STAT 542 Project Report

Reuven Birnbaum (birnbam2) – **leader**,
Ashley Shang (shumin2),
Youngwook Jung(yj10)

Contents

1 Project Description and Summary

Our main goals within this report, was to performed two kinds of analysis using microbiome data. One is to find clusters that share some similarity based on the OTU (Operational Taxonomic Unit) variables in observations. The other is to estimate how OTU variables, along with basic demographic variables (race, sex, age, etc.), can explain physical characteristics (BMI) or habit characteristics (alcohol frequency).

One of the biggest challenges in achieving these goals was data wrangling. There are only 1,876 columns out of the total 32,954 have more than 1% non-zero values. We assumed that if including the OTUs, which consist of almost all zero values, the computation cost will be much more complex and we could not earn much information benefit. In other words, even though we chose to only use the OTU variables with more than 1% non-zero values, it can still give enough useful information.

In addition to our data cleaning process, our next approach is dimension reduction. Rather than using all OTU variables in every analysis, PCA was applied on OTU features with more than 1% non-zeros to speed up the process as well as extract the useful information. Specifically, by evaluating the cumulative percentage that gauges how much the PCA components contribute to the total variation, the top 69 components, so called projected OTUs, were selected.

Our innovative method has helped us to run various analyses and create explanatory visualization. However, unfortunately, in unsupervised learning analysis, we found it difficult to find distinct clusters that could clearly explain even one of the categorical variable. But we chose to label the data set to see if the clustering has further predictive power.

In supervised learning section, our first task is to predict the BMI. After removing some outlier observations, we implemented least square regression with Lasso penalty as well as with AIC and BIC model selection criteria. As expected, there is a strong linear relationship between weight and BMI in both models. Next, to predict alcohol consumption frequency, logistic and KNN classification models are employed. However, the misclassification rates for logistic regression and KNN are up to 0.63 and 0.74, respectively, which is not ideal compared to other methods. Finally, in the BMI category classification section, tree models are used, specifically gradient boosting tree and random forest. The highest accuracy reached nearly 80%, but we found that the k-means labels are not useful, even could lower the model performance sometimes. The hierarchical clustering result can slightly improve the model performance, but not tremendously.

In conclusion, we tried to understand the data using various methods of unsupervised learning. It was difficult to find potential clusters due to low predictive power. We also analyzed the effects on BMI, BMI category, and alcohol frequency by performing supervised learning methods with newly selected variables, although it is a somewhat naive. As many previous studies emphasized, it was not easy to deal with the intertwined microbiome factors, but this paper made a new contribution to selecting and using the informative projected-OTU variables based on the explanatory power of the overall variation.

2 Literature Review

An overwhelming portion of the data we are using in this report was collected and processed from the American Gut Database [1]. As such it its imperative that we first provide some background on the properties of this data and the common approaches used to analyze this data. The human body is composed of many trillions of cells. Most of which are non-native prokaryotes (bacteria), archaea, fungi, and viruses. The composition of these "foreign" bodies are collectively referred to as the microbiome. As these microbial cells almost always function in some cooperative and competitive balance with other microbial cells and their environment, the effect on the individual they reside in is a much studied topic. As a single individual OTU's values are ratios of specific microbial species within the host, not only do the values ad up to one but many of them are 0. As such, performing supervised and unsupervised learning tasks on this dataset compromises a tricky task, as not only are there a very large number of possible features to learn from, most of the data is therefore sparse. As such we report on a few studies and review articles within the field that address these challenges.

Hongzhe Li in his 2015 review [2], discusses a plethora of methodologies, and what advantages and limitations they have. Of particular interest to us was his reporting of phylogenetic tree–based distances, and differential abundance analysis. In phylogenetic tree-based distance, distances between samples of gut

bacteria are calculated based on their photogenic tree (evolutionary similarity). To explains this method further, if you consider the reasonable expectation that if two species are evolutionary close to each other then the amount that they appear in a sample will be correlated. To account for this, Li presents a few metrics that weight certain closely related organism less so than distantly related organism, thus removing the correlation. This phylogenetic tree–based distance can then be used to perform clustering methods, or can be used to define a kernel functions to be used in general regression methodologies that can naturally account for these covariates. In Differential abundance analysis, the purpose is to find if any organisms have different distributions given certain conditions. The sparsity or microbiome data, makes this task especially hard but Li references a regression model which first models the distribution of an certain organism with a gamma function with a log-link mean function, and a set of covariates that and associated regression parameter. Therefore a log likelihood test of the mean can be performed to see if two distributions given a separate condition. Finally, we should also not that Li explains the issues with using many regression techniques to analyze microbial data, as not only do the total parameters for an observations should always equal one, there are many complex interactions that are possible that can not be easily found. These two conditions make many protocols of regression difficult. namely considering a single feature changing while others remain the same. Nevertheless, for this report we will proceed by employing a number of regression tasks to show their use and power with such a complex dataset.

One of the dependent variables in our analysis is a alcohol consumption frequency. We will show the variables that affect the frequency in detail later. In our analysis, we had to take into account a variety of factors, especially the tens of thousands of genetic factors. When a certain result was shown, it was necessary to clarify from which cause the effect. Dubinkina, et al.([3])'s work is particularly interesting in this respect. They found the impact of alcohol consumption on the gut community structure, and especially focused on the identification which distinguishes the impact from psychiatric symptoms caused by alcohol abuse and that from the liver dysfunction.

