Human gut microbiome for pediatric Autism Spectrum Disorder (ASD) Report

Gutsy Graduates: Yixiang Qu, Jialiu Xie, Jose S. Lopez, Yifan Dai, Jane She

2023-04-24

Introduction

Backgound & Motivation

Autism spectrum disorder (ASD) is characterized by behaviors in the areas of social communication and restricted repetitive sensory-motor behaviors. While ASD results from altered brain development and neural reorganization, there are no genetic biomarkers for diagnosis. Instead, diagnosis is characterized by a patient's behavior which are assessed on different scales by physicians through clinical observation as well as caregiver reports. However, diagnosis screening processes are often different for young children, older children, adolescents, and adults. The majority of children with ASD in North America and northern Europe will have already been diagnosed by early school age, but there are still those without a diagnosis (Lord et al. 2018). Validity of self-report and self-diagnosing instruments are also questionable.

In 2012, it was estimated by the WHO that the global prevalence of ASD was around 1% which was concluded based on reviewing epidemiological surveys worldwide (Elsabbagh et al. 2012). While it was once considered rare, current estimates suggest that these rates have risen in prevalence, and there is controversy as to what is causing these increased rates (Matson and Kozlowski 2011). These changes may possibly be attributed to new assessment instruments, improved awareness, and changes in diagnostic criteria.

Despite the increased awareness about the disorder, its underlying causes remain unclear. There have been genetic components linked to ASD, but determining the specific mutations associated with ASD is very difficult due to the high variation of underlying gene mutations (Sutcliffe 2008). Additionally, such genetic anomalies only account for a small fraction of cases, around 7% from current estimates (Landrigan 2010). However, this small percentage is also due to currently undiscovered genetic associations and is expected to increase with the advancement of genetic research. It has also been hypothesized that environmental exposures may also contribute to the causation of the disease.

One of the commonly seen comorbidities in patients with ASD is the occurrence of gastrointestinal (GI) related disturbances. Recently, links have been formed between the human gut microbiome and ASD

symptoms (Vuong and Hsiao 2017). Changes in the gut microbiome have been demonstrated to modulate behavior through the gut-microbiome-brain axis.

Microbiome data is known to be challenging to work with for several reasons. Microbiome data faces challenges in data quality, as different measurement techniques from experimental replicates may yield high variability in results. Additionally, data is high dimensional naturally inner-correlated, violating the independence assumption in many traditional analysis approaches. Zero-inflation is also present in the data due to zeroes from sampling, rounding, and normalization from preprocessing. Finally, microbiome data is always positive, so users need to be intentional in choosing a modeling strategy that allows for positive results.

To tackle the issue of high dimensionality, dimensionality reduction techniques have been applied using methods such as principal component analysis (PCA). More recently, preprocessing methods for microbiome datasets have been developed which borrow ideas from the field of natural language processing (NLP). (Tataru and David 2020) utilized the GloVE word embedding algorithm to preprocess microbiome data prior to using a random forest based method to predict a binary outcome. This paper concluded that performanced based on NLP-inspired preprocessing was superior to that of PCA.

Data Overview

The dataset was found on Kaggle, which originally drew from (Dan et al. 2020), and contains information from a cohort of 143 Autism Spectrum Disorder (ASD) patients and typically developed (TD) patients spanning ages from 2-13 years old which were recruited from May, 2016 to August, 2017. The 143 ASD patients were age and sex-matched with the 143 TD children who were recruited. Additionally, metagenomic analysis of gut microbiota was performed for 30 constipated ASD patients and 30 TD patients, as constipation is another major symptom presenting in ASD patients.

Kaggle has available a 16S rRNA sequence of gut microbiota with 1322 rows and 256 columns. The rows are numbered according to Operational Taxonomic Units (OTUs) which are used to classify groups of closely related individuals, similar to the concept of Linnaean taxonomy or evolutionary taxonomy. OTUs construct "mathematically" defined taxa and their use is widely accepted and applied to describe bacterial communities using amplicon sequencing of 16S rRNA gene (Lladó Fernández, Větrovskỳ, and Baldrian 2019). So, there were 1322 of these OTUs identified. The columns then include a "taxonomy" column, which are lists of the microbiotic species belonging to that OTU. For the rest of the columns, we have 143 columns denoting sample IDs from TD children, and 111 columns denoting sample IDs from ASD children. Down the columns as observations, we see numbers corresponding to each row (OTU) which indicate a measurement of gene abundance.

Additionally, there is a second dataset which contains data from the metagenomic analysis performed for 30 ASD patients and 30 TD patients. The data is presented similarly with patient IDs for 30 ASD and 30 TD children across the top. There are 5619 rows which identify different bacteria present, this time as individual rows rather than in groups like before. Similar to the previous dataset, down the columns as observations, we see measures of abundance in patients. For our project purposes, we will not be using the second dataset, although it could yield interesting results as an extensionn of this project.

Typically, the data will need to be preprocessed by transforming the abundance data into proportional data, as well as removing species/groups with a high proportion of 0. Additionally, the data may require transformation such as a log transformation.

