut microbiome, defined by biologist Joshua Lederberg, is the total collections of microorganisms, bacteria, viruses, protozoa, fungi, and their collective genetic materials present in the gastrointestinal tract. The gut microbiota is consist of all the bacteria, both commensal and pathogenic, residing in the gastrointestinal tract. There are tens of trillions of microorganisms in human gut, including more than 1000 bacterial species [6][7]. With the development of bioinformatics and sequencing technologies in recent decades, researchers have received much more opportunities to explore gut microbiome. Investigations of human gut microbiota have been extended beyond classical infectious diseases. Studies have shown changes in the gut microbiota during obesity, diabetes, metabolic disorder and other diseases [8–11]. An well-proportioned microbiome plays an important role in immune system preventing some specific disease [10][11].

Furthermore, various studies have indicated links between gut microbiota and neurodegenerative diseases, such as PD. Intestinal microbiome is altered in PD and is related to motor phenotype [12]. Microbial dysbiosis happens during PD and can lead to inflammation-

induced misfolding of α-synuclein and development of PD pathology [13]. Compared with healthy controls, PD patients tend to have less bacterial phylum Bacteroidetes and the bacterial family Prevotellaceae is also less, while Enterobacteriaceae are more abundant in fecal samples from PD patients [14]. Changes in abundances of 9 genera and 15 species of microorganisms happen during the process of PD [15], triggering local inflammation and consequent aggregation of α-synuclein and Lewy bodies, which are typical features of PD.

All in all, a variety of studies reveal the evidence that PD is close related with gut microbiome. On this basis, this thesis continues to conduct further exploration on the association between PD and microbiome, from a statistical machine learning perspective. We attempt to find the decisive microbiome patterns that are highly related with PD. When it comes to predictive analysis of individuals' phenotype based on microbial data, finding ideal models is challenging but important. An optimal predictive model can not only guarantee a numerically accurate result, but also provide variables or factors that are related to PD.

1.2 Main Methods

Typically, microbial abundance is quantified by Operational Taxonomy Unit (OTU). The OTU dataset is high-dimensional and over-dispersed , with excessive zeros. Effective predictive models that can deal with OTU data properly and do feature selection are necessary. To explore the association between PD phenotype and gut microbiome, we use logistic regression and linear support vector machine (SVM) combined with Lasso regularization and elastic-net regularization. Lasso is able to conduct feature selection after a process of shrinkage. By adding L1 penalty to loss function, this regularization enables the coefficients of non-significant features to be zero. Elastic-net regularization is a linear combination of L1 penalty and L2 penalty. When there is a group of correlated variables, Lasso often keeps one of them and ignores others, while elastic-net can consider all correlated variables in the group. In Lasso and elastic-net regularizations, different tuning parameter λ has different predictive performances. Hence, we need to find the most optimal λ for each model.

It's incorrect to use the same dataset for learning the parameters of a predictive model and testing predictive performance, because by doing this, models having perfect predictive results

may fail to predict accurately on yet-unseen data. This phenomena is called overfitting. In order to avoid overfitting, cross validation (CV) is conducted to learn the parameters of our predictive models. Internal cross validation (ICV) is widely used. However, the accuracy of ICV is cheating, because the process of selecting features has already involved the test samples. Thus, we apply external cross validation (ECV), which keeps test samples external during the process of variable selection. The predictive result of ECV looks like worse than that of ICV, but ECV's result is more convincing and reliable. Leave one out cross validation (LOOCV) is a special case of CV, which can make full use of all samples in dataset. In this thesis, we apply LOOCV to regularized classification models, for the purpose of finding optimal λ. Test error rate is the main criterion for choice of λ. According to test error rates, we choose the best λ of each model. Then, in each optimal model, variables with non-zero coefficients are chosen as significant features. AMLP, ROC curves and PRC curves are alternative metrics for evaluation of predictive models.

1.3 Summary of Results

After applying regularized logistic regression and SVM, we obtain the error rates of these models with different λ. A λ associated with minimum error rate is chosen as the optimal one. For optimal logistic models with L1 penalty and elastic-net penalty, error rates are 0.223 and 0.238, respectively; AUC are 0.826 and 0.820, respectively; AUPR are both 0.879; AMLP are 0.515 and 0.506, respectively. For optimal SVM with L1 penalty and optimal SVM with elastic-net penalty, error rates are 0.278 and 0.275; AUC are 0.777 and 0.760; AUPR are 0.832 and 0.831; AMLP are 0.550 and 0.562.

In different folds of LOOCV, coefficients of OTUs are different. Thus, for each OTU, we find the median of absolute value of its all coefficients. After ranking the median, the OTUs associated with non-zero median of absolute value are considered as significant features related to PD. Boxplot is a proper tool to present the different distributions of OTU's relative abundances between PD patients and control cases. Finally, we can find some consistencies between our results and other published studies.

1.4 Outline

Chapter 2 is an explicit description of predictive models for OTU dataset related to PD. including logistic regression, SVM, Lasso regularization, elastic-net regularization, and cross validation. We can see how regularizations help models to do feature selection. As for cross validation, comparison of ECV and ICV tells us why ECV has more reliable results than ICV. Evaluation metrics are also described explicitly in Chapter 2, including error rate, AMLP, ROC, AUC, PRC, and AUPR. These metrics provide the comprehensive evaluation of our models. In Chapter 3, we give the introduction to our real dataset. The raw dataset is 16r RNA sequencing data. The tool to extract OTU from sequencing data is also introduced in Chapter 3. In Chapter 4, we report our results. The performances of our models are discussed in this chapter. What's more, we also present the decisive OTUs selected by our model. Discussion on the comparison of our results with other researchers' results is needed, to assess how credible our results are. The conclusion is in Chapter 5. Besides, we also talk about drawbacks of our work and future work.

Chapter 2

Methodology

2.1 Models

2.1.1 Data Structure

Table 2.1: Overview of Data for Moldes

	OTU_1	OTU_2	.....	OTU_q	Age	Sex	Parkinson	Total reads
$Sample_1$	$otu_1^{(1)}$	$otu_1^{(2)}$	.....	$otu_1^{(q)}$	age_1	sex_1	Y_1	$T_1 = \sum_{j=1}^{q} otu_1^{(j)}$
.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....
$Sample_n$	$otu_n^{(1)}$	$otu_n^{(2)}$	.....	$otu_n^{(q)}$	age_n	sex_n	Y_n	$T_n = \sum_{j=1}^{q} otu_n^{(j)}$

In this thesis, we explore the association between PD and human gut microbiome, by establishing models for OTU variables, environmental variables and response variable, the phenotype. Environmental variables are age and sex. Age is the numerical variable to measure how old the subject is. Sex is the binary categorical variable to denote the gender of a subject, male or female. Phenotype variable is also binary categorical variabale to describe whether a subject is PD patient, Parkinson cases or healthy control.

OTU, the operational taxonomic unit, is an operational definition for classifying groups of closely related individuals. Originally, this notation was proposed by Robert R. Sokal and Peter H. A. Sneath [16]. At first, "Operational Taxonomic Unit" was a pragmatic definition to group individuals according to similarity. Nowadays, OTU is also widely seen in different backgrounds. The meaning of OTU has been extended into a different context. Presently,

OTU also refers to clusters of microorganisms, grouped by DNA sequence similarity to a specific taxonomic marker gene. Assigning OTU is the process to cluster similar sequences to into operational taxonomic units based on a defined similarity threshold. Sequences that are similar at or above the threshold level are grouped into a taxonomic unit in the sequence collection. After assigning OTU, we obtain OTU dataset. OTU variables can be denoted by $(otu_i^{(1)}, otu_i^{(2)}, .., otu_i^{(q)})$. $otu_i^{(j)}$ means that, for the i_{th} subject, the number of sequences that are grouped into the j_{th} OTU. Apparently, OTU dataset is consist of non-negative integers, including lots of zero.

2.1.2 Logistic Regression

Logistic regression, a kind of generalized linear model, is used to predict the probabilities of certain classes or events, such as pass/fail, healthy/sick, negative/positive. Logistic regression is widely applied to analyze data, especially in medical areas [17–21].

Logistic regression can be extended for multiple classes. In our study, it's a binary classification, Parkinson/non-Parkinson, so we discuss the binary case of logistic regression. Suppose for each observation, we have a series of features including OTU features, age and sex, denoted by $X_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(p)})$. The response variable Y means whether it's Parkinson case or not, with 1 denoting getting PD, 0 denoting healthy case. Then the prediction for the probability of getting PD is:

$$P(y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... \beta_p x_i^{(p)}}}{1 + e^{\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + ... \beta_p x_i^{(p)}}}. \tag{2.1}$$

Its negative log-likelihood function is:

$$l = -\sum_{i=1}^{n} log(P_\beta(Y_i|X_i)) = -\sum_{i=1}^{n}(y_i(\beta_0 + \sum_{j=1}^{p}\beta_j x_i^{(j)}) - log(1 + e^{\beta_0 + \sum_{j=1}^{p}\beta_j x_i^{(j)}})). \tag{2.2}$$

Its loss function is:

$$J = -\frac{1}{n}\sum_{i=1}^{n}(y_i(\beta_0 + \sum_{j=1}^{p}\beta_j x_i^{(j)}) - log(1 + e^{\beta_0 + \sum_{j=1}^{p}\beta_j x_i^{(j)}})). \tag{2.3}$$

We obtain coefficients of logistic regression by minimizing its loss function.

2.1.3 Linear Support Vector Machine

Support vector machine is a supervised machine (SVM) learning algorithm that finds a hyperplane to distinctly separate data points into two groups. A hyperplane is a subspace of dimension p-1 in a p-dimensional space. Particularly, in a 2D space, the hyperplane is a line, and in a 3D space, it's a plane. For SVM, the number of dimensions, p, is the number of features. Given dataset (X_i, Y_i) for $i = 1...n$, with $X_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(p)}) \in R^p$ and $Y_i \in \{-1, 1\}$, SVM is a classification denoted by

$$\begin{cases} \hat{y_i} = +1, & f(X_i) \geq 0 \\ \hat{y_i} = -1, & f(X_i) < 0 \end{cases} \quad i = 1, 2, 3...n, \tag{2.4}$$

where $f(X) = \boldsymbol{w}^T X + b$ is the hyperplane. Here, $y = +1$ denotes getting PD and $y = -1$ denotes the healthy case.