To deal with more than 30,000 OTUs, we adopt two steps to refine the data as explained below. So we analyzed over 30,000 OTU variables by reducing them to a maximum of 69 projected-OTUs. Pollock et al.(2018) gave us an intuitive explanation for our approach ([4]). They argued that if consistency of dataset is assured (any new observations will follow old patterns), the method has the potential to solve the problems, even if bias can be shown into microbiome studies from sampling. Determining whether a methodology that we used guarantees consistency is beyond the scope of this study, so it was not possible to proceed explicitly. However, when plotting the results using various methodologies, the results did not fluctuate much. It can be argued that some degree of consistency is guaranteed in our paper.

Although we haven't been able to find a study which uses the same method with ours, our methodology is similar to described in Debelius et al. (2016) ([5]). They found that for both biological and technical covariates setting effect sizes of them can improve the quality of results. Though they used Principal Coordinates Analysis (PCoA), plotted them, and analyzed the key components using P values, we similarly utilized PCA and set optimal sizes of covariates.

These previous studies could not prove that the method we used was theoretically rigorous, but what they pointed out that we should be careful when reducing a large dimension data. From this reason, our study did not blindly remove OTU variables, but instead selected variables according to one clear criterion, and adopted a method of reducing the dimension via PCA. The detailed explanation is given in the next section.

3 Data Properties

The microbiome data was processed from the American Gut database [1]. This dataset contains 9,511 samples and 32,954 OTU (Operational Taxonomic Unit) variables at the species level. This specific dataset consists of demographic variables *Race* and *Sex*, a continuous health outcomes *BMI* (body mass index), and two categorical health outcomes *BMI category*, and *alcohol consumption frequency*. Among these variables, *BMI* will be used for the regression task in Supervised Learning section, and *BMI category* and *alcohol consumption frequency* for classification tasks in the same section.

OTU variables have compositional nature of microbiome data, meaning that the sum of all OTUs for each subject is 1. Under this concept, the dataset is extremely sparse, with 92.83% of the variables only has less than 1% nonzero entries. Data processing part can be found at section 4.1.

4 Unsupervised Learning

4.1 Pre-processing and Pipeline

There are around 12% records containing one or more missing values in the original dataset. Details for each column can be found below.

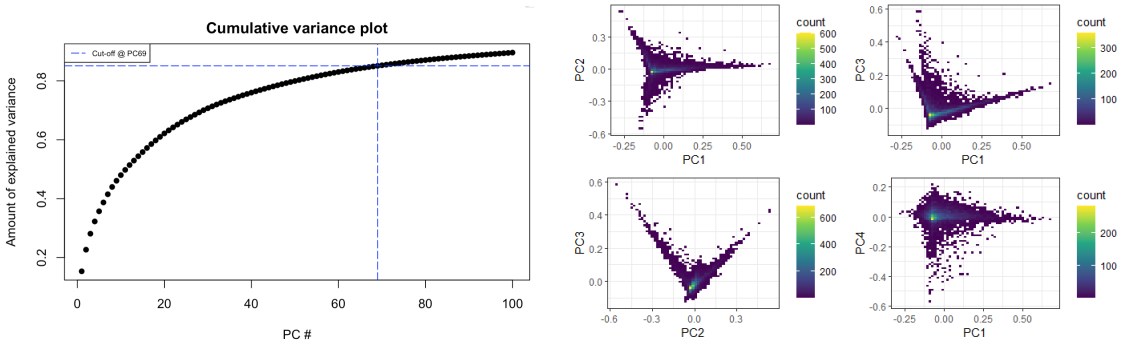
Table 1: Missing values in the dataset

variable	age_cat	bmi	bmi_cat	weight_kg	race	sex	alcohol_frequency
count	547	308	578	252	189	368	215
portion	5.75%	3.24%	6.08%	2.65%	1.99%	3.87%	2.26%

On the other hand, rows only have invalid BMI/BMI_Cat are 6% of the total data. To keep the tasks consistent between our supervised learning tasks, we decided to drop those values considering the BMI is the response for the machine learning tasks. We furthermore dropped any row that contained missing values to make our regression tasks more concrete.

From the original 32,954 OTU columns, only 1,876 columns have more than 1% non-zero values. We decided to only use those features, mainly because we think only with more than 1% valid values can a feature provide useful information for modeling. We further checked this claim by checking the amount of unique non-zero values within each column of the OTU attributes. We found that all of columns have more than 80% unique nonzero values, which further cements that by removing any attribute with a large number of zeros, we will eliminate any feature that would likely be useles for our supervised and unsupervised tasks.

After the pre-processing, PCA was applied on the remaining OTU features to extract the significant components in order to lower the dimension and fasten the training process. Our strategy to select the top components from PCA is based on the variance contribution, with cutoff 85% set. Figure 1(a) shows that the first 69 components reached 85% contribution on the total variance.



(a) PCA component selection

(b) Pairs of PCA Projections

Figure 1: PCA result

For the clustering tasks, we only considered clusters within those top 69 components. Figure 1(b) shows how data projected is projected on the first 4 principle components, and as can be seen, there does not seem to be any graphical evidence of distinct clusters within the data. However, the data is organized in interesting star shapes, which will likely allow our clustering tasks to pick up on these formations.