Project Aims

The primary aim of this project is to assess the association between gut microbiome data and ASD patients to see if there are certain bacteria (or bacterial groups) associated more highly with ASD children vs. TD children. After building this model, we would also like to predict whether a child has ASD based on their gut microbiome data.

There exist several obstacles that require attention while analyzing microbiome data. Three of these issues are outlined below:

- Microbiome data exhibits a high-dimensional nature- our dataset comprises of over 1000 species. This high dimensionality necessitates the use of dimensionality reduction techniques to streamline the analysis.
- Microbiome data is over-dispersed, therefore for modeling, the Negative Binomial distribution should be used instead of the Poisson distribution to model the count data
- Microbiome data is zero-inflated. In order to account for this problem, we will use a zero-inflated model for this feature (Xia et al. 2018).

Data Preprocessing

As mentioned, typical methods of preprocessing for microbiome data includes removing species (OTUs in this case) with high counts of 0 as well as transforming into proportional data from abundance data.

However, the model we are implementing makes use of the abundance counts through zero inflated negative binomial (ZINB) regression, so there will be no need for preprocessing in this case.

Methods

For our data set, the response observed are counts, many of which are zero counts. Furthermore, the data set is high dimensional, as the number of OTUs (taxa consisting of "similar" species) exceeds the number of subjects from which data was collected. To counteract the high dimensionality of the data, the first step in our approach is to identify OTUs with significant differential abundance using a zero-inflated negative binomial model that can take into account the numerous zeros observed. In the second step, we include the OTUs in a logistic regression model to classify participants as individuals with or without ASD. In our third step, we will use random forest to also classify participants as individuals with or without ASD. Finally, we compare the two approaches to classifying participants in our fourth step.

In the sections below, we describe in detail our model and methods.

Step 1: Find differential abundance species using the zero inflated negative binomial (ZINB) model

Probability Distribution Functions (PDFs)

PDF of Negative Binomial distribution with mean parameter $\mu_{i,g}$ and dispersion parameter θ_g:

$$f\left(Y_{i,g} = y_{i,g} \mid u_{i,g}, \theta_g\right) = \frac{\Gamma\left(y_{i,g} + \theta_g^{-1}\right)}{y_{i,g}!\Gamma\left(\theta_g^{-1}\right)} \left(\frac{\theta_g \mu_{i,g}}{1 + \theta_g \mu_{i,g}}\right)^{y_{i,g}} \left(\frac{1}{1 + \theta_g \mu_{i,g}}\right)^{\theta_g^{-1}} \tag{1}$$

Suppose $Y_{i,g} \sim ZINB\left(\mu_{i,g}, \theta_g, \pi_{i,g}\right)$,

$$f\left(Y_{i,g} = y_{i,g} \mid u_{i,g}, \theta_g, \pi_{i,g}\right) = \left[\pi_{i,g} + (1 - \pi_{i,g}) \left(\frac{1}{1+\theta_g \mu_{i,g}}\right)^{\theta_g^{-1}}\right]^{I[y_{i,g}=0]} \left[(1 - \pi_{i,g}) \frac{\Gamma\left(y_{i,g}+\theta_g^{-1}\right)}{y_{i,g}!\Gamma\left(\theta_g^{-1}\right)} \left(\frac{\theta_g \mu_{i,g}}{1+\theta_g \mu_{i,g}}\right)^{y_{i,g}} \left(\frac{1}{1+\theta_g \mu_{i,g}}\right)^{\theta_g^{-1}}\right]^{I(y_{i,g}>0)} \tag{2}$$

The indicator exponents mark which parts belong to or don't belong to the zero-inflated part. Next, in order to use the EM algorithm, we need to include a latent variable which indicates membership to the

zero-inflated part. Suppose $Z_{i,g}$ indicates whether or not $Y_{i,g}$ belongs to the zero-inflated part. We can also use the complete data likelihood (CDL) to rewrite 2.

$$f\left(Y_{i,g}=y_{i,g},Z_{i,g}=z_{i,g}\mid u_{i,g},\theta_g,\pi_{i,g}\right)=\left\{\pi_{i,g}^{z_{i,g}}\left[(1-\pi_{i,g})\left(\tfrac{1}{1+\theta_g\mu_{i,g}}\right)^{\theta_g^{-1}}\right]^{1-z_{i,g}}\right\}^{I[y_{i,g}=0]}\left[(1-\pi_{i,g})\tfrac{\Gamma\left(y_{i,g}+\theta_g^{-1}\right)}{y_{i,g}!\Gamma\left(\theta_g^{-1}\right)}\left(\tfrac{\theta_g\mu_{i,g}}{1+\theta_g\mu_{i,g}}\right)^{y_{i,g}}\left(\tfrac{1}{1+\theta_g\mu_{i,g}}\right)^{\theta_g^{-1}}\right]^{I(y_{i,g}>0)\cdot(1-z_{i,g})} \tag{3}$$

Complete Data Likelihood (CDL)

For the gth species, the CDL can be written as

$$\prod_{i=1}^{n} f\left(Y_{i,g}=y_{i,g},Z_{i,g}=z_{i,g}\mid u_{i,g},\theta_g,\pi_{i,g}\right) \tag{4}$$

Since we now have the complete data likelihood, we just need to take the log of it to obtain the log-likelihood to be used in the EM algorithm. Therefore, we can use the EM algorithm to find the MLE for the gth OTU.