2.1.3.1 Hard Margin

A good SVM model can not only correctly separate points into two sides but also separate them as stable as possible, which means points in two sides are divided by a clear distance. The distance between hyperplane and the closet point should be as large as possible. This idea is called SVM with maximum margin. Maximization of the margin distance can make sure the stability and reinforcement, so that data can be linearly separable with more confidence. Based on the idea of separating data points as accurately as possible with a maximization of the margin, SVM can be modeled as:

$$\begin{aligned} &max_{\boldsymbol{w},b} \ \frac{2}{\|\boldsymbol{w}\|}, \\ &subject\ to\ y_i(\boldsymbol{w}^T X_i + b) > 1, i = 1, 2, ..., n. \end{aligned} \tag{2.5}$$

Function 2.5 is equivalent to to:

$$\begin{aligned} &min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2, \\ &subject\ to\ y_i(\boldsymbol{w}^T X_i + b) > 1, i = 1, 2, ..., n. \end{aligned} \tag{2.6}$$

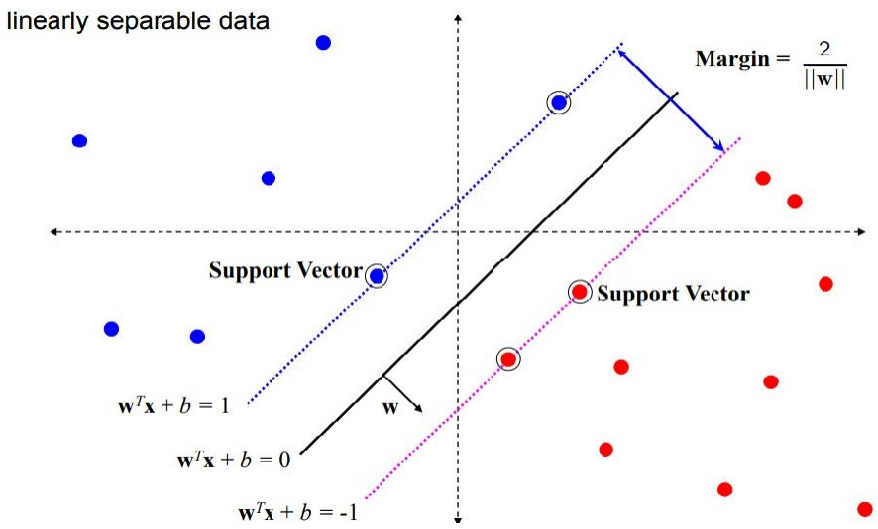


Figure 2.1: Overview of SVM

2.1.3.2 Soft Margin

When we deal with a real dataset, data points are not always as ideal as we expect. It's almost impossible that all points can be separated into two distinct parts without any mistakes. Thus, the soft margin is proposed for fault tolerance. On the basis of hard margin in section 2.1.3.1, the intuitive idea of soft margin is to allow SVM to have a certain number of misclassified points for building an margin as wide as possible, so that the remaining points can be separated correctly. After slight alterations based on function 2.5, the more flexible SVM with soft margin can be written as:

$$\begin{aligned} &min_{\boldsymbol{w},b}\ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\varepsilon_i \\ &subject\ to\ y_i(\boldsymbol{w}^T X_i + b) > 1 - \varepsilon_i, \\ &\varepsilon_i \geq 0, i = 1, 2..., n. \end{aligned} \tag{2.7}$$

ε is called slack variable, which is the distance between misclassified point and the side of its true class in the margin. If a point is classified correctly, then its corresponding ε is 0. $C > 0$ is the error penalty parameter.

2.1.3.3 Hinge Loss Function

From a SVM model like:

$$\begin{cases} \hat{y}_i = +1, & \boldsymbol{w}^T X_i + b \geq 0 \\ \hat{y}_i = -1, & \boldsymbol{w}^T X_i + b < 0 \end{cases} \quad i = 1, 2, 3...n, \tag{2.8}$$

$y_i(\boldsymbol{w}^T X_i + b) \geq 1$ means the point i is classified correctly, so the "loss" of this case is 0. Otherwise, $y_i(\boldsymbol{w}^T X_i + b) < 1$ indicates the classification is not perfect enough and then the "loss" is defined as $1 - y_i(\boldsymbol{w}^T X_i + b)$. The "loss" of point i is $max(0, 1 - y_i(\boldsymbol{w}^T X_i + b))$. Therefore, the loss function of SVM is denoted by:

$$L_{hinge\ loss} = \sum_{i=1}^{n} max(0, 1 - y_i(\boldsymbol{w}^T X_i + b)) = \sum_{i=1}^{n} [1 - y_i(\boldsymbol{w}^T X_i + b)]_+. \tag{2.9}$$

The definition of $[x]_+$ is as the following:

$$[x]_+ = \begin{cases} x & x > 0 \\ 0 & x \leq 0. \end{cases}$$

This loss function is termed as Hinge Loss. Alternatively, the function of 2.7 can be written as

$$min[L_{hinge\ loss} + \lambda \|\boldsymbol{w}\|^2]. \tag{2.10}$$

They are equivalent.

2.1.4 Lasso Regularization

The least absolute shrinkage and selection operator (LASSO) is a famous algorithm for regularization and feature selection. By adding the L1 penalty into loss function, LASSO can set a constraint on the sum of the absolute values of coefficients. The sum needs to be less than a upper bound. Thus, Lasso applies a shrinking process that penalizes the coefficients of features, enabling some of them to zero [22]. Lasso regularization is also named as L1 regularization. For Logistic Regression, Lasso regularization can be conducted by penalizing the loss function with L1-norm:

$$J_{lasso} = -\frac{1}{n} \sum_{i=1}^{n} (y_i(\beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)}) - log(1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)}})) + \lambda \|\boldsymbol{\beta}\|_1 \tag{2.11}$$

Lasso estimator of Logistic Regression is defined as:

$$argmin(-\frac{1}{n}\sum_{i=1}^{n}(y_i(\beta_0+\sum_{j=1}^{p}\beta_j X_i^{(j)})-log(1+e^{\beta_0+\sum_{j=1}^{p}\beta_j X_i^{(j)}}))+\lambda\|\boldsymbol{\beta}\|_1). \tag{2.12}$$

$\|\boldsymbol{\beta}\|_1$ is L1 norm of $\boldsymbol{\beta}$, equal to $\sum_{j=1}^{p}|\beta_j|$. β_0 is the intercept item, which is not included in the process of penalizing.

Similarly, Lasso estimator of linear SVM is defined as:

$$argmin[\sum_{i=1}^{n}[1-y_i(\boldsymbol{w}^T X_i+b)]_+ +\lambda\|\boldsymbol{w}\|_1]. \tag{2.13}$$

L1 penalty can shrink the coefficients associated with less important features into exactly zeros. In our study, OTUs with non-zero coefficients are treated as significant OTUs. λ is the tuning parameter, controlling the strength of the penalty. When λ is sufficiently large, then strength of penalty becomes more intensive, resulting in more coefficients being forced to be zero. Intuitively, after Lasso, higher the absolute value of a coefficient is, more significant the associated variable is.

2.1.5 Elastic-net Regularization

The elastic-net penalty is a combination of a L1-norm penalty with a L2-norm penalty.

$$\lambda_1\|\boldsymbol{\beta}\|_1+\frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2. \tag{2.14}$$

$\|\boldsymbol{\beta}\|_2^2$ is $\sum_{j=1}^{p}\beta_j^2$, called L2 norm. λ_1 and λ_2 are the tuning parameters. In a package in R [23], function 2.14 is denoted by

$$\lambda[\alpha\|\boldsymbol{\beta}\|_1+(\frac{1-\alpha}{2})\|\boldsymbol{\beta}\|_2^2]. \tag{2.15}$$

Here $\alpha=\frac{\lambda_1}{\lambda_2+\lambda_1}$ and $\lambda=\lambda_2+\lambda_1$. Then α is the mixing parameter and λ is the tuning parameter. When $\alpha=1$, the regularization becomes L1 regularization, and when $\alpha=0$, it's ridge regularization only with L2 penalty. In R, we can choose an appropriate value of α according to our dataset, and then choose optimal tuning parameter.

In elastic-net regularization, L1 norm penalty is to enable feature selection, whereas L2 norm penalty is to enable correlated features to be selected together. By L2 penalty, highly

related features tend to get similar coefficients [24]. This effect is the grouping effect. In our case, the presence of PD is characterized by patterns of gut microbial data. Microbial data is featured by high-dimensional OTU. It's possible that some OTUs are highly correlated. With grouping effect of elastic-net regularization, those correlated OTU features will be selected together, and non-zero coefficients will become more interpretable.

So, elastic-net estimators of logistic regression and linear SVM are:

$$argmin(-\frac{1}{n}\sum_{i=1}^{n}(y_i(\beta_0+\sum_{j=1}^{p}\beta_j X_i^{(j)})-log(1+e^{\beta_0+\sum_{j=1}^{p}\beta_j * X_i^{(j)}}))+\lambda(\alpha\|\boldsymbol{\beta}\|_1+(\frac{1-\alpha}{2})\|\boldsymbol{\beta}\|_2^2)) \tag{2.16}$$

and

$$argmin(\sum_{i=1}^{n}[1-y_i(\boldsymbol{w}^T X_i+b)]_+ +\lambda(\alpha\|\boldsymbol{\beta}\|_1+(\frac{1-\alpha}{2})\|\boldsymbol{\beta}\|_2^2)). \tag{2.17}$$

From the function 2.10, we can see that the original maximum-margin SVM model has already had L2 regularization inherently. Algorithms to solve out functions 2.12, 2.13, 2.16 and 2.17 have already been explicitly introduced by Trevor Hastie and Rob Tibshirani [22][25].

2.1.6 Transformation of OTU Data

In microbiome analysis, people always assume that the phenotype can influence the relative composition or relative abundance of OTUs, rather than original counts of OTUs. Therefore, when we fit models to OTU dataset, a reasonable transformation for OTU data is necessary, changing count number into information of relative abundances. Here, we choose one variance-stability transformation of the proportion:

$$\widetilde{X}_i^{(j)} = log(X_i^{(j)}+1)-log(\sum_{j=1}^{p}X_i^{(j)}+p). \tag{2.18}$$

2.2 Cross Validation

Cross-validation is a statistical method to evaluate the performance of learning algorithms by dividing data into two segments. One is for learning or training the model and the other is for validating the model. There are two popular forms of Cross Validation, k-fold cross-validation and Leave p out cross-validation. k-fold cross-validation means randomly partition

all samples into k equal-sized and non-overlapping parts. One of k parts is regarded as the validation data for testing the model, and the remaining k-1 parts are used as training set. K-fold cross-validation process repeats k times, with each of the k subsets being used once as the validation set. Leave p out cross validation means using p observations as the validation set and all remaining observations as the training set. This process repeats on all ways and iterates for each p points of dataset. Leave p out cross validation trains and validates a model C_p^n times. When $p = 1$, it's leave one out cross validation, which iterates for each data point.

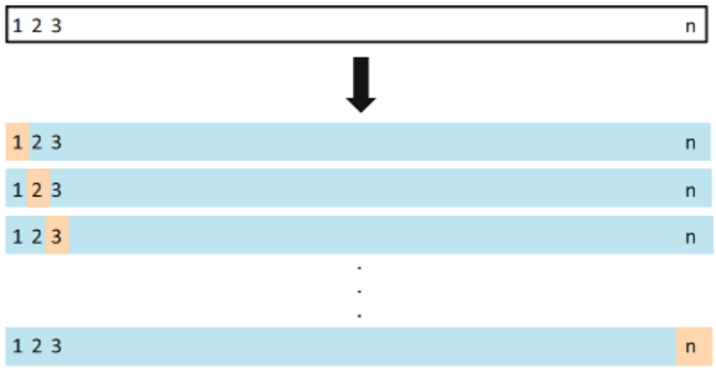


Figure 2.2: Overview of Leave-One Out Cross Validaion

2.2.1 Internal Cross Validation

Wen we conduct CV for dataset, internal cross validation (ICV) makes test samples internal to the feature selection process. The steps of ICV are:

- Do feature selection based on all samples, keep the selected features.
- Divide the samples into K equal-sized and non-overlapping groups randomly.
- For each k = 1, 2, ..., K, use the selected features to conduct CV, choose tuning parameters, train models based on the samples except those in fold k, and make predictions on group k.

Actually, using ICV is not "fair" enough. The result of test accuracy may be higher but it's a cheating result. Selecting features based on all samples before CV means that, at the beginning, these selected features have already known all left out test samples. Then, during CV process, test samples are not really independent and unknown.

2.2.2 External Cross Validation

Unlike ICV, external cross validation (ECV) makes test samples external to the step of feature selection. ECV is a more reasonable way to carry out cross validation. The steps of ECV are:

- Divide the samples into K equal-sized and non-overlapping groups randomly.
- For each k=1,2..K, use all of the groups except group k to train the model and select significant features.
- Make predictions on samples in group k based on the model built by training dataset, evaluate performance and choose optimal tuning parameter.

While ICV involves selecting features based on all samples, ECV is to use different samples in different training fold to do feature selection. Generally, CV should be applied into the entire process of modeling steps. Samples should be "left out" before selection or filtering steps. Obviously, the test error rate of ICV would be lower than ECV. However, it's a deceptive result, because those selected variables have already seen test samples before. ECV selects different subsets of features in different fold. For ECV, the test set is totally left out, and is only seen in the processes of validation and evaluation. In short, ECV is more reliable than ICV.