4.2 Self-Organized Maps

Our first unsupervised learning task was to employ self organized maps as they can not only be used to cluster observations but also for dimensional reduction by projecting the data into a large number of interconnected centroids that can give pointers to our later clustering tasks. We could then use these maps to identify the number of clusters if we see that a portion of nodes are closely connected but are surrounded by nodes that are not. A general rule of thumb for the total number of centroids to use is $5\sqrt{\text{observations}} \approx 457$,

which we can represent with a 21×2 hexagonal grid where each centroid/node is connected to at most 6 other nodes. The results of those parameters are shown in Figure 2. We can see that there does not seem to be large evidence of distinct clusters, as we do not see any areas of highly connected nodes surrounded by nodes that are minimally connected, and the area in the middle that is highly connected does not show any evidence of being more populated than others. As such we can proceed with using more exhaustive searches to find optimum clustering.

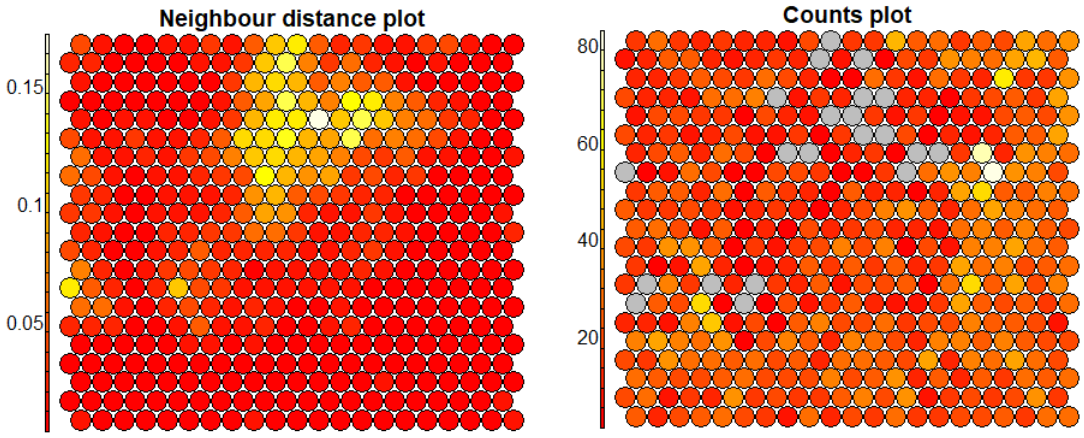


Figure 2: Self Organized Map Results

We further employed self-organized maps to perform lower node clustering to be used in our future classification tasks. In particular we employ clusters of 2 through 8, we can see the results of these clusters in Figure 3, and as we hypothesized above the clusters did organize themselves into the spikes of these formations. We will use these results in the future with various supervised learning tasks.

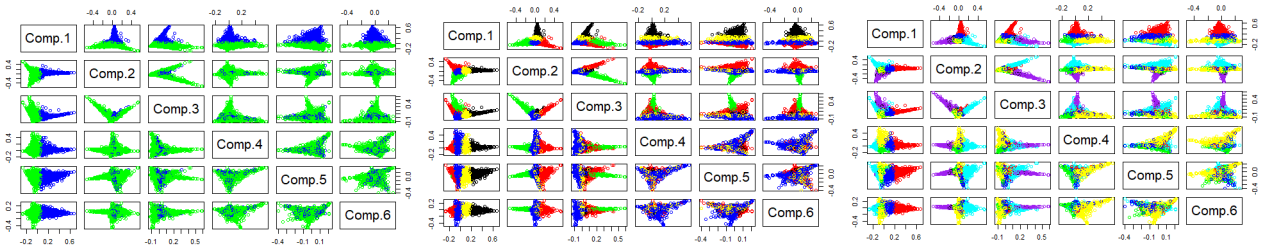
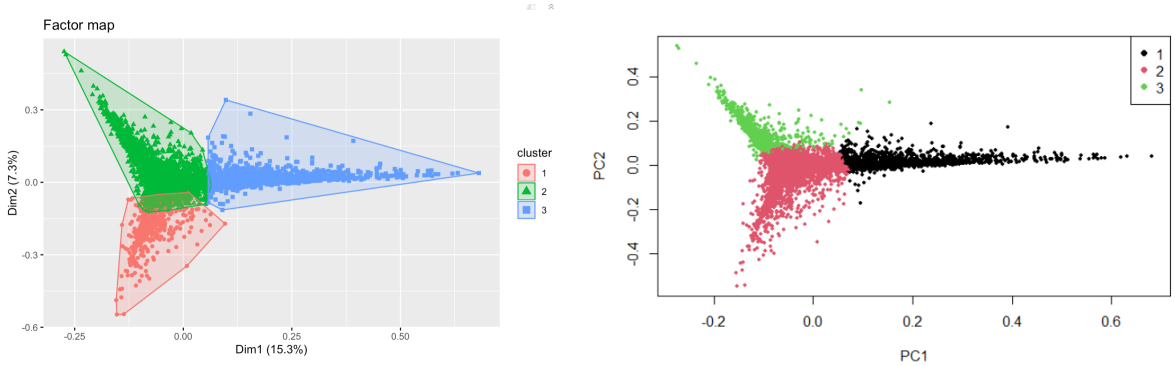


Figure 3: SOM Clustering Results.
Left: K=2, Center: K=5, Right: K=7

4.3 Hierarchical clustering with PCA

Applying hierarchical clustering on the first 69 components of the PCA results on OTU features eventually gives us 3 clusters. Visualization on the principal component map shown as figure 4(a) shows the criterion hierarchical clustering followed in this case this the main component distribution. Just like we analyzed before, the clustering was based on the star-shaped formation of the PCA.