EM Algorithm

E-Step Let

- $l_{c,g}$ denote the g^{th} OTU's complete data log-likelihood (CDLL)
- $\zeta_{i,g}=(\mu_{i,g},\pi_{i,g})$
- $\zeta_g=(\mu_{1,g},\pi_{1,g},\mu_{2,g},\pi_{2,g},...,\mu n,g,\pi n,g)$

We want to compute

$$E[l_{c,g}|\zeta_g^{(k)},Y_g]=\textstyle\sum_{i=1}^{n}E[l_{c(i,g)}|\zeta_{i,g}^{(k)},Y_{i,g}]=\sum_{i=1}^{n}E[z_{i,g}|\zeta_{i,g}^{(k)},Y_{i,g}],$$

so we begin by solving $E[z_{i,g}|\zeta_{i,g}^{(k)},Y_{i,g}]$ and obtaining

$$E[z_{i,g}|\zeta_{i,g}^{(k)},Y_{i,g}]=P(z_{i,g}=1|\zeta_{i,g}^{(k)},Y_{i,g})=\frac{\pi_{i,g}^{(k)}}{\pi_{i,g}^{(k)}+(1-\pi_{i,g}^{(k)})(\frac{1}{1+\theta_g^{(k)}\mu_{i,g}^{(k)}})^{[\theta_g^{(k)}]^{-1}}}I(y_{i,g}=0) \tag{i}$$

Using equation (i), we can calculate

$$E[l_{c,g}|\zeta_g^{(k)},Y_g]=\textstyle\sum_{i=1}^{n}E[l_{c(i,g)}|\zeta_{i,g}^{(k)},Y_{i,g}]$$

M-Step To optimize $E[l_{c,g}|\zeta_g^{(k)},Y_g]$ with respect to $\beta^{(k)}$, $\gamma^{(k)}$, and $\theta^{(k)}$ we implement Newton-Raphson's algorithm. Hence, the derivation of the score function and Fisher's information matrix is necessary.

Score Function: We derive the score function by solving the following:

- $\frac{\partial l_{c,(i,g)}}{\partial\beta_g}=\frac{\partial l_{c,(i,g)}}{\partial\mu_{i,g}}\frac{\partial\mu_{i,g}}{\partial\beta_g}=\frac{1-z_{i,g}}{(1+\theta_g\mu_{i,g})\mu_{i,g}}(y_{i,g}-\mu_{i,g})\exp(x_i^{\intercal}\beta_g)x_i$,
- $\frac{\partial l_{c,(i,g)}}{\partial\gamma_g}=\frac{\partial l_{c,(i,g)}}{\partial\pi_{i,g}}\frac{\partial\pi_{i,g}}{\partial\gamma_g}=\frac{z_{i,g}-\pi_{i,g}}{\pi_{i,g}(1+\pi_{i,g})}\frac{\exp(x_i^{\intercal}\gamma_g)}{(1+\exp(x_i^{\intercal}\gamma_g)}$, and

- $\frac{\partial l_{c,(i,g)}}{\partial \theta_g} = (1 - z_{i,g})\theta_g^{-2}\psi(\theta_g^{-1}) + \psi(y_{i,g} + \theta_g^{-1}) + \log(1 + \theta_g\mu_{i,g}) + \frac{1}{\theta_g(1+\theta_g\mu_{i,g})(y_{i,g}-\mu_{i,g})}$, where $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$

Let $S(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)})$ be the score function for the current iteration k.

$$S(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) = \begin{bmatrix} X^\mathsf{T} V(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)} e_1(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) \\ X^\mathsf{T} e_2(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) \\ \frac{\partial l_c}{\partial \theta_g}|_{\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}} \end{bmatrix}, \tag{ii}$$

where $e_1(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) = \begin{bmatrix} (1 - z_{1,g})(y_{1,g} - \mu_{1,g}^{(k)}) \\ (1 - z_{2,g})(y_{2,g} - \mu_{2,g}^{(k)}) \\ \vdots \\ (1 - z_{n,g})(y_{n,g} - \mu_{n,g}^{(k)}) \end{bmatrix}$, $e_2(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) = \begin{bmatrix} z_{1,g} - \pi_{1,g}^{(k)} \\ z_{2,g} - \pi_{2,g}^{(k)} \\ \vdots \\ z_{n,g} - \pi_{n,g}^{(k)} \end{bmatrix}$, and $\mathrm{V}(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) = \mathrm{diag}\{\frac{1}{1+\theta_g \exp(x_i^\mathsf{T}\beta_g^{(k)})}\}$.