2.3 Predictive Metrics

2.3.1 Error Rate and AMLP

In order to evaluate predictive performance, we need to compare our predicted results with the true cases, and to rely on several evaluation criteria. One of the criteria is error rate . Error rate means the proportion of wrong predictions, which is defined as:

$$ER = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i), \tag{2.19}$$

where y_i is the observed case of sample i, and $\hat{y}_i$ is predicted case of sample i.

Another metric, the average of minus log predictive probabilities (AMLP) is:

$$AMLP = -\frac{1}{n}\sum_{i=1}^{n} log(\hat{P}_i(y_i|x_i)). \tag{2.20}$$

AMLP is more sensitive than error rate because it measures not only the correctness of a point estimate $\hat{y}_i$ but also the level of correctness expressed by $\hat{P}_i(y_i|x_i)$.

Both error rate and AMLP should be compared with their baseline values. The baseline is the result of a very intuitive and straightforward solution. Generally, baseline is only a random guess based on the frequency of observed y_i without using any models and predictors. For all observations, the frequency of $y_i = 1$ is denoted by $f_1 = \frac{1}{n}\sum_{i=1}^{n} I(y_i = 1)$, and the frequency of $y_i = 0$ is denoted by $f_0 = \frac{1}{n}\sum_{i=1}^{n} I(y_i = 0)$. Then the baseline error rate is $ER_{(0)} = min\{f_0, f_1\}$ and the baseline AMLP is $AMLP_{(0)} = -[f_0 log(f_0) + f_1 log(f_1)]$. For instance, if there is a dataset with 95% patients and 5% controls, then the baseline error rate is 5% and the baseline AMLP is -[0.05log(0.05)+0.95log(0.95)]. If the test error rate of a predictive model is 6%, it looks like very good with only 0.06 error rate. However, even if we use the random guess without any predictors, we can obtain the baseline error rate, 5%. Actually it's a very bad predictive model. In brief, to explore the level of predictivity of a model, we should compare error rate and AMLP with baseline values. Thus, we define the relative error rate as:

$$R^2_{ER} = \frac{ER_{(0)} - ER}{ER_{(0)}} \tag{2.21}$$

and the relative AMLP as:

$$R^2_{AMLP} = \frac{AMLP_{(0)} - AMLP}{AMLP_{(0)}} \tag{2.22}$$

2.3.2 AUC

The predicted labels of all samples are either positive (P) or negative (N). For a sample whose predicted label is P, if its actual label is also P, then it's called a true positive (TP); if the actual label is N, then it's a false positive (FP). Similarly, a true negative (TN) means both the predicted label and the actual label are N, and false negative (FN) means the predicted

label is N, whereas its actual label is P. We can summary these notations by building a confusion matrix:

Table 2.2: Confusion Matrix

		Predicted Labels	
		Class 1	Class0
Actual Labels	Class 1	TP	FN
	Class 0	FP	TN

Receiver operating characteristic curve (ROC) is the cut-off between sensitivity and specificity. They are defined as:

$$Sensitivity = \frac{the\ number\ of\ TP}{the\ number\ of\ TP + the\ number\ of\ FN} \tag{2.23}$$

$$Specificity = \frac{the\ number\ of\ TN}{the\ number\ of\ TN + the\ number\ of\ FP} \tag{2.24}$$

1-Specificity is called false positive rate. ROC space is defined by 1-Specificity and Sensitivity as x axis and y axis, respectively, which draws relative trade-offs between true positives and false positives. The predictive model provides predicted probabilities of being positive for all samples. Suppose c is the threshold, and $\hat{P}_i$ is the predicted probability of a certain sample being positive. If $\hat{P}_i \geq c$, then the predicted label of this sample is P; if $\hat{P}_i < c$ then its predicted label is N. By iteratively using each predicted probability as the threshold, we calculate corresponding sensitivity and specificity. Corresponding to a particular threshold, each point in ROC space is located by sensitivity as y-coordinate and 1-specificity as x-coordinate.

AUC is the area under the ROC curve. The baseline AUC is 0.5, from a random guess. A prediction having a AUC closer to 1, is considered more perfect. The theoretical definition of AUC is the probability that a randomly selected actual positive has a higher test result than a randomly selected actual positive [26]. AUC expresses the degree of separability, indicating how much a model is able to distinguish two classes. Higher AUC means a stronger ability to predict labels correctly.

2.3.3 AUPR

Precision recall curve (PRC) is a cut-off between Precision and Recall. They are defined as:

$$Precision = \frac{the\ number\ of\ TP}{the\ number\ of\ TP + the\ number\ of\ FP} \tag{2.25}$$

$$Recall = \frac{the\ number\ of\ TP}{the\ number\ of\ TP + the\ number\ of\ FN} \tag{2.26}$$

Recall is another name of sensitivity. PRC space is defined by Recall and Precision as x axis and y axis, respectively. Similar with ROC, each point in PRC is positioned by recall as x coordinate and precision as y coordinate. Precision and recall for each point are obtained by using corresponding predicted probability as the threshold. The baseline of PRC is a horizontal line at y=f_1. Here, f_1 is the frequency of positive observations in our dataset. This line divides the precision-recall space into two parts. Curves in the area above the line denote relatively good predictive performance.

AUPR is the area under the precision-recall curve. The best is 1.0. If a predictive model has a perfect AUPR, it means this model can find all positive samples without incorrectly classifying any negative samples to be positive. If we classify all points as P, we will get a perfect recall but a bad precision, and we will achieve a perfect precision but a bad recall, by classifying all points as N.Therefore, we should consider recall and precision together to evaluate a model.

Chapter 3

Dataset

The raw dataset is consist of 16s rRNA sequencing data from a case-control study of 327 participants. The full name of 16s rRNA is 16S ribosomal RNA, which is the component of the small subunit of the ribosome in prokaryotic. The gene coding for 16s rRNA is referred as 16S rRNA gene. 16S rRNA gene has 9 variable regions (V1–V9). Usually, we do amplicon analysis and sequencing on v3-v4 and v6. Sequences of 16S rRNA gene are widely used in bacterial phylogeny [27] and classification of bacteria. Reasons for its extensive usage are that it's present in most bacteria and microbes [28], that the function of the 16S rRNA gene has proper changes over time [28], and that the 16S rRNA gene is large enough for general information [29] [30]. Sequences and metadata are in the European Bioinformatics Institute European Nucleotide Archive, accession number ERP016332.

Quantitative Insights Into Microbial Ecology-version 1 (QIIME1) is a tool for pre-processing our 16s RNA raw sequencing dataset downloaded from EBI website. QIIME1 is a bioinformatics pipeline based Linux operating system, specially designed for conducting microbiome analysis on 16s rRNA sequencing data. QIIME1 can perform a series of analyses on raw sequencing data, including demultiplexing, quality filtering, OTU picking, taxonomic assignment, phylogenetic reconstruction, diversity analysis and visualizations. These steps can be conducted by direct commands and scripts [31]. In this thesis, operational taxonomic units (OTUs) are picked based on the closed reference option, with SortMeRNA [32] against the Green genes 16S rRNA gene sequence database released on August 2013 [33]. We set 94% as the threshold of similarity. OTUs with 94% similarity are in genus level.

From EBI website, we can download 376 samples, of which 28 are blank samples with scientific name "metagenome", and 348 are with scientific name "human gut metagenome". We just need to pay attention to those 348 samples. During the process of picking OTU, 21

samples are excluded from these 348 samples, and then 327 samples are kept. In these 21 exclusions, 3 samples are deleted automatically by QIIME in pick-OTU step. The output of "pick OTU" step just don't keep these 3 samples. 14 samples are deleted because of their insufficient total reads. A sample whose total reads are less than 5000 is treated as a non-informative enough. 5 samples should be deleted because 2 of them are with "bad metadata" and 3 of them are "duplicate samples". In addition, The group of 14 samples with total reads less than 5000 has a overlap with the group of 2 samples with bad metadata. Then, we exclude 21 samples (3+14+5-1=21) from 348 ones. Finally 327 samples are kept.

Outputs of QIIME1 contain 382 different OTUs (genera) for 327 samples.

Chapter 4

Results

4.1 Evaluation of Predictive Models

4.1.1 Choice of the Optimal Values of Tuning Parameters

This chapter demonstrates the investigation on fitting regularized SVM and logistic regression to dataset mentioned in section 3. Our models include SVM with L1 penalty, SVM with elastic net penalty, logistic regression with L1 penalty, and logistic regression with elastic net penalty. For these four models, LOOCV is applied to find the optimal values of tuning parameters.

We fit models sing R packages glmnet [23] and sparseSVM [34]. Before fitting models, the OTU dataset is transformed by formula $\widetilde{X}_i^{(j)} = log(X_i^{(j)} + 1) - log(\sum_{j=1}^{p} X_i^{(j)} + p)$. After transformation, we get input dataset. Functions in these two pakages can do standardization by default. There are cv.glmnet() and cv.sparsesvm() functions. Depending on our input dataset, these two functions generate a series of λ for SVM and logistic regression, with L1 and Elastic net regularizations. Then, we implement LOOCV to select the optimal one of these λ, using error rate as the criterion.

The total number of observations is $n = 327$. With each λ, we select 326 ones as train, to construct models. The one left is regarded as test set. After 327 iterations, we get test result of each observation. Comparing with true labels of 327 observations, evaluation metrics can be calculated, including error rate, AMLP AUC, AUPR. As we know, we need predicted probabilities to generate AMLP, ROC curve and PR curve. However, the predicted results of the SVM model are only distances to hyperplane and predicted labels, without probabilities. Thus, to get probabilistic results from SVM, we fit logistic transformation, using distance as

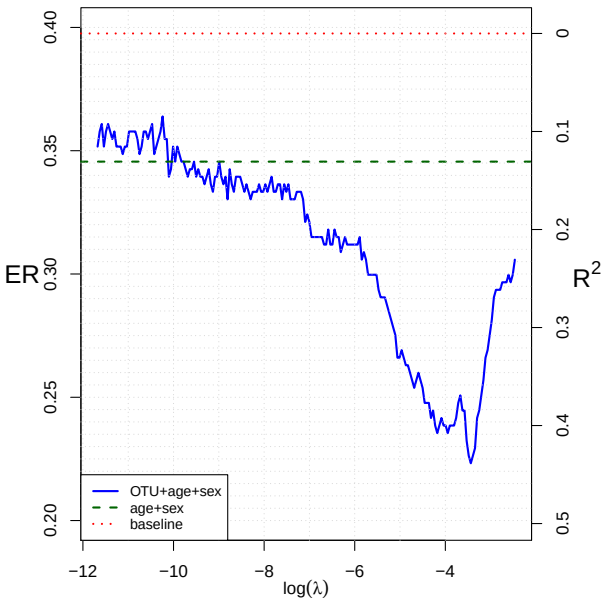
the predictor and true label as response variable. λ associated with smallest test error rate are is selected as the optimal λ of corresponding model. Consequently, we further explore models with these optimal λ.

Blue lines in figure 4.1 display test error rates of regularized logistic regression and SVM, with different values of tuning parameter λ. According to smallest error rate, we choose optimal λ for each model. The smallest error rate of logistic regression with L1 penalty, logistic regression with elastic net penalty, SVM with L1 penalty and SVM with elastic net penalty are 0.223, 0.238, 0.278, 0.275, respectively.