(a) Hierarchical Clusters Visualization

(b) K-means Visualization

Figure 4: Clustering visualization

To further investigate the predictive power of the clustering result, we looked into the categorical variables separately. Below tables and figure show us the total amount of records per category falls in each cluster.

Table 2: Clusters per race

cluster	African American	Asian or Pacific Islander	Caucasian	Hispanic	Other
cluster 1	1	23	336	12	11
cluster 2	38	271	5558	118	174
cluster 3	15	89	1618	42	53

Table 3: Clusters per alcohol frequency

cluster	Daily	Never	Occasionally	Rarely	Regularly
cluster 1	38	77	90	98	80
cluster 2	645	1383	1399	1524	1208
cluster 3	181	403	395	475	363

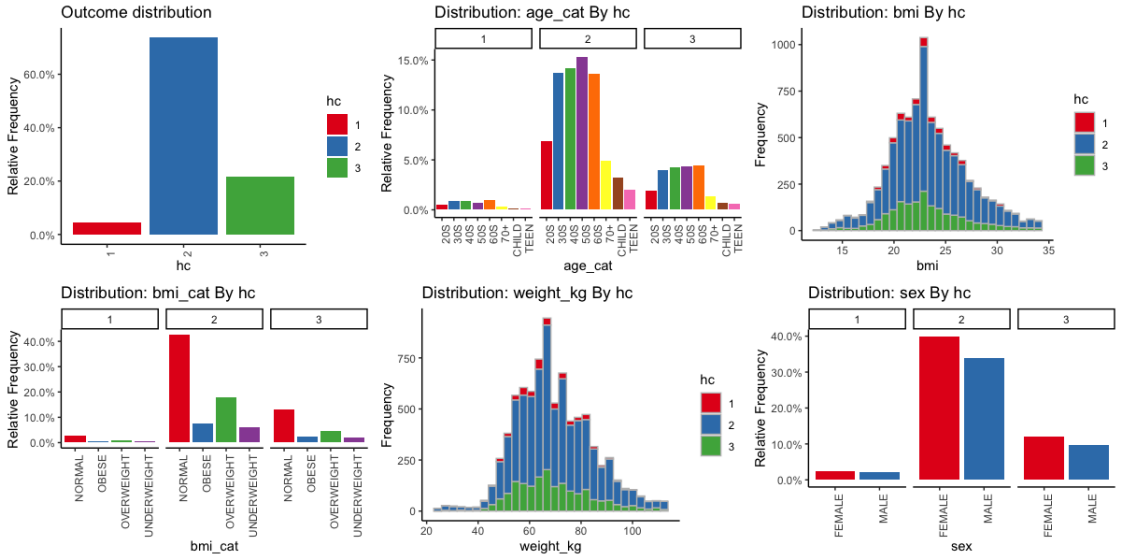


Figure 5: Hierarchical clustering predictive power

Figure 5 visually shows us the distribution of more variable values by hierarchical clusters. As we can see, the predictive power of this clustering is really low. In other words, it cannot contribute enough explanation

on either variables. At this time, we still think this clustering result is useful on some level. We decide to use the clusters as labels and add them to the data for further classification use.

4.4 K-means

K-means clustering splits the total observations into k clusters, and then each observation belongs to the cluster with the closest centroid. It minimizes within-cluster variances, which is very intuitive and reasonable, but it is computationally difficult to run. Therefore, it is unlikely to get result within a reasonable period of time utilizing all 32,962 variables. Thus, we used the filtered OTU instead, as explained before.

Here, we used five types of K. That is, the numbers of centers are 3, 4, 5, 6, and 7. For instance, in case of 3 centers, each observation is assigned between 1 to 3. Similar to the other unsupervised learning before, it shows the star-shaped (Figure 4(b)). In addition, by assigning k equals to 4, 5, and 8, we tried to match categories of variables *bmi_cat*, *alcohol_frequency*, and *age_cat*, respectively. However, the predictive power of this clustering is also really low. For example, in Table 4, all four clusters contain a large portion of Normal BMI category, and we cannot find a specific cluster that can distinguish BMI categories well.

Table 4: K-means (Centroid=4)

	BMI category			
Cluster	**Normal**	**Obese**	**Overweight**	**Underweight**
1	143	18	59	19
2	239	36	78	32
3	1075	183	375	152
4	3432	609	1432	477

Table 5: K-means (Centroid=5)

	Alcohol frequency				
Cluster	**Daily**	**Never**	**Occasionally**	**Rarely**	**Regularly**
1	91	165	202	197	149
2	19	47	52	55	45
3	223	508	508	569	475
4	62	136	136	172	118
5	469	1007	986	1104	864

Table 6: K-means (Centroid=8)

	Age category							
Cluster	**20s**	**30s**	**40s**	**50s**	**60s**	**70+**	**child**	**teen**
1	322	629	650	702	661	225	155	84
2	18	39	39	34	33	14	4	5
3	61	143	144	170	115	48	27	27
4	107	216	220	236	210	66	48	23
5	14	20	30	20	16	11	2	10
6	40	74	71	54	72	29	12	8
7	54	116	94	110	97	22	16	13
8	164	314	370	378	388	137	77	51