Fisher Information Matrix: We derive the fisher information matrix by solving the following:

- $-E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \beta_g \beta_g^\mathsf{T}}\big] = X^\mathsf{T} P_1(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) X$, where $P_1(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) = \mathrm{diag}\{\frac{(1-z_{i,g})\exp(x_i^\mathsf{T}\beta_g^{(k)})}{1+\theta_g \exp(x_i^\mathsf{T}\beta_g^{(k)})}\}$
- $-E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \beta_g \gamma_g}\big] = 0$,
- $-E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \beta_g \theta_g}\big] = 0$,
- $-E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \gamma_g \gamma_g^\mathsf{T}}\big] = X^\mathsf{T} P_2(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) X$, where $P_2(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) = \mathrm{diag}\{\frac{(\exp(x_i^\mathsf{T}\gamma_g^{(k)})}{1+\theta_g \exp(x_i^\mathsf{T}\gamma_g^{(k)})^2}\}$,
- $-E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \gamma_g \beta_g}\big] = 0$,
- $-E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \gamma_g \theta_g}\big] = 0$, and
- $-E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \theta_g^2}\big] \approx (1 - z_{i,g})\theta_g^{-4}(2\theta_g\psi(\theta_g^{-1}) - 2\theta_g\psi(y_{i,g} + \theta_g^{-1}) + 2\theta_g(1 + \theta_g\mu_{i,g}) + \psi'(\theta_g^{-1}) - \psi'(y_{i,g} + \theta_g^{-1} - 2\theta_g^2(\frac{\mu_{i,g}}{1+\theta_g\mu_{i,g}})]$

Let $I_n(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)})$ be the fisher information matrix for the current iteration k.

$$= \begin{bmatrix} X^\mathsf{T} P_1(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) X & & I_n(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) \\ & X^\mathsf{T} P_2(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) X & \\ & & E\big[\frac{\partial^2 l_{c,(i,g)}}{\partial \theta_g^2}\big]|_{\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}} \end{bmatrix} \tag{iii}$$

We update $\beta_g^{(k)}$, $\gamma_g^{(k)}$, and $\theta_g^{(k)}$ by solving the following:

$$(\beta_g^{(k+1)}, \gamma_g^{(k+1)}, \theta_g^{(k+1)})^T = (\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)})^T + [I_n(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)})]^{-1} S(\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}) \tag{iv}$$

The following steps summarize the EM algorithm:

1. Initialize $\beta_g^{(0)}$, $\gamma_g^{(0)}$, and $\theta_g^{(0)}$

For the k^{th} iteration

2. E-step: Calculate equation $E[l_{c,g}|\beta_g^{(k)}, \gamma_g^{(k)}, \theta_g^{(k)}, Y_g]$
3. M-step: Calculate the score function and fisher information matrix to solve equation (iv).

Steps 2 and 3 are repeated until convergence.

Likelihood Ratio Test

In order to test the differential abundance status, we need to model $\log(\mu_{i,g}) = X_i^\top \beta_g$, where $\beta_g = (\beta_g^{(Intercept)}, \beta_g^D)^\top$, $\text{logit}(\pi_{i,g}) = X_i^\top \gamma_g$, where $\gamma_g = (\gamma_g^{(Intercept)}, \gamma_g^D)^\top$.

We just need to test if $H_0 : \beta_g^D = \gamma_g^D = 0$ holds. And we will use LRT to test this hypothesis.

$$\text{LRT} = \prod_{i=1}^{n} \frac{f\left(Y_{i,g} = y_{i,g} \mid u_{i,g}^{(0)}, \theta_g^{(0)}, \pi_{i,g}^{(0)}\right)}{f\left(Y_{i,g} = y_{i,g} \mid u_{i,g}^{(1)}, \theta_g^{(1)}, \pi_{i,g}^{(1)}\right)} \tag{5}$$

Under H_0, $-2\log(\text{LRT}) \to \chi^2(2)$. We can test the species one by one and find the significantly differential abundance OTUs as the input data for the next step.

Step 2: Using logistic regression to predict ASD status using the differential abundance species

Using the identified OTUs from step 1, we will fit a logistic regression to predict ASD status, a binary outcome, using the measurements as count data.

Step 3: Compare to Random Forest

Finally, we will compare the performance of our model with a Random Forest model from the "caret" package in R. Random forest classification will be implemented to assess the ability of the gut microbiome data to predict ASD status. The implementation of random forest classification improves the disadvantages of just having one classification tree. Any one classification tree tends to not "have the same level of predictive accuracy as some of the other regression and classification approaches" and are susceptible to large changes because of small changes in data (James et al. 2021). Additionally, random forests run efficiently on large datasets and have high performance on classification algorithms. Rather than performing this random forest with the step 1 selected differential abundant OTU species, we will use the entire dataset, as the algorithm will decide for itself which of the OTUs is important as a feature for predicting ASD status.

Tuning parameters that can be adjusted include the number of trees generated, the subset of the number of predictors considered at each split, and the measure used to assess the quality of the split, among others.

Before performing the analysis, the data will be split into a training and testing set following an 80/20 train/test split. The training will be performed on the training data for both the random forest and the logistic regression model.