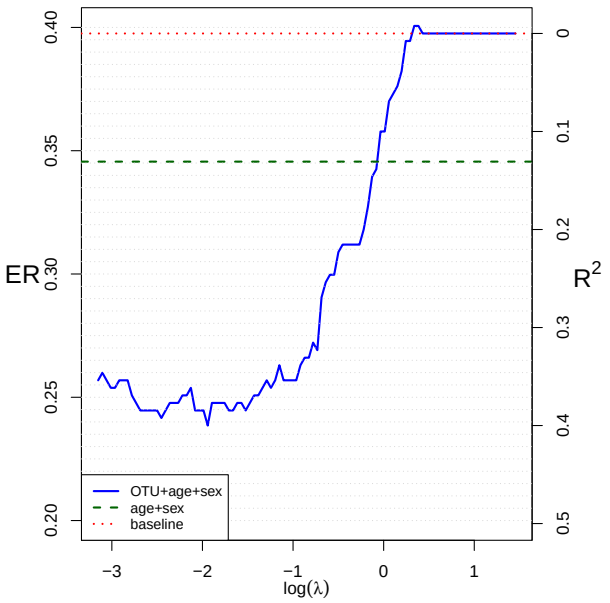
Blue lines in figure 4.2 and figure 4.3 show AMLP, AUC and AUPR of regularized logistic regression and SVM, with different values of tuning parameter λ. Under the optimal λ, the optimal AMLP of logistic regression with L1 penalty, logistic regression with elastic net penalty, SVM with L1 penalty and SVM with elastic net penalty are 0.515, 0.506, 0.550, 0.562, respectively. The optimal AUC of logistic regression with L1 penalty, logistic regression with elastic net penalty, SVM with L1 penalty and SVM with elastic net penalty are 0.826, 0.820, 0.777, 0.760, respectively; the optimal AUPR are 0.879, 0.879, 0.832, 0.831, respectively.

These figures provide not only the choice of optimal λ, but also the degree of difficulty in predicting the PD. In figure 4.1 and 4.2, we use red lines to denote the baseline error rate and AMLP. As it's mentioned in section 2.3 , R^2 is the percentage of the reduction of error rate or AMLP from the null model based on the frequency of observed y_i without any predictors. We show R^2 on the right vertical axes.

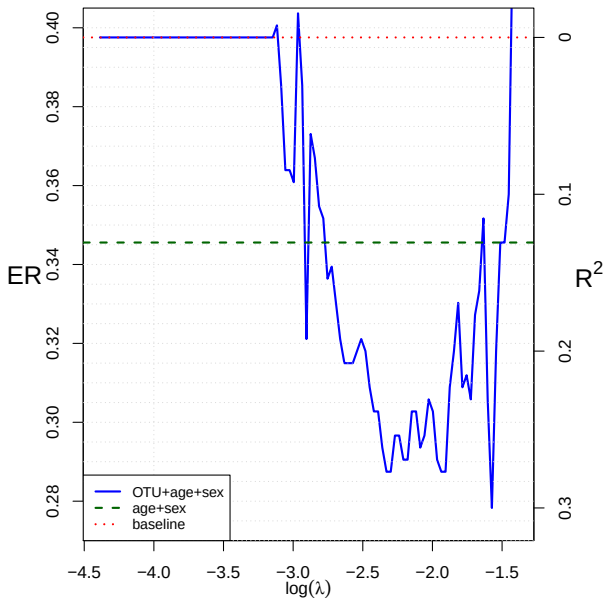
R^2 is a comparison of predictive models versus a random guess without variables, so, R^2 can indicate the predictivity of selected variables for PD, showing how much PD is related to these variables. Under the optimal λ, the R^2 of error rate for logistic regression with L1 penalty, logistic regresion with elastic net penalty, SVM with L1 penalty and SVM with elastic net penalty are 43.9%, 40.1%, 30.1%, 30.8%, respectively; the R^2 of AMLP are 25.7%, 27.0%, 20.7%, 18.9%, respectively. We can see R^2 of regularized logistics regression models are larger than those of regularized SVM, which means the predictivity of regularized logistics regression models with selected OTUs is stronger than that of regularized SVM. In our study, the effect of R^2 seems to be not obvious enough because we only use one dataset. Particularly, for those studies with different datasets, R^2 is much more useful because different datasets



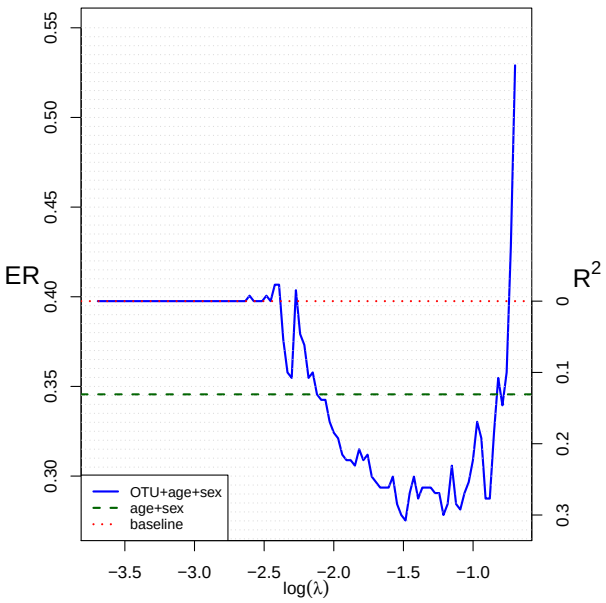
(a) Error Rate of of Logistic Regression with L1



(b) Error Rate of of Logistic Regression with Elastic-net

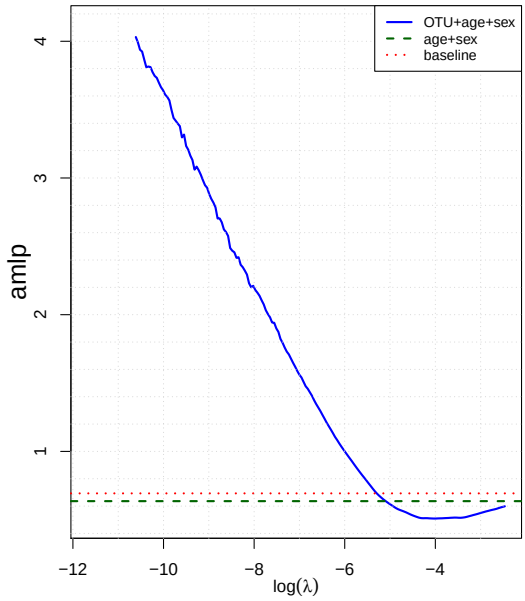


(c) Error Rate of of SVM with L1

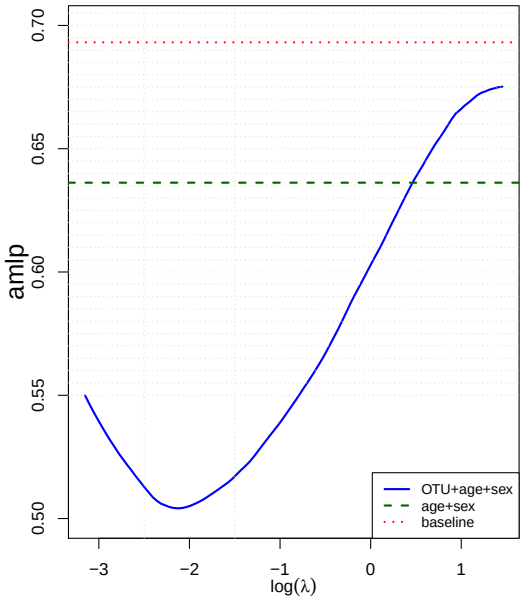


(d) Error Rates of of SVM with Elastic-net

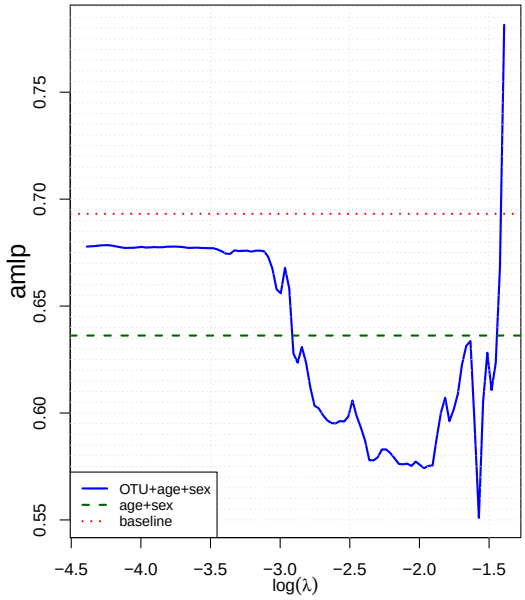
Figure 4.1: Error Rate of Regularized Logistic Regression and Regularized SVM, with Different Values λ.



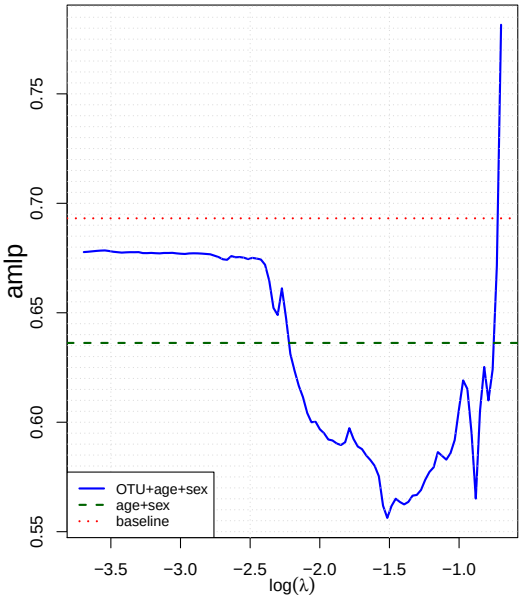
(a) AMLP of Logistic Regression with L1



(b) AMLP of Logistic Regression with Elastic-net

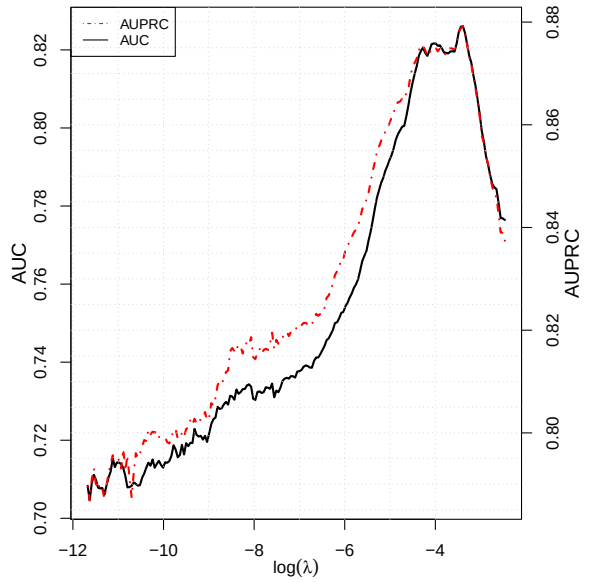


(c) AMLP of SVM with L1

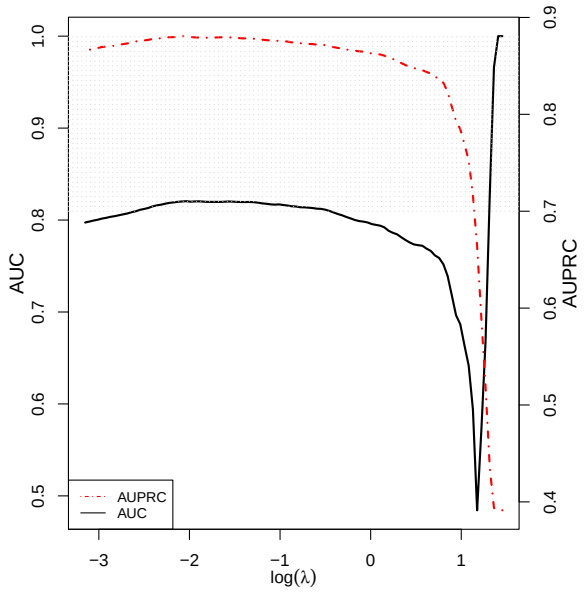


(d) AMLP of SVM with Elastic-net

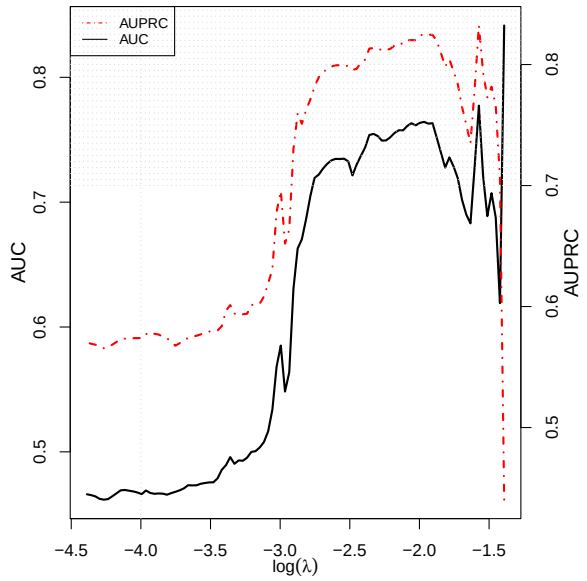
Figure 4.2: AMLP of Regularized Logistic Regression and Regularized SVM, with Different Values λ.