4.5 Cluster Results Comparison

To compare the clustering we present Table 7, which shows the count of observations assigned to one cluster label vs. another. As the actual cluster label is meaningless on it's own, we are interested in observing if there is a label that is highly correlated with another. As we can observe, the K-means labels 1, 2, and 3 are highly correlated with the SOM labels 3, 1, and 2 respectively. while hierarchical does not appear to show any distinct pattern with SOM or K-means. This makes sense as cluster centroids are generated similarly for k-means and SOM. To compare these results further, the total within cluster distance came out to 1,663.28, 1,646.49, and 1,646.40 for hierarchies, K-means, and SOM respectively. Not surprisingly, the K-means and SOM results are quite similar to one another while the hierarchical results differ from them both.

Table 7: Side-by-Side Cluster Comparison

	Hierarchical		
SOM	1	2	3
1	382	5167	1
2	0	875	7
3	1	117	1809

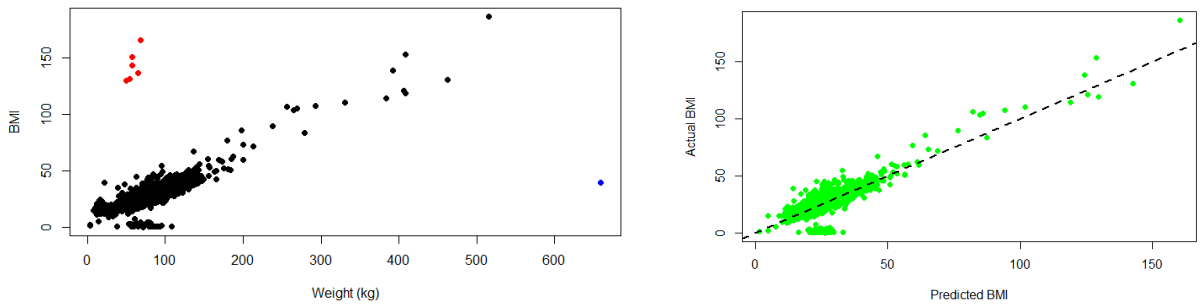
	K-Means		
SOM	1	2	3
1	0	5546	4
2	0	33	849
3	1724	202	1

	Hierarchical		
K-Means	1	2	3
1	1	1	1722
2	381	5312	87
3	0	846	8

5 Supervised learning

5.1 Regression

Our first task in our supervised learning task is to predict the numerical value for the body mass index (BMI). We can expect that BMI to be proportional to one of the explanatory variables, weight_kg, as the general formula to calculate BMI is $mass/height^2$ where mass and height is measured kg and meters respectively. As such we have plotted it out in Figure 8(a), and we can see that BMI is highly linear-correlated with that attribute but there does exist some outliers. In regards to these outliers, we decided it was important to consider removing these observations as they can affect the parameters greatly if there is not a factor within the model to account for it. We inspected those observations and found that there was no similarities between the cluster of points in red among their demographic features, and there were many other observations that matched their features. As such, we can classify these observations as outliers and removed those red and blue observations.



(a) linear model summary

(b) stepwise model summary

Figure 6: Weight V. BMI

5.1.1 Lasso

For our first regression model, we are interested in performing least square regression with a lasso penalty on the non-OTU features as well as our scores from the 69 principle components. This will allow us to perform the general linear regression while considering model selection. We select the penalty term λ by performing 6-fold cross validation on the data set. Our result are presented in Figure 7, where we have plotted the root mean squared difference of all testing data for each lambda. As can be seen the graphs the RMSE term is stationary for $\lambda > .5$, which is where all parameters are present. However after some careful manual searching, we found that there exists a minimum for $\lambda \approx .002$. For this penalty, all non OTU factors are present, but out of the 70 score features only the second and sixth were not set to 0.

We further did a final test with this lambda value by training on a 5/6ths of our dataset and testing on the remaining portion. The result of the predicted v.s. actual BMI are plotted in figure 6, and we find that we get a training error of 2.78 and a testing error of 2.90. which are similar to our results from the lasso search.