The data will be split into 10 folds within the training data. The logistic regression will be trained on 9 folds, with 1 fold held back as the testing data. This process will be repeated until we have 10 different

performance scores and the model has been trained 10 times. We will perform the same process to assess the performance of the random forest algorithm. A 10-fold cross validation procedure where participants a part of the study are randomly allocated to 1 of 10 independent folds of approximately equal size. The training and testing are iterated over such that each individual fold was treated as a testing data set on its own once, while the remaining 9 folds are treated as the training data set. Measurements used to assess the random forest algorithm include accuracy, sensitivity, and specificity.

- Accuracy will be calculated by summing the number of correctly predicted individuals with ASD divided by the total number of individuals.
- Sensitivity will be calculated by taking the total number of correctly predicted individuals with ASD and dividing by the total number of individuals with ASD.
- Specificity will be calculated by taking the total number of correctly predicted individuals without ASD and dividing by total number of individuals without ASD.

In order to tune the parameters, a grid search will be conducted during the 10-fold cross validation procedure in the random forest.

Finally, to compare logistic regression and random forest classification performance, ROC AUCs will be computed and compared for the two, as well as the aforementioned sensitivity and specificity calculations for both models, after applying the trained models on the unseen test data.

Package

The "caret" package is a pre-written package that allows us to use a pre-coded random forest model and allows us to perform cross validation simultaneously.

Our package, the "ZINB" package includes the following functions:

- LRT1D: Tests the intercept model against the model fitted by the design matrix. Outputs parameters of the fitted model, chisquared statistics, degrees of freedom, and p-values.
- fit.zinb: Conducts ZINB negative binomial regression on the data. Estimates parameters and their covariance using the EM algorithm. Outputs estimates for the coefficients of the negative binomial model, the zero inflated model, the reciprocal of the size parameter for the negative binomial distribution, the covariance matrix for all estimated parameters, aand the log-likelihood of the estimated model.
- LRTnD: Tests for the intercept model against the model fitted by the design matrix– allows for multiple responses. Outputs parameters of the fitted model, chisquared statistics, degrees of freedom, and p-values.

Results

As a preface, while the original model intention was written as above, there were still too many OTU's selected by the ZINB model for the logistic regression to converge. Thus in our final analysis, we used logistic LASSO regression, which can perform variable selection and thus further reduce the dimensions of our predictors.

```
## Loading Package
library(devtools)
library(Rcpp)
library(dplyr)
library(ZINB)
library(caret)
library(glmnet)
library(pROC)

## load function which trains model and calculates confusion matrix
source("function.R")

## read in data
microbiome = t(read.csv("Data/microbiome.csv",row.names = 1)[,-1])

# split data into train and test
set.seed(1)
train.num = sort(sample(1:nrow(microbiome),0.8*nrow(microbiome)))
train = microbiome[train.num,]
test = microbiome[-train.num,]

## analyze the data
result = model_compare(train,test, seed = 1)
```

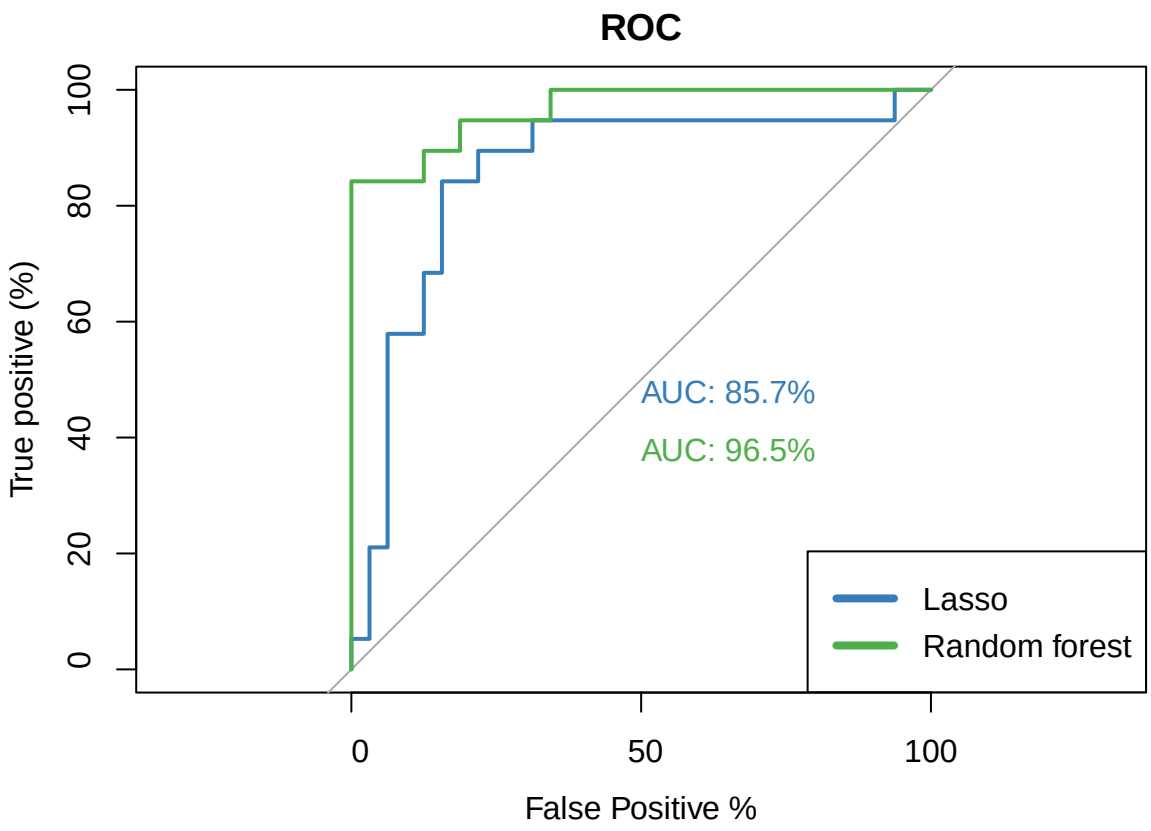
The model_compare function does the following:

- Implements the ZINB package to subset OTUs that are then included in the logistic LASSO model. Trains the logistic LASSO model using a 10 fold cross validation and tests the best fitted model that is selected based on accuracy.
- Trains random forest using a 10 fold cross validation and tests the best fitted model that is selected based on accuracy.

```
## generate ROC curve and AUC score
roc(result$status,as.numeric(result$pred.lasso),plot=T,percent = T,
    legacy.axes = T,print.auc = T,col = "#377eb8",
    ylab = "True positive (%)",xlab = "False Positive %",main = "ROC")
```

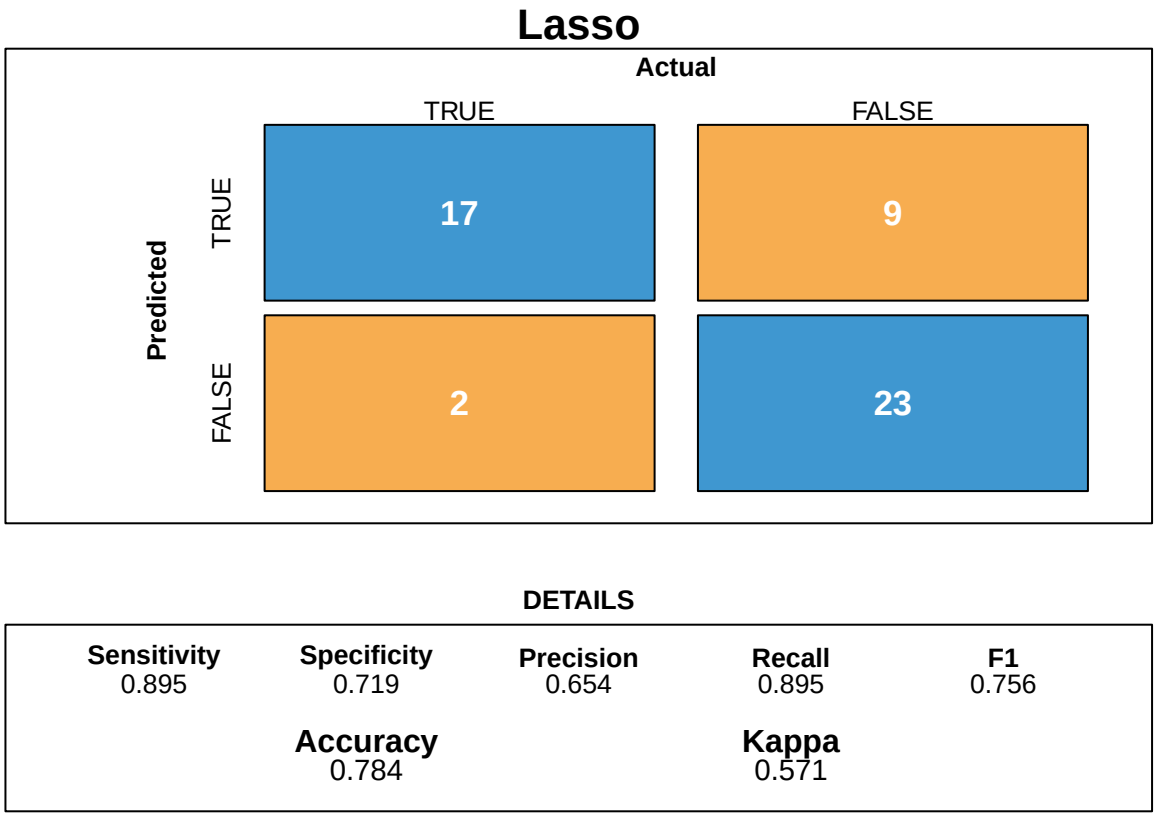
```
##
## Call:
## roc.default(response = result$status, predictor = as.numeric(result$pred.lasso),     percent = T, pl
##
## Data: as.numeric(result$pred.lasso) in 32 controls (result$status FALSE) < 19 cases (result$status T
## Area under the curve: 85.69%
```

```
plot.roc(result$status,as.numeric(result$pred.rf),percent = T,add=T,
         print.auc = T,print.auc.y = 40,col="#4daf4a")
legend("bottomright",legend = c("Lasso","Random forest"),col=c("#377eb8","#4daf4a"),lwd=4)
```