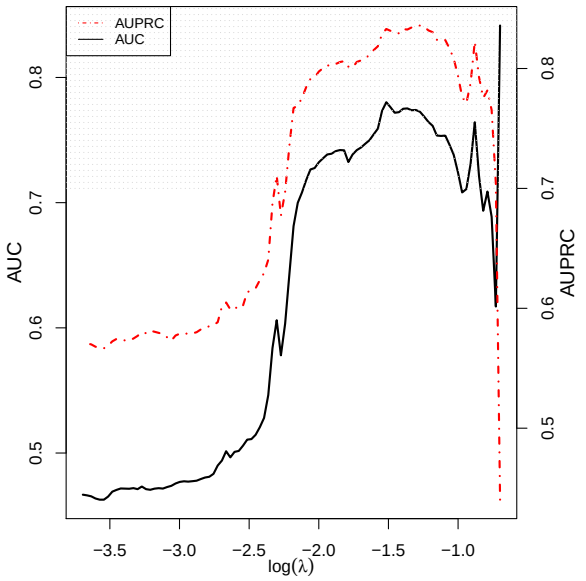
(a) AUC and AUPR of Logistic Ression with L1



(b) AUC and AUPR of Logistic Ression with Elastic-net



(c) AUC and AUPR of SVM with L1



(d) AUC and AUPR of SVM with Elastic-net

Figure 4.3: AUC and AUPR of Regularized Logistic Regression and Regularized SVM, with Respective λ.

have different baselines of error rate and AMLP. Hence, we cannot directly compare error rates and AMLPs from different dataset, but the comparison of R^2, a relative predictive metric, is reasonable.

Relative comparison with baseline is also shown in figure 4.2 and 4.3. In these figures, the red line denotes the baseline, and the green line is result of those predictive models only with age and sex as predictors, without OTUs. They are parallel to horizontal line and not affected by the value of λ because their corresponding models don't contain any regularizations. In these figurs, we can see the lowest point of those blue lines are much lower than the red line, and red line is higher than green line. Thus, with each optimal λ, OTU features are very predictive, making error rate and amlp much less than baselines. Prediction only based on age and sex is reliable than random guess, which indicates that age and sex are also factors related to PD. This is consistent with the basic factors of PD, discussed in section 1.1

4.1.2 Performance of Models with Respective Optimal Tuning Parameters

In this section, we further investigate optimal models with their own ideal λ. Table 4.1 summarizes specific values of evaluation metrics for each model with its optimal λ. Table 4.1 is a straightforward demonstration for predictive performances of these models. From the table, we can see that logistic regression with L1 penalty performs best. Although the AMLP of logistic model with L1 is a little larger than that of logistic regression with elastic-net, its other metrics are still the best of all. If we only focus on regularizations, we can see that the performances of two penalties are very close, hard to be distinguished; if we pay attention to models, it's clear that logistic regression performs better than SVM.

Table 4.1: Evaluation of Predictive Models with Respective Optimal λ

Metric / Model	ER	AUC	AUPR	AMLP
LR+L1	0.223	0.826	0.879	0.515
LR+Elastic-net	0.238	0.820	0.879	0.506
SVM+L1	0.278	0.777	0.832	0.550
SVM+Elastic-net	0.275	0.760	0.831	0.562

Nevertheless, there is no clear answer for the question whether logistic regression does better than linear SVM. Actually it depends on what dataset we analyze. In our study, we use linear SVM without kernel tricks, which performs similar with logistic regression, but there are still differences between them. The loss function of SVM is hinge loss whereas logistic regression has a cross entropy loss function. The corresponding convex optimization of SVM involves in Lagrange Duality and Karush Kuhn Tucker (KKT) conditions, resulting in only points within the margin (support vectors) being decisive for optimal solution. In SVM, only points within margin effect model, so alterations of other points cannot change solutions of the model. Therefore, SVM has a stronger generalization ability and then it is preferred for high-dimensional dataset with many noises.

Unlike SVM, logistic regression considers all training data. Thus, logistic regression is more sensitive to outliers and alterations. However, it's perfect for a dataset that doesn't have so many dimensions and noises. Sometimes SVM may ignore some significant information by only relying on points near the margin, especially when the dataset is not so large. Our dataset has 327 observations and 382 features. Compared with datasets having thousands of dimensions and observations, actually our dataset is not a real high-dimensional dataset. Hence, we'd better let all samples and features are fully used, to make our models stable and reliable enough. Under this circumstance, using SVM only relying on support vectors may miss some significant information.

Figure 4.4 displays ROC and PRC for models. ROC reveals the rate of correctly classified positive observations (true positive rate) versus the rate of incorrectly classified negative

observations (false positive rate), under each threshold. The diagonal line represents that true positive rate is always equal to false positive rate. If we don't use any variables as predictors, and just do random guess, then true positive rate will be always equal to false positive rate. The diagonal line denotes the baseline situation.

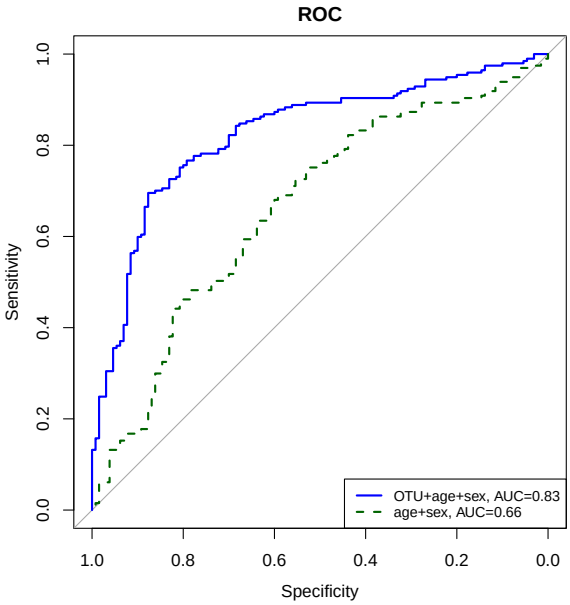
Intuitively, we expect that true positive rate can be as high as possible whereas false positive rate can be as low as possible. Then, in ROC space, the closer a curve is to the left-top corner, the better the predictive performance is. Curves close to left-top corner and enclose the diagonal line are indicators of favorable predictions, while curves close to diagonal line imply worse predictive performance.

In figures 4.4a, 4.4b, 4.4c, 4.4a, blue lines are ROC curves of regularized models with their optimal λ; green lines are ROC curves of predictive models only with age and sex as predictors; diagonal lines are baselines of random guess. We can see all blue lines are close to left-top corner, enclosing diagonal lines, and green lines are between blue lines and diagonal lines. These relative positions suggest that models with OTU, age and sex are most predictive. Besides, compared with models based on OTUs, age and sex, those models only with age and sex are a little worse, but they are still better than baseline situations. Logistic regression with L1 is the best with largest AUC value. ROC curves of logistic regression models are a little more left-top oriented than those of SVM, and AUC values of logistic regression models are also higher than those of SVM. Results of ROC curves are consistent with indications of figure 4.1 and 4.2.

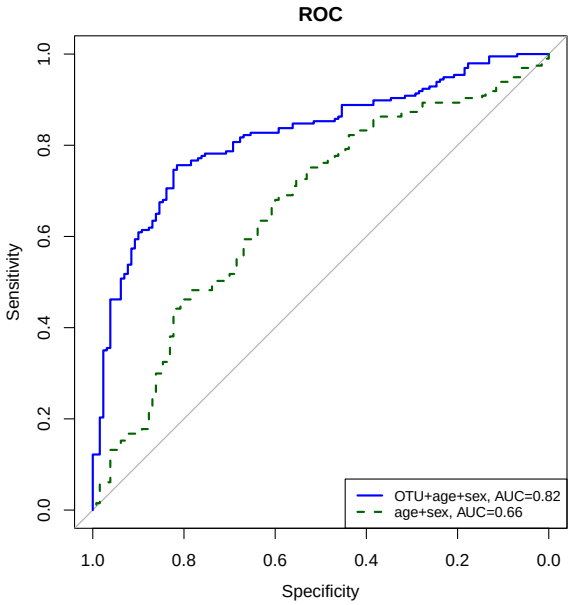
Figure 4.4a, 4.4b, 4.4c, 4.4d are PR curves for models. PR curve depicts precision versus recall under each threshold. The baseline of PR curve depends on the distribution of data. In PRC space, baseline is a horizontal line whose y-coordinate is the frequency of positive observations in a dataset. If we don't use any variables as predictors, and just do random guess, then the precision will be always equal to the ratio of positive samples to all samples. The higher precision and recall are, the more advantageous a model is. In PR space, the closer a curve is to the right-top corner, the better the predictive performance is, while curves closer to baseline imply relatively worse predictive performances.

Blue lines 4.4a, 4.4b, 4.4c, 4.4d are PR curves of regularized models with the optimal λ; green lines are PRC curves of predictive models with age and sex as features; red lines are

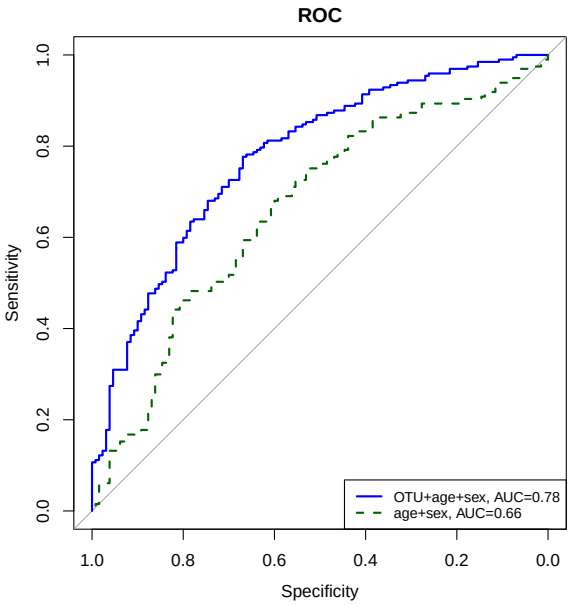
baselines of random guess. All blue lines are close to the right-top corner, and they are located far above baselines. Green lines are also above red lines obviously, but still underneath blue lines. These also suggest that age, sex are factors of PD, and that models with age, sex and OTU as predictors would be much more predictive. When we compare logistic regression with SVM, though their AUPR values are too close to distinguish, the shapes of their blue lines are obvious different. PR curves of logistic regression are more closer to the right-top corner than those of SVM.