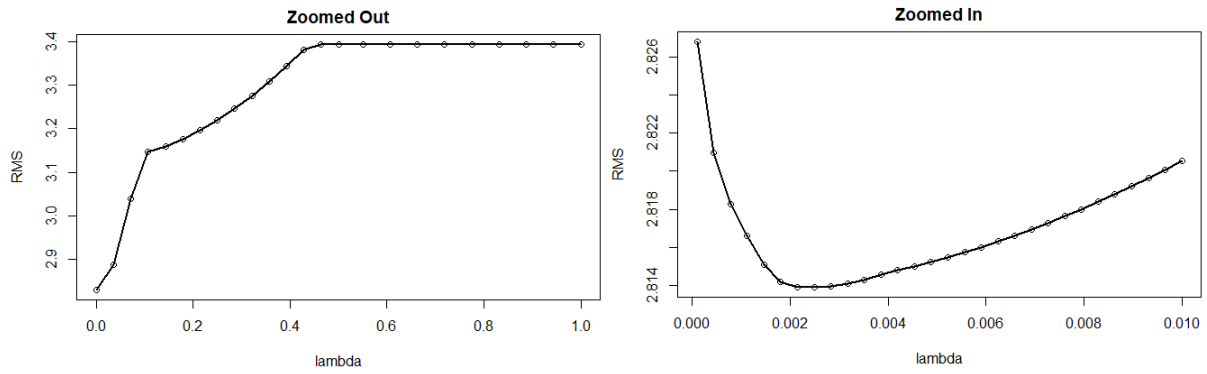
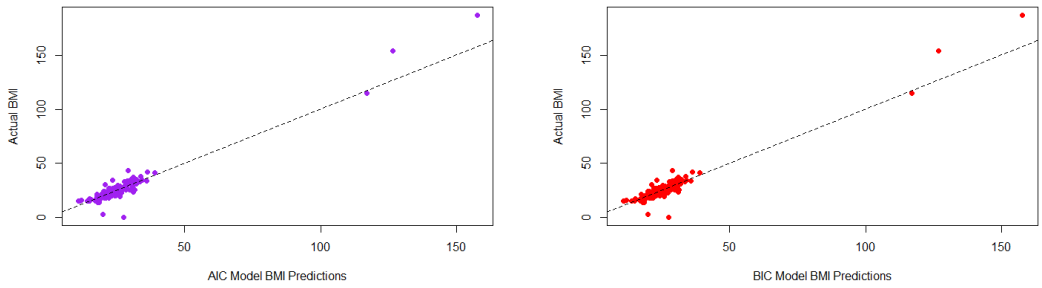


Figure 7: Lasso penalty

5.1.2 Model Selection

For our second task we also employed model selection using AIC and BIC criteria on our dataset but to switch things up, we decided to employ the clusters we obtained beforehand. Using the step function in R, we find that the optimal linear model from the AIC condition, was one that included all factors within the demographic features along with the hierarchical clustering, as well as the K=2 self organized map results. For the BIC criteria, we obtain a similar model except it did not select any of the the clustering results. This is not to surprising as BIC does favor smaller models. The testing errors for these models were found to be 2.893, and 2.891 for the AIC and BIC criteria respectively, which shows that neither of these models performed any better than the other or lasso we performed above. Additionally, the disadvantage of AIC is that it becomes inaccurate when the sample n increases, and the index that compensates for this is BIC. The calculated AIC and BIC look very similar, inferring that the sample size used in the analysis did not significantly influence our model selection.



(a) AIC Comparison

(b) BIC Comparison

Figure 8: Predictions vs. Expectation

5.2 Classification

5.2.1 Logistic regression for Alcohol Consumption Frequency

Multinomial logistic regression is a classification method to deal with multi-class problems. Here, '*alcohol frequency*' has five categories. Daily (864 obs), Never (1,863 obs), Occasionally (1,884 obs), Rarely (2,097 obs), and Regularly (1,651 obs). In this chapter, using 74 variables (3 categorical variables: *age*, *race*, and *sex*, and 71 numeric variables: *weight* and 70 projected-OTU variables), it will predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, '*alcohol frequency*'. 8,359 observations were randomly divided into two groups and used as training samples and testing samples.

In the analysis, *multinom* function in *nnet* package is specifically utilized. To make a clear interpretation, *Never* is set as a basic category. Due to the space limit, only the first five variable result are shown. Standard errors are presented within the parentheses.

Table 8: Multinomial logistic regression

	(Int')	**30s**	**40s**	**50s**	**60s**	**70+**	**child**	**teen**	**weight**	**race 1**
Daily	-11.45	0.84	1.16	1.96	2.00	2.24	-2.73	-2.27	0.01	7.15
	(0.39)	(0.44)	(0.43)	(0.42)	(0.42)	(0.45)	(1.11)	(1.09)	(0.00)	(0.38)
	race 2	**race 3**	**race 4**	**male**	**female**	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
Daily	8.41	-2.66	7.23	0.31	-5.94	0.12	1.70	0.57	-0.91	-0.51
	(0.24)	(0.00)	(0.50)	(0.14)	(0.00)	(0.57)	(0.83)	(0.98)	(1.24)	(1.25)

Note: race 1 (Asian or Pacific Islander), race 2 (Caucasian), race 3 (Hispanic), race 4 (Other), (Occ: Occasionally, Reg: Regularly)

Although it is well known that interpretation of the mulinomial logistic regression is a bit complex, here, we can give some intuitive explanation. Let's focus on the row, 'Daily'.
$\rightarrow logit(\text{“}Daily\text{”}) = \frac{P(Y=\text{“}Daily\text{”})}{P(Y=\text{“}Never\text{”})} = -11.45 + 0.84 \times age(30s) + \cdots + 0.01 \times weight + \cdots$

The equation above means that given other variables are fixed, if a person is his/her 30s, the odds that the person's alcohol frequency will change from "Never" to "Daily" increase by $exp(0.84) = 2.3$ times. Or, when the person's weight increases by 1kg, the odd increases by $exp(0.01) = 1.01$ times.

In addition, we can get a fitted value for each observation and set the value as cutoff points. It enables us to predict the classification of alcohol frequency. The table below shows that which category of alcohol frequency each observation is most likely to fall into. For instance, the 1018-th observation may be classified as "Rarely", and the 8005-th observation is probably classified as "Never".

Table 9: Multinomial logistic regression (Classification probability)

Obs id	**Never**	**Daily**	**Occ'**	**Rarely**	**Reg'**
1018	0.19	0.12	0.20	0.26	0.23
8005	1.00	0.00	0.00	0.00	0.00
4776	0.13	0.15	0.23	0.22	0.26

Using this method, a confusion table is presented as below, and accuracy is around 0.63.