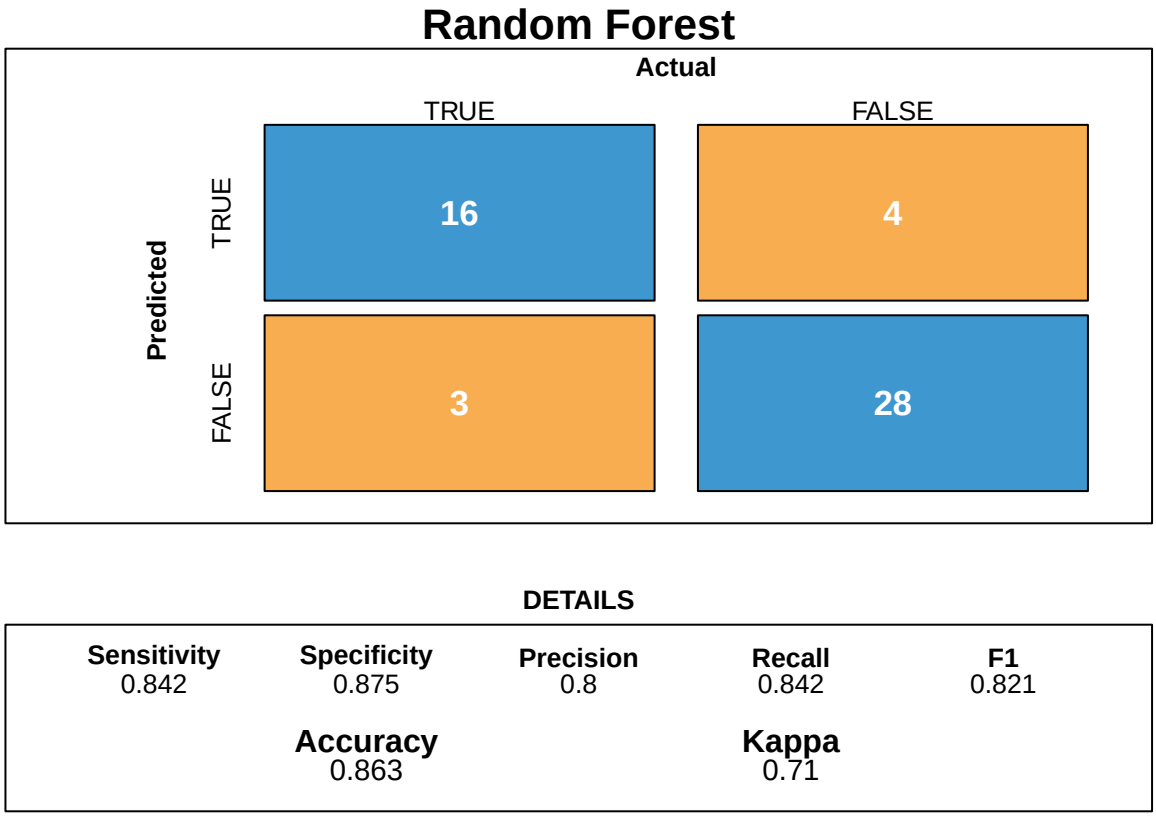


The curve (AUC) for the logistical LASSO model and random forest are 85.7% and 96.5% respectively. From the figure we note that for most thresholds random forest performs better than the logistic LASSO model at classifying individuals with and without ASD.

```
## generate information of confusion matrix
## calculate accuracy, specificity, sensitivity metrics for both model

result$status = relevel(result$status,"TRUE")
result$pred.lasso = as.factor(result$pred.lasso > 0.5)
result$pred.lasso  = relevel(result$pred.lasso,"TRUE")
result$pred.rf = as.factor(result$pred.rf > 0.5)
result$pred.rf  = relevel(result$pred.rf,"TRUE")

cm.lasso <- confusionMatrix(data = result$pred.lasso, reference = result$status)
cm.rf <- confusionMatrix(data = result$pred.rf, reference = result$status)

# Confusion matrix for penalized logistic
draw_confusion_matrix(cm.lasso,title = "Lasso")
```




```
# Confusion matrix for random forest
draw_confusion_matrix(cm.rf,title = "Random Forest")
```



The two confusion matrices and corresponding performance metrics of the logistic LASSO model and random forest reinforce that random forest outperforms the logistic LASSO model. Specifically, the confusion matrices are based on individuals being classified with ASD if their prediction probability is greater than 0.50. The accuracy of the logistic LASSO model (0.784) is considerably smaller than the accuracy of random forest (0.863). The specificity of the logistic LASSO model is 0.719 and 0.875 for random forest. The sensitivity of the logistic LASSO model was 0.895 and the sensitivity ofthe random forest was 0.842.