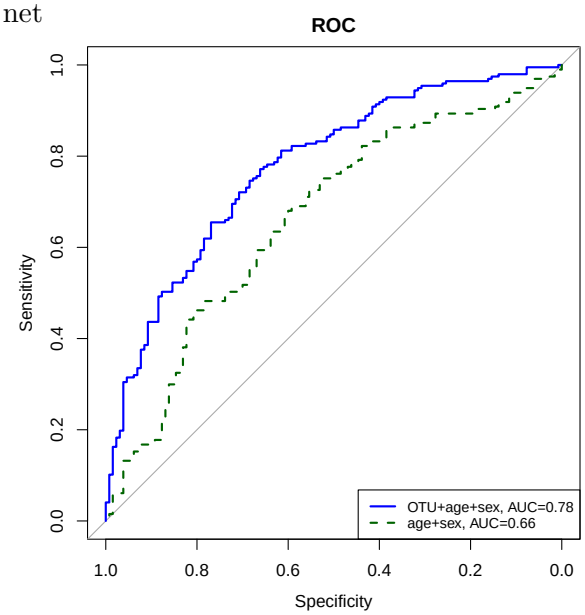
(a) ROC of Logistic Regression with L1



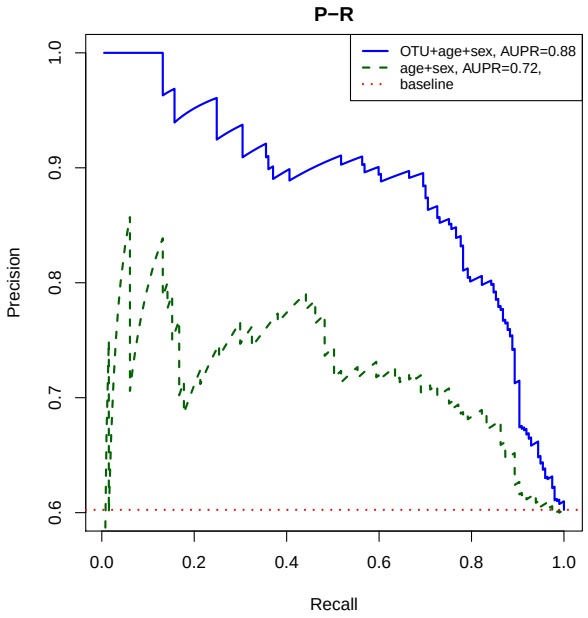
(b) ROC of Logistic Regression with Elastic-net



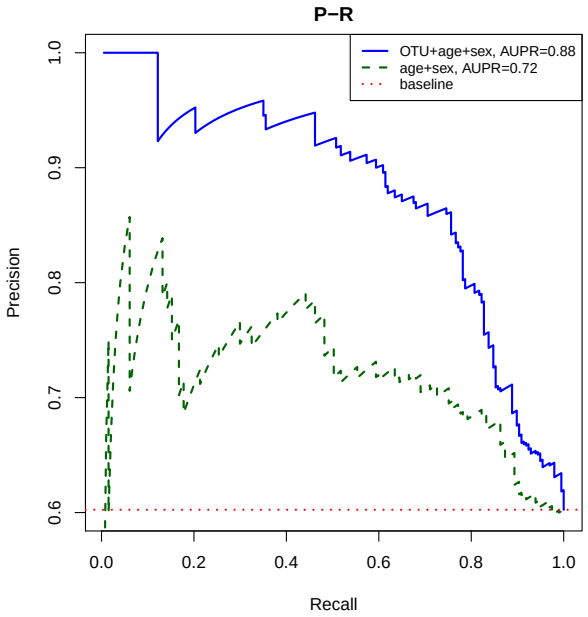
(c) ROC of SVM with L1



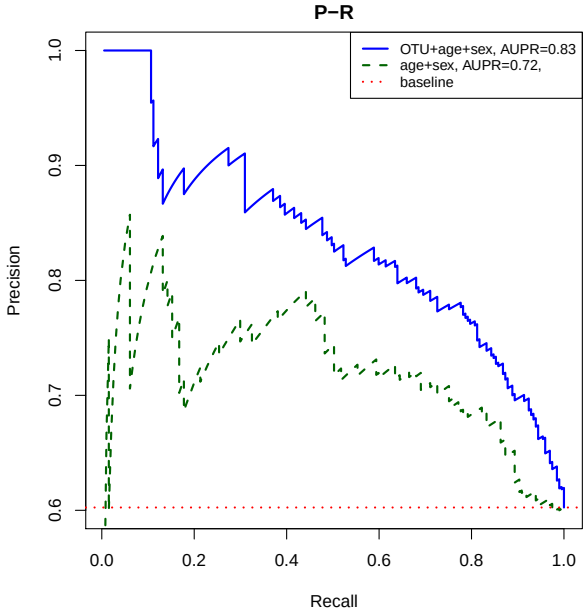
(d) ROC of SVM with Elastic-net



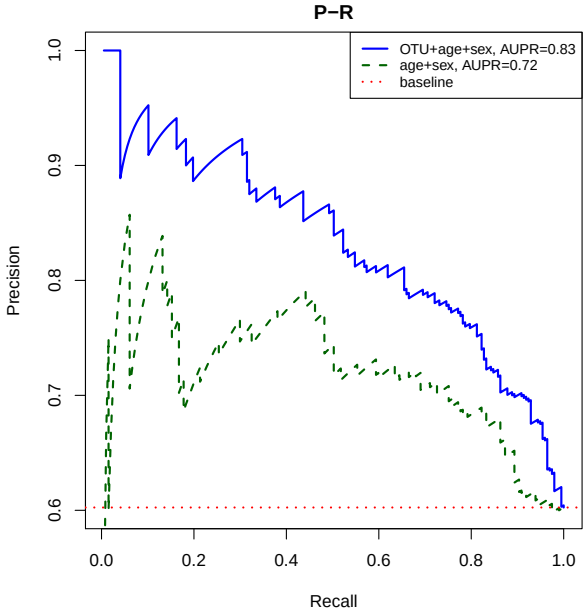
(a) PRC of Logistic Regression with L1



(b) PRC of Logistic Regression with Elastic-net

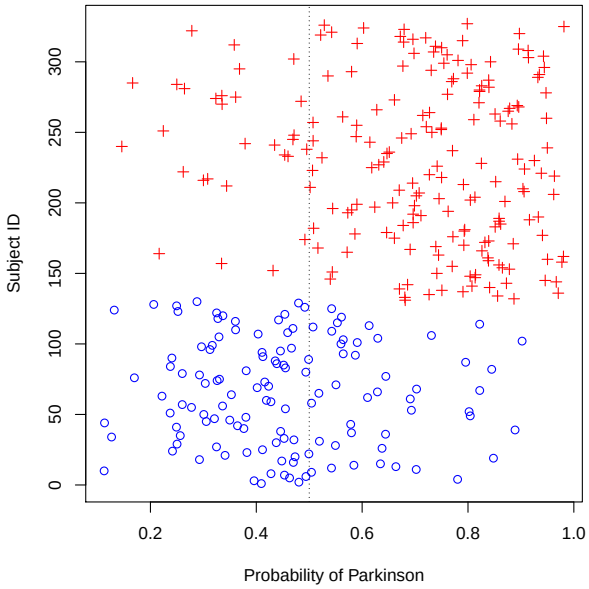


(c) PRC of SVM with L1

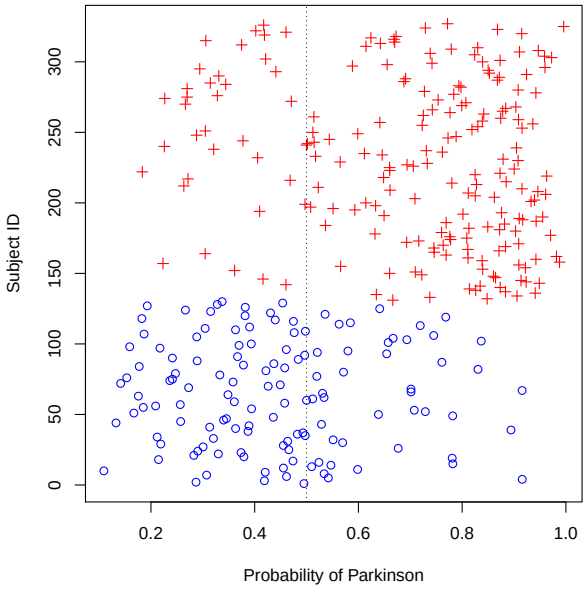


(d) PRC of SVM with Elastic-net

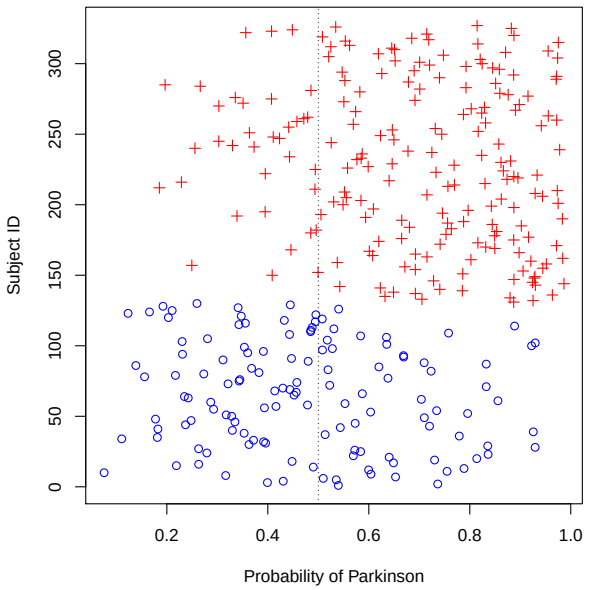
Figure 4.4: ROC and PRC of Regularized Logistic Regression and Regularized SVM, with Optimal λ.



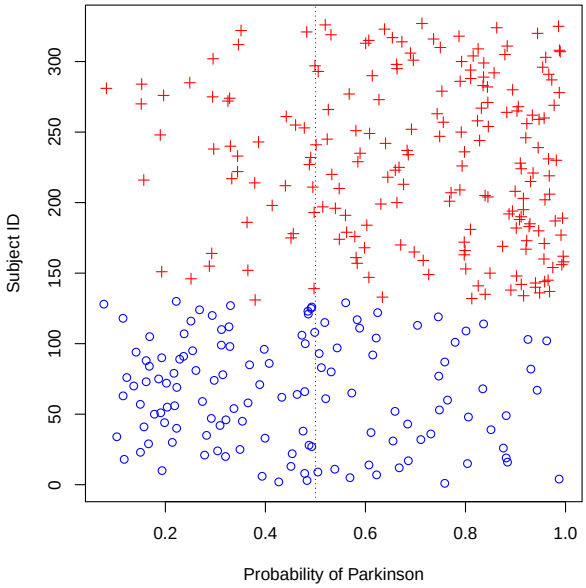
(a) Predicted Probabilities of Logistic Regression with L1



(b) Predicted Probabilities of Logistic Regression with Elastic-net



(c) Predicted Probabilities of SVM with L1



(d) Predicted Probabilities of SVM with Elastic-net

Figure 4.5: Probability of Parkinson Plots of Regularized Logistic Regression and Regularized SVM, with Optimal λ.

Figure 4.5 can be treated as a visual illustration of the confusion matrix for these models

with optimal λ. Red crosses and blue circles are positive observations and negative observations, respectively. X-coordinate of a point is the predicted probability of the corresponding observation being a PD patient. Y-coordinates just mean the index of points. Here, the threshold is 0.5. If an observation's predicted probability is no less than 0.5, then the predicted label is positive. Otherwise, the predicted label is negative. The middle vertical line at x=0.5 is the probabilistic decision line. From these figures, we can conclude that logistic models have higher sensitivity and specificity than SVM models, because logistic models have more red crosses located in right of threshold line, and more blue circles in left. Red crosses in left side are positive observations that are incorrectly classified as negative, and blue circles in right are negative samples that are falsely classified as positive. Logistic regression models have a less amount of red crosses in left and blue circles in right. Thus, error rates of logistic models are smaller than those of SVM models.