Table 10: Multinomial logistic regression (Confusion table)

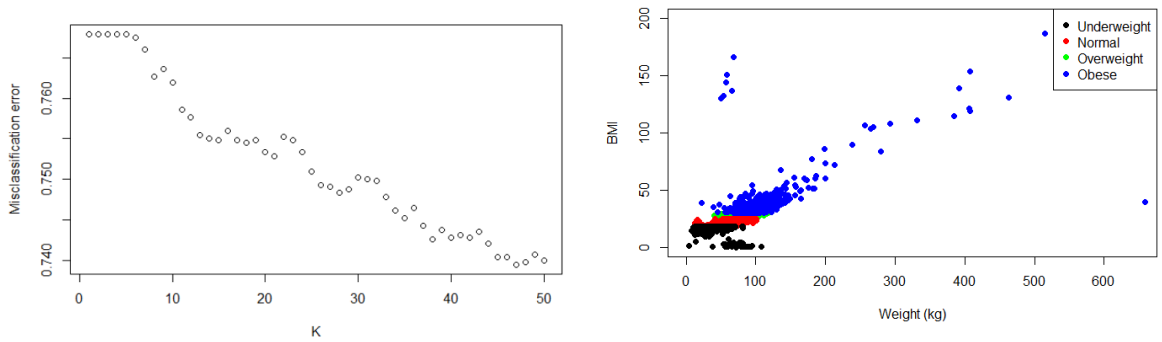
		Alcohol frequency (Prediction)				
		Daily	**Never**	**Occ'**	**Rarely**	**Reg'**
Alcohol frequency (Test)	**Daily**	87	24	72	128	128
	Never	37	305	163	261	143
	Occ'	29	54	384	324	161
	Rarely	59	69	262	484	195
	Reg'	43	31	191	278	268

5.2.2 KNN for Alcohol Consumption Frequency

The k-nearest neighbors algorithm (KNN) is a non-parametric method for classification as well as regression. The k closest training samples are used as input, and output is the assigned class. Similar to the classification using the logistic regression, the total observations were randomly divided into two groups, and one is used as training data and the other is used as testing data. The confusion table when K=10 is represented as below, and misclassification rate is around 0.76, which is higher than the analysis from the logistic regression. misclassification rates of K from 1 to 50 are also shown in the plot below. misclassification rate has the lowest value when K=47.

Table 11: KNN (Confusion table, K=10)

		Alcohol frequency (Prediction)				
		Daily	**Never**	**Occ'**	**Rarely**	**Reg'**
Alcohol frequency (Test)	**Daily**	42	62	67	92	85
	Never	88	257	191	193	157
	Occ'	97	180	229	249	195
	Rarely	109	245	262	297	204
	Reg'	103	165	203	238	170



(a) Accuracy based on K

(b) Weight Vs. BMI with BMI Cat labels

Figure 9

5.3 Gradient Boosting Tree for BMI category

The BMI Category response is expressed in four categories; Underweight, Normal, Overweight, and Obese. We decided to tackle this multi-classification task by using ensembled tree models, specifically gradient boosting tree and random forest. The response is BMI category as categorical variable, and the predictors we chose all the features except for BMI that are not OTUs, along with all the clustering labels from the last section.

For these models we used sex, race, weight, alcohol frequency, age as base predictors, with 5-fold cross-validation to train the model, and used accuracy as the metric. We then performed the same classification with hierarchical cluster labels as a feature to see if there exists any improvement. The accuracy of the models can be found in Table 12.

Table 12: Gradient boosting tree performance

model	accuracy
gbm	0.7851586
gbm with hc	0.7815679
gbm with hc and k-means	0.7983244

Table 13: Gradient boosting tree with hc confusion table

		BMI category (Reference)			
		Normal	**Obese**	**Overweight**	**Underweight**
BMI Category (Prediction)	**Normal**	872	7	114	45
	Obese	0	108	33	1
	Overweight	97	38	261	4
	Underweight	7	1	1	82

The confusion matrix in Table 13 shows that the highest misclassification rate occurs when trying to classify overweight category, with only 0.7878 accuracy while other categories all have more than 0.8 accuracy. We think there might be some information got lost during the data filtering. Some columns with more then 99% zeros might provide useful information for the overweight.

5.4 Random Forest for BMI category

We used random forest model with the same predictors and response as section 5.3. The model was trained based on the same training dataset using 5-fold cross validation. The evaluation was based on the same test set as gradient boosting tree model. The prediction result and model performance can be found at Table 14 and Table 15 below.

Table 14: Random forest tree performance

model	accuracy
random forest	0.7767804
random forest with hc	0.7863555
random forest with hc and k-means	0.7582286

Table 15: Random forest with hc confusion table

		BMI category (Reference)			
		Normal	**Obese**	**Overweight**	**Underweight**
BMI Category (Prediction)	**Normal**	870	7	135	46
	Obese	3	102	35	2
	Overweight	90	44	238	3
	Underweight	13	1	1	81

Same pattern happens with random forest models as shown in table 15, overweight got lowest accuracy when testing the model. As we can see here, k-means labels do not necessarily provide useful information for the classification, sometimes they can even lower the model performance. But hierarchical clusters still offer some predictive information, which can slightly improve the predicting accuracy.