Discussion

In summary, we found that all measurements of performance (area under the curve, accuracy, sensitivity, and specificity) on the test data indicate that random forest outperforms the logistic LASSO model that used a subset of predictors selected by the ZINB model.

The test data included a total of 51 participants of which 19 were individuals with ASD and 32 were individuals without ASD. The confusion matrices in the results provide context for the performance measurements, showing that logistic LASSO model and random forest both correctly identified 17 individuals as having ASD. With respect to correctly identifying individuals without ASD random forest correctly identified nine more individual compared the logistic LASSO model.

We can also look at variable importance scores, which were output from the random forest, to give insight into which OTUs specifically were the most important predictors in the ASD classification problem. We can see that the following OTUs were the most important in prediction along with the genus of the OTU members.

Weighted sums of the absolute regression coefficients. The weights are a function of the reduction of the sums of squares across the number of partial least square components and are computed separately for

each outcome. Therefore, the contribution of the coefficients are weighted proportionally to the reduction in the sums of squares (Kuhn and Max 2008).

First Five Most Important Variables by Partial Least Squares		
OTU Number	Importance Value	Genus
1225	100.00	Lachnoclostridium
625	93.30	Prevotella 2
1301	81.47	Megasphaera
913	77.04	Ruminococcaceae NK4A214 group
784	76.27	Eubacterium xylanophilum group

Figure 1: Variable Importance Scores

Limitations

1. ZINB is unstable when there are fewer zeros than expected in the data.
2. We used a ZINB model and iteratively included 1 of 1322 OTUs as a covariate in the model to determine whether there is significant differential abundance based on the data. The iterative procedure was done to determine which OTUs to include in a final logistic model to predict ASD status. The procedure substantially reduced the number of OTUs to be considered in the the logistic model to 206 OTUs. 206 OTUs was still too many OTUs to be considered without introducing a penalty that would result in natural variable selection.
3. The test set is small with only 51 individuals.

Future Directions

As an extension of our approach in order to use the model as intended, we may consider merging OTUs by taxonomy to create higher order clades and therefore fewer OTUs to be used as predictors. This may solve the instability we saw in the ZINB model, where if the counts contained too few or too many 0 values, the information was almost singular. However, this does not appear to be a problem specific to our package, as other pre-written packages encounter this issue as well when we compare our results. To eliminate the two step approach, one could consider a ZINB model with a penalty such as LASSO, "smoothly clipped absolute deviation (SCAD), or minimax concave penalty (MCP)" (Wang, Shuangge, and Wang 2015).

References

Dan, Zhou, Xuhua Mao, Qisha Liu, Mengchen Guo, Yaoyao Zhuang, Zhi Liu, Kun Chen, et al. 2020. "Altered Gut Microbial Profile Is Associated with Abnormal Metabolism Activity of Autism Spectrum Disorder." *Gut Microbes* 11 (5): 1246–67.

Elsabbagh, Mayada, Gauri Divan, Yun-Joo Koh, Young Shin Kim, Carlos Kauchali, Cecilia Montiel-Nava, Vikram Patel, Cristiane S Paula, Chongying Wang, et al. 2012. "Global Prevalence of Autism and Other Pervasive Developmental Disorders." *Autism Research* 5 (3): 160–79.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Landrigan, Philip J. 2010. "What Causes Autism? Exploring the Environmental Contribution." *Current Opinion in Pediatrics* 22 (2): 219–25.

Lladó Fernández, Salvador, Tomáš Větrovskỳ, and Petr Baldrian. 2019. "The Concept of Operational Taxonomic Units Revisited: Genomes of Bacteria That Are Regarded as Closely Related Are Often Highly Dissimilar." *Folia Microbiologica* 64: 19–23.

Lord, Catherine, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele. 2018. "Autism Spectrum Disorder." *The Lancet* 392 (10146): 508–20.

Matson, Johnny L, and Alison M Kozlowski. 2011. "The Increasing Prevalence of Autism Spectrum Disorders." *Research in Autism Spectrum Disorders* 5 (1): 418–25.

Sutcliffe, James S. 2008. "Insights into the Pathogenesis of Autism." *Science* 321 (5886): 208–9.

Tataru, Christine A, and Maude M David. 2020. "Decoding the Language of Microbiomes Using Word-Embedding Techniques, and Applications in Inflammatory Bowel Disease." *PLoS Computational Biology* 16 (5): e1007859.

Vuong, Helen E, and Elaine Y Hsiao. 2017. "Emerging Roles for the Gut Microbiome in Autism Spectrum Disorder." *Biological Psychiatry* 81 (5): 411–23.

Wang, Zhu, Ma Shuangge, and Ching-Yun Wang. 2015. "Variable Selection for Zero-Inflated and Overdispered Data with Application to Health Care Deman in Germany." *Biometrical Journal* 57 (5): 867–84.

Xia, Yinglin, Jun Sun, Ding-Geng Chen, Yinglin Xia, Jun Sun, and Ding-Geng Chen. 2018. "Modeling Zero-Inflated Microbiome Data." *Statistical Analysis of Microbiome Data with R*, 453–96.