4.2 Selection of Significant OTUs

4.2.1 The process of Selection

In addition to the predictive performances of models, the selection of significant OTUs is of great importance. As it's mentioned in section 2, L1 penalty does feature selection by enabling coefficients associated with non-significant variables to be exactly zero. Elastic-net does grouping effect on the basis of L1 penalty. Next, we pay attention to the coefficients of each variable. Intuitively, the larger absolute value of a coefficient is, the more significant the associated variable is. One thing should be noted is that we conduct ECV to do feature selection. Unlike ICV, ECV makes coefficients of variables differ in each fold. In our dataset, 327 observations mean that we have 327 folds in LOOCV, and then, each variable has 327 coefficients. For a certain variable, absolute value of coefficient in a fold may be zero, while in another fold, it might be non-zero. Thus, if the median of absolute values of coefficients from all folds is zero, then the associated variable is eliminated; if the median of absolute values of coefficients is non-zero, the associated variable is retained. In each regularized model with optimal λ, OTUs retained are different. In logistic models, with L1 penalty, the number of

remaining OTUs is 25, and with elastic-net penalty, the number is 196. SVM with L1 keeps only 2 OTUs remained, while SVM with elastic-net has 27 OTUs remained.

The number of remaining OTUs varies so much. As we discuss in section 4.1.2, the solution of SVM only relies on a subset of training data, named as support vectors. The sparseness of solution is obtained by solving out optimization problems conditional on KKT. Therefore, the solution of SVM tends to be sparser than that of logistic models. Besides, in section 2.1.5 we have mentioned that if there is a group of highly related variables, L1 penalty tends to choose one of them and ignore others, while elastic-net often considers all of them in the group. Thus, under the same model, the solution of L1 penalty will be sparser than that of elastic-net penalty.

We rank the top 20 remaining OTUs from most significant to less significant, by ranking the median. Boxplots in figure 4.6a, 4.6c, 4.7a, and 4.7c show distributions of absolute values of coefficients associated with top 20 variables. Boxplots in figure 4.6b, 4.6d, 4.7b, and 4.7d show the distributions of standardized log relative abundances of selected OTUs. Here, log relative abundances are from the formula 2.18. We can see that for most selected OTUs, the size and length of blue boxes are different from those of red boxes, which means distribution of relative abundances differs obviously between controls and PD cases. For some selected OTUs, shape of blue box is similar as red box, but they have obviously divergent outliers. These outliers are very likely to make some differences for the prediction on PD.

4.2.2 Consistency with Other Studies

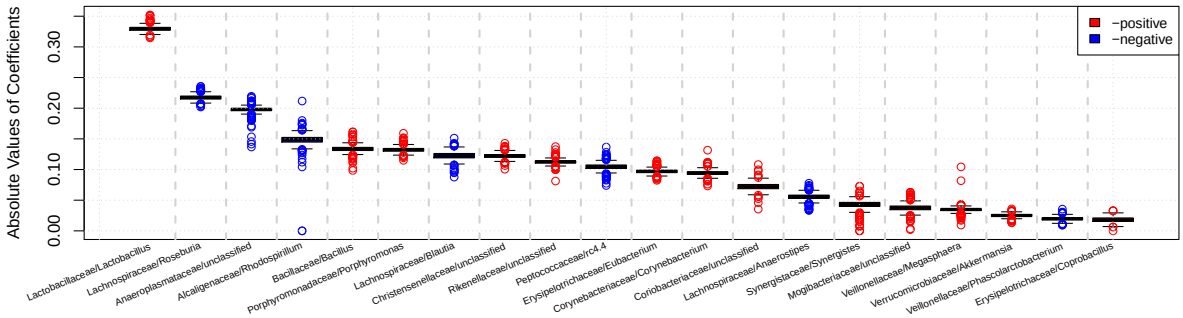
In our study, the sparsest model is SVM with L1, making only 2 OTUs remained, Roseburia and Lactobacillus. These two genera are also at the top positions in rankings of selected OTUs by other models. Lactobacillus is ranked at 1st by logistic regression with L1 penalty, at 6th by logistic regression with elastic-net penalty, and at 1st by SVM with elastic-net penalty. As for Roseburia, it's ranked at 2nd by logistic regression with L1, at 5th by SVM with elastic-net, and at 27th by logistic regression with elastic-net. There are some evidences indicating that in PD patients, the level of relative abundances of Roseburia in gut is decreased [35][13], while some studies conclude that Lactobacillus is more abundant in PD patients [36] [15].

[13] and [15] also imply that the levels of relative abundances of Bluatia, Akkermansia and Bifidobacterium are significantly changed in the course of PD. PD patients tend to have a higher level of Akkermansia and Bifidobacterium, but a lower level of Bluatia. In figures 4.6b, 4.6d and 4.7d, we can see that Blautia is ranked at 7th by logistic regression with L1, at 19th by logistic regression with elastic-net, and at 16th by SVM with elastic-net. Shapes of boxplots show that Blautia of PD patients is relatively more abundant than that of healthy cases. For Bifidobacterium and Akkermansia, we can see that only SVM with elastic-net ranks Bifidobacterium in top 20, at 10th, and only logistic regression with L1 ranks Akkermansia in the top 20, at 18th. However, we have already mentioned in section 4.2.1 that logistic regression with elastic-net has 196 OTUs retained, that logistic regression with L1 has 25 OTUs retained, and that SVM with elastic-net has 27 OTUs retained. We only display the boxplots of top 20. Some significant OTUs are not displayed in figures, but it doesn't mean that they are not selected. Bifidobacterium is ranked at 24th and 98th, by logistic regression with L1 and elastic-net, respectively. Akkermansia is ranked at 23th and 127th by SVM with elastic-net and logistic regression with elastic-net, respectively. They are all in the list of retainment from each model.

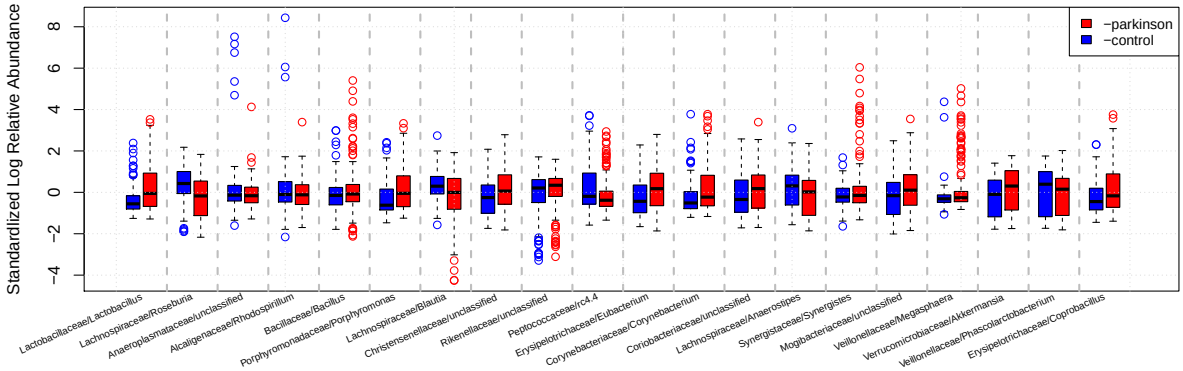
Additionally, a study [37] that uses the same dataset as ours, also explores the relation between PD and gut microbiome. Though methodologies are different, our study still has some consistencies with theirs. [37] summarizes 8 genera that are highly related with PD. Lactobacillus, Blautia, Roseburia, Bifidobacterium, Akkermansia are all in the list of these 8 genera. One of these 8 genera is a genus without a specific name, under the family Christensenellaceae. In our results, all the lists include this genus called as Christensenellaceae/unclassified. It's ranked at 8th and 12th by logistic regression with L1 and SVM with elastic-net, respectively. We can see it from our figures 4.6b and 4.7d. It's ranked at 47th by logistic regression with elastic-net, so we cannot see it in figure 4.6d, but it's still in the list of retainment from logistic model with elastic-net.

All in all, the results of our study do have some consistencies with conclusions of other published studies.

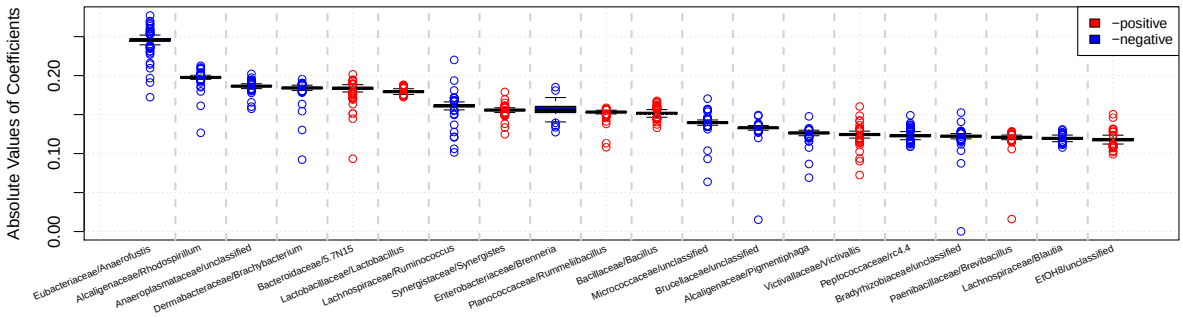
Figure 4.6: Selected OTUs of Regularized Logistic Regression



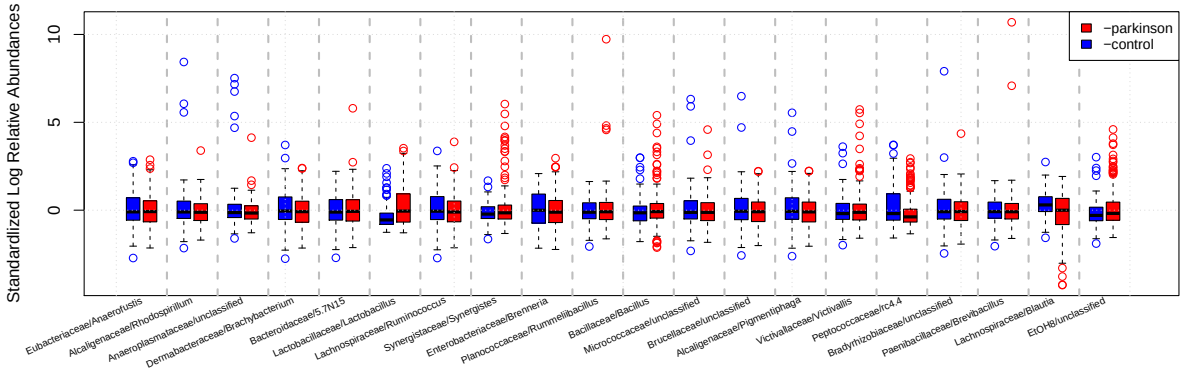
(a) Absolute Coefficients of Selected OTUs by Logistic Regression with L1



(b) Relative Abundances of Selected OTUs by Logistic Regression with L1

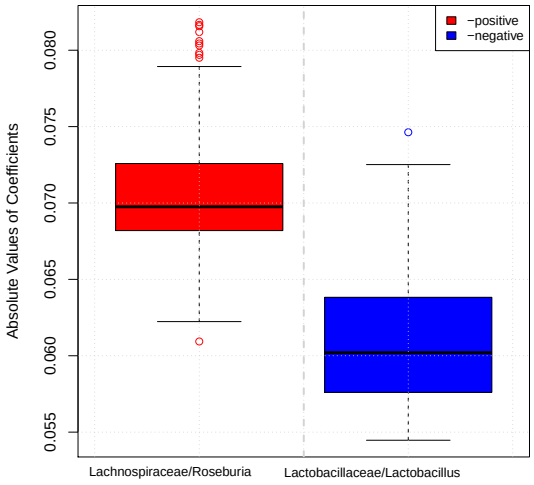


(c) Absolute Coefficients of Selected OTUs by Logistic Regression with Elastic-net