Best model for this task is the same as the regression with AIC. using all the non-OTU variables combined with hierarchical clusters provides the highest accuracy.

6 Questions

The data is very noisy and based on our current analysis, there is very little signal. Hence the collaborator concerns about whether the findings are real. Below are some questions regarding the data properties and the feature analysis:

1. How to determine if there are taxa in this data set with correlated abundances. Since we were using a post-processed dataset, to determine the correlation, we might need more prior knowledge about this specific dataset. In a compositional approach, the variance between ratios of 2 taxa should be 0 or nearly so for the 2 taxa to be counted as correlated. The simplest approach is to transform the OTUs into log-ratio-transformed data. The question is whether or not the data had already been transformed using this method, if not, what method was applied on it?

2. How could the statement "compositional data sum up to 1" give us more information. Considering the dimension was too large, we dropped some columns before modeling, which basically broke the validity of this statement. But if we choose not to simply drop the columns with too much 0, how could this attribute help us to do further analysis, especially based on so many features?

3. How could we determine noises in OTU variables? We could easily detect any outliers or invalid values within variables like BMI and weight. But with so many OTU variables, how to guarantee those values are clean and ready for the models?

4. Is it feasible to simply drop columns with too many zeros? What kind of information are we losing? One concern is that maybe one OTU determines whether a person is overweight or not, if the data is imbalanced, that OTU variable might got dropped. But if so, to determine such variables faces computational challenge. How can deal with such problems or how can we exclude this idea?

5. In figure 8(a), we could see there are some data points (red ones) far beyond the main trend. What is the possible reason for those data points? In our case, we treat them as outliers and remove them for our regression tasks. But can we treat them as invalid/wrong values considering there was not only one data point following this pattern (unlike the blue one)?

6. As we can see, the predictive power of the hierarchical clustering is very low, even though it could slightly improve the classification performance. What is the true meaning for this clustering result? Where can we start to focus on to answer this question. It surely does not have much to do with all the existing categorical variables in the dataset, such as race/sex/bmi cat/age etc., but how can we further interpret this result in-depth?

7. In this research we did not use most of the libraries that were created and mentioned in the papers for microbiome data. It is mainly because the the property of the data is unknown after the previous processing before we got the data. And the tasks are different also. Some questions may be raised regarding those tools. Are there actually some tools/packages/open-source resources that we can use for this project? If so, what are the main benefits of those tools in terms of computational cost and methodology?

8. In our analysis, we found that weight is highly correlated with BMI, but with different intercepts. Our results could not capture the overweight pattern perfectly. One possible reason was the model bias being large in terms of the intercept. Some useful information might be dropped in OTUs. How can we capture those information afterwards. Because from our results, some categories already had good predicting performance without OTUs. Further research could consider a bias reduction model combining with the current model to improve the prediction results.

7 Conclusion

Many studies in general have a problem with the quality rather than the quantity of data, so there are many methodologies to deal with this. However, in the case of dealing with large data with tens of thousands of variables, as in this study, on the contrary, it is important to analyze the data without compromising the quality of the information. We have shown in this report the use of multiple supervised and unsupervised tasks on a large and sparse dataset. We reduced the dimension of the OTU features from around 33,000 feature to just 70 by selecting the features with lower sparsity, performing principle component analysis, and using the top components to describe the complex and large OTU feature set.

We utilized Self-Organized Maps, K-means, and Hierarchical Clustering methodologies, and although we found K-means and Self-Organized Maps to be consistent with one another, the Hierarchical results were proved to be the most useful one to our regression/classification tasks. We performed a number of supervised learning methodologies using the demographic data as well as our clustering results and the scores from the PCA projections. We found three models that can predict a person bmi with accuracy ≈ 2.9, which is unfortunately quite high considering the ranges of the BMI categories for normal and underweight are around 5. We also chose to use K-Nearest Neighbors and Logistic Regression to predict alcohol consumption and found the accuracies of these models less than 50%. Finally we used tree ensembled methods to predict BMI category and found models with different feature sets reached highest prediction accuracy more than 79%. All together we were able to perform these regression and classification tasks with a highly complex and large dataset and found decent models. We would expect that a more rigorous treatment with collaboration and direct aid from a microbiology researcher could enable us to form more complex models with better accurate predictions.

8 Appendix

All the code can be found at GitHub: https://github.com/ReuvenBirnbaum/STAT542Project

9 References

[1] American gut database, https://github.com/biocore/american-gut.

[2] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, 2015.

[3] Veronika B Dubinkina, Alexander V Tyakht, Vera Y Odintsova, Konstantin S Yarygin, Boris A Kovarsky, Alexander V Pavlenko, Dmitry S Ischenko, Anna S Popenko, Dmitry G Alexeev, Anastasiya Y Taraskina, et al. Links of gut microbiota composition with alcohol dependence syndrome and alcoholic liver disease. *Microbiome*, 5(1):141, 2017.

[4] Jolinda Pollock, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. The madness of microbiome: attempting to find consensus "best practice" for 16s microbiome studies. *Applied and environmental microbiology*, 84(7), 2018.

[5] Justine Debelius, Se Jin Song, Yoshiki Vazquez-Baeza, Zhenjiang Zech Xu, Antonio Gonzalez, and Rob Knight. Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome biology*, 17(1):217, 2016.