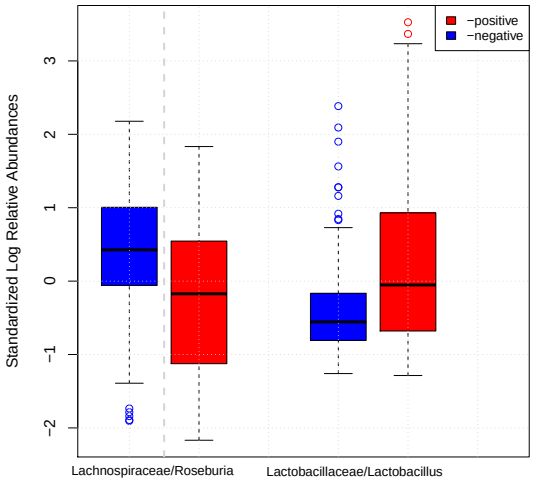


(d) Relative Abundances of Selected OTUs by Logistic Regression with Elastic-net

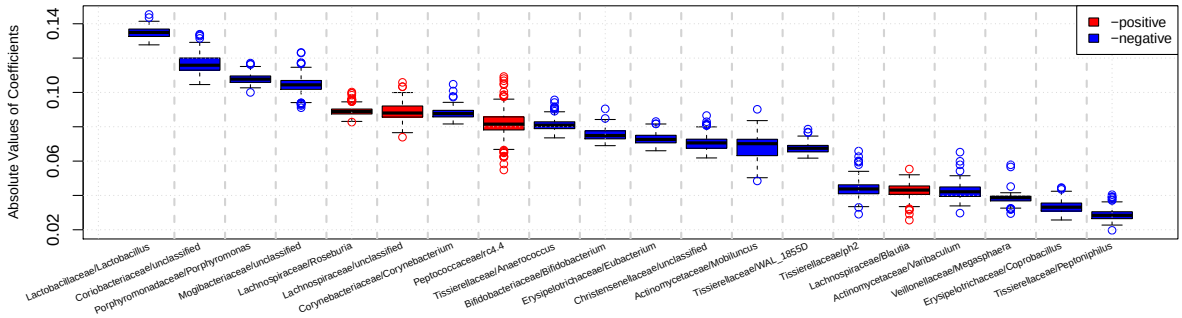
Figure 4.7: Selected OTUs of Regularized SVM



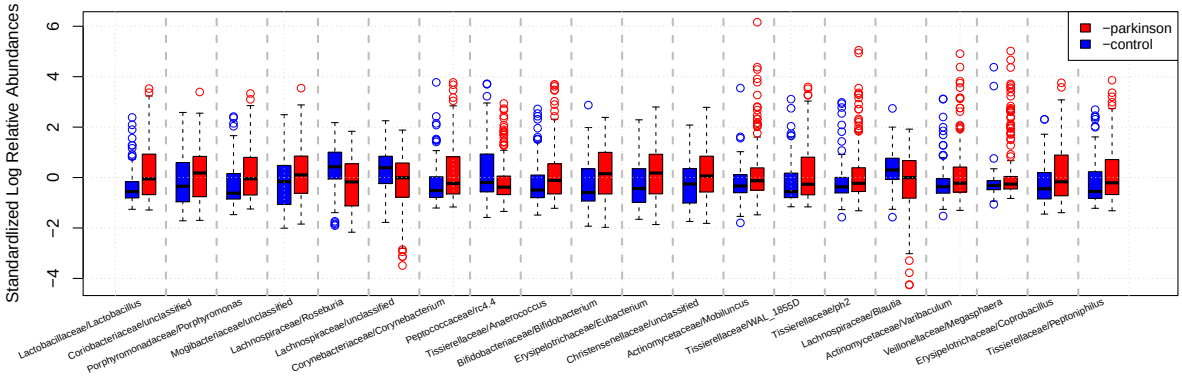
(a) Absolute Coefficients of Selected OTUs by SVM with L1



(b) Relative Abundances of Selected OTUs by SVM with L1



(c) Absolute Coefficients of Selected OTUs by SVM with Elastic-net



(d) Relative Abundances of Selected OTUs by SVM with Elastic-net

Chapter 5

Conclusion and Feature Work

In this thesis, we explore the relationship of PD and gut microbiota by fitting SVM and logistic regression to predict PD, with OTUs as predictors. To select significant OTUs, regularizations are applied into predictive models, including L1 regularization and elastic-net regularization. LOOCV is conducted to choose the optimal values of tuning parameter λ in regularizations. We also discuss the difference between ICV and ECV, and then point out ICV is not reliable enough, causing underestimated error rate. Although the error rate of ECV looks like higher than that of ICV, ECV's result is much more reliable. After choice of the optimal λ for each model, we obtain that for logistic models regularized by L1 penalty and elastic-net penalty, error rates are 0.223 and 0.238, respectively; AUC are 0.826 and 0.820, respectively; AUPR are both 0.879; AMLP are 0.515 and 0.506, respectively. In results of SVM with L1 penalty and elastic-net penalty, error rates are 0.278 and 0.275; AUC are 0.777 and 0.760; AUPR are 0.832 and 0.831; AMLP are 0.550 and 0.562. Of course we should compare these evaluation values with baselines by looking into R^2. Under the corresponding optimal λ, the R^2 for error rate of logistic regression with L1, logistic regresion with elastic net, SVM with L1 and SVM with elastic net are 43.9%, 40.1%, 30.1%, 30.8%, respectively; R^2 for AMLP of them are 25.7%, 27.0%, 20.7%, 18.9%, respectively.

Then, we select significant OTUs related with PD, according to the median of absolute values of coefficients associated OTU variables. If absolute values of coefficients have a non-zero median, then the corresponding variable is retained.

After selection of significant OTUs, we compare our results with reported studies. Other studies have found that abundances of some gut genera in PD patients differ significantly between patients and controls. Selected genera in our study are in accordance with most of their results, including Lactobacillus, Blautia, Roseburia, Bifidobacterium, Akkermansia and

a genus without a specific name, under the family Christensenellaceae.

Although the consistency with other studies exists, we still fail to exactly figure out a group of specific OTUs that are highly related with PD. Our body is a complex system. Each OTU has complicated interactions with gut microbial environment in the progress of PD. It's hard to figure out all details and interactions. In those published studies, there are some other genera recognized as very significant, but in our results, they are not selected as significant variables. It means our models may have apparent drawbacks. Furthermore, correlation does not imply causation. We confirm that some patterns of gut microbiota are related with PD, but we still cannot know what causes changes of gut mocrobiota in the course of PD. There are many other models suitable for OTU data, such as zero-inflated regression models. We can further explore our dataset using other models in the future. Besides, some medical experiments can be also done by injecting related genera into animal's gut, and observing the changes.

References

[1] National Institute of Neurological Disorders and Stroke. Parkinson's disease information page. *https://www.ninds.nih.gov/Disorders/All-Disorders/Parkinsons-Disease-Information-Page#disorders-r1*, 2016.

[2] Sigurlaug Sveinbjornsdottir. The clinical symptoms of parkinson's disease. *Journal of Neurochemistry*, 139(1):318–324, 2016.

[3] Lorraine Kalia and Anthony Lang. Parkinson's disease. *The Lancet*, 386(9996):896–912, 2015.

[4] Maria Cersosimo, Gabriela Raina, et al. Gastrointestinal manifestations in parkinson's disease: prevalence and occurrence before motor symptoms. *Journal of Neurology*, 260(5):1332–13328, 2013.

[5] David Devos, Thibaud Lebouvier, et al. Colonic inflammation in parkinson's disease. *Neurobiology of Disease*, 50:42–48, 2013.

[6] Gwen Falony, Marie Joossens, et al. Population-level analysis of gut microbiome variation. *Science*, 352(6285):560–564, 2016.

[7] Alexandra Zhernakova and Alexander Kurilshikov. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, 2016.

[8] Ömrüm Aydin, Max Nieuwdorp, and Victor Gerdes. The gut microbiome as a target for the treatment of type 2 diabetes. *Current Diabetes Reports*, 18(8), 2018.

[9] Patrice Cani. Human gut microbiome: hopes, threats and promises. *Gut*, 67(9):1716—-1725, 2018.

[10] Andrew Shreiner, John Kao, and Vincent Young. The gut microbiome in health and in disease. *Current Opinion in Gastroenterology*, 31(1):69—-75, 2015.

[11] Maria Carmen Cénita, Vasiliki Matzaraki, et al. Rapidly expanding knowledge on the role of the gut microbiome in health and disease. *Neurobiology of Disease*, 1842(10):1981–1992, 2014.

[12] Filip Scheperjans, Velma Aho, et al. Gut microbiota are rlated to parkinson's disease and clinical phenotype. *Movement Disorders*, 30(3):350–358, 2015.

[13] Ali Keshavarzian, Stefan Green, et al. Colonic bacterial composition in parkinson's disease. *Movement Disorders*, 30(10):1351–1360, 2015.

[14] Marcus Unger, Jorg Spiegel, et al. Short chain fatty acids and gut microbiota differ between patients with parkinson's disease and age-matched controls. *Parkinsonism and Related Disorders*, 32:66–72, 2016.

[15] Vjacheslav Petrov, Irina Saltykova, et al. Analysis of gut microbiota in patients with parkinson's disease. *Bulletin of Experimental Biology and Medicine volume*, 162:734–737, 2017.

[16] Herbert Ross et al. Principles of numerical taxonomy. *Systematic Biology*, 13:106–108, 1964.

[17] Mary Ann Tolson Carl Boyd and Wayne Copes. Evaluating trauma care: The triss method. trauma score and the injury severity score. *The Journal of Trauma*, 27(4):370–378, 1987.

[18] Murat Kologlu, Doruk Elker, et al. Validation of mpi and pia ii in two different groups of patients with secondary peritonitis. *Hepato-Gastroenterology*, 48(37):147–151, 2001.

[19] Sebastiano Biondo, Emilio Ramos, et al. Prognostic factors for mortality in left colonic peritonitis: a new scoring system. *Journal of the American College of Surgeons*, 191(6):635–642, 2000.

[20] John Marshall, Deborah Cook, et al. Multiple organ dysfunction score: A reliable descriptor of a complex clinical outcome. *Critical Care Medicine*, 23(10):1638–1652, 1995.

[21] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *The Journal of the American Medical Association*, 270(24):2957–2963, 1993.

[22] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal, Series B (Methodological)*, 58(1):267—-288, 1996.

[23] Jerome Friedman, Rob Tibshirani, and Trevor Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1—-22, 2010.

[24] Li Wang, Ji Zhu, and Hui Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16:589–610, 2006.

[25] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22, 2010.

[26] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[27] Carl Woese and George Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *roceedings of the National Academy of Sciences of the United States of America*, 74(11):5088–5090, 1977.

[28] Pablo Yarza, Pelin Yilmaz, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. *Nature Reviews. Microbiology*, 12(9):635–645, 2014.

[29] Jean Baldus Patel. 16s rrna gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular Diagnosis*, 6(4):313–321, 2001.

[30] Michael Janda and Sharon Abbott. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9):2761–2764, 2007.

[31] Justin Kuczynski, Jesse Stombaugh, et al. Using qiime to analyze 16s rrna gene sequences from microbial communities. *Current Protocols in Bioinformatics*, page Doi:10.1002/0471250953.bi1007s36, 2011.

[32] Evguenia Kopylova, Laurent Noe, and Helene Touzet. Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217, 2012.

[33] Daniel McDonald, Morgan Price, et al. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *Multidisciplinary Journal of Microbial Ecology*, 6(3):610–618, 2012.

[34] Congrui Yi and Jian Huang. Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. *Journal of Computational and Graphical Statistics*, 26(3):547–557, 2017.

[35] Janis Bedarf, Falk Hildebrand, et al. Functional implications of microbial and viral gut metagenome changes in early stage l-dopa-naive parkinson's disease patients. *Genome Medicine*, 9(1), 2017.

[36] Satoru Hasegawa, Sae Goto, et al. Intestinal dysbiosis and lowered serum lipopolysaccharide-binding protein in parkinson's disease. *PLoS One*, 10(11):e0142164, 2015.

[37] Erin Hill-Burns, Justine Debelius, et al. Parkinson's disease and parkinson's disease medications have distinct signatures of the gut microbiome. *Movement Disorders*, 32(5):739–749, 2017.