COMPARATIVE ANALYSIS OF METHODS FOR MICROBIOME STUDY

BY

MIHIR VISHWANATH IYER

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Adviser:

Professor Ravishankar K. Iyer

Abstract

Microbiome analysis is garnering much interest with benefits including improved treatment options, enhanced capabilities for personalized medicine, greater understanding of the human body, and contributions to ecological study. Data from these communities of bacteria, viruses, and fungi are feature rich, sparse, and have sample sizes not appreciably larger than the feature space, making analysis challenging and necessitating a coordinated approach utilizing multiple techniques alongside domain expertise. This thesis provides an overview and comparative analysis of these methods, with a case study on cirrhosis and hepatic encephalopathy demonstrating a selection of methods. Approaches are considered in a medically motivated context where relationships between microbes in the human body and diseases or conditions are of primary interest, with additional objectives being the identification of how microbes influence each other and how these influences relate to the diseases and conditions being studied. These analysis methods are partitioned into three categories: univariate statistical methods, classifier-based methods, and joint analysis methods. Univariate statistical methods provide results corresponding to how much a single variable or feature differs between groups in the data. Classifier-based approaches can be generalized as those where a classification model with microbe abundance as inputs and disease states as outputs is used, resulting in a predictive model which is then analyzed to learn about the data. The joint analysis category corresponds to techniques which specifically target relationships between microbes and compare those relationships among subpopulations within the data. Despite significant differences between these categories and the individual methods, each has strengths and weaknesses and plays an important role in microbiome analysis.

Acknowledgments

I thank my adviser, Professor Ravishankar K. Iyer, for constantly pushing me to grow and learn along with his guidance throughout the course of my research.

I also thank the entire Depend research group for creating an encouraging and team-oriented environment.

I thank the ECE department at the University of Illinois and the Champaign-Urbana community for six years of continuous learning both inside and outside the classroom as well as great friendships over the years.

I am grateful to my family all over the world for incredible support throughout my life which has helped me with every challenge I have faced.

Table of Contents

Chapter 1: Introduction 1
Chapter 2: Microbiome study background 4
2.1 Benefits of microbiome study 4
2.2 Data overview 8
Chapter 3: Univariate analysis 11
3.1 Univariate methods 12
3.2 Univariate methods in the context of microbiome study 13
Chapter 4: Classifier-based analysis 15
4.1 Considering classifier-based approaches 15
4.2 Classification methods 16
Chapter 5: Joint analysis 22
5.1 Considering joint analysis methods 22
5.2 Joint analysis methods 23
Chapter 6: Case study: Cirrhosis and hepatic encephalopathy 33
6.1 Background 33
6.2 Dataset characteristics 35
6.3 Methods 36
6.4 Results 39
6.5 Methods discussion 51
Chapter 7: Conclusion 53
References 55

Chapter 1: Introduction

The study of microbes inhabiting the human body represents a new frontier in medicine with anticipated benefits in personalized medicine, pulmonology, and obstetrics among others [1]–[3]. Multiple microbial communities are present in various parts of the body, with different areas containing their own microbiomes (e.g. intestine, lungs, skin, mouth). Despite not consisting of human cells, the microbes are essential for the proper function of the human body. A multitude of factors including diseases, diet, medication, and genetics can affect the compositions of the various microbiomes throughout a person [4], [5]. Numerous techniques have been employed to investigate the relationships between these factors and microbiome composition, ranging from simple diversity comparisons to metagenomic studies involving the genetic makeup of the microbes present [6]. Efforts utilizing a combination of univariate statistical methods, supervised and unsupervised learning techniques, and other multivariate or joint analysis approaches are necessary to tackle the challenges presented by microbiome data which is often feature rich and sparse with limited sample sizes [7].

The development of techniques and procedures for microbiome analysis presents a unique challenge with a wide variety of potential benefits, both short and long term. Including microbiome analysis as a standard component of disease study can allow for a more detailed understanding of the underlying mechanisms and potentially result in improvements to existing treatment protocols. Additionally, microbial information may provide more details on the long-term effects of current medications and therapies. Precision medicine appears to also be a promising beneficiary of improvements to microbiome study due to the personalized nature of

the microbiome. With improved collection and sequencing capabilities contributing to an increase in available microbiome data, it is more important than ever to understand and be informed about the methods available to analyze microbiomes and their connections to diseases. In conjunction with the aforementioned medical advances enabled by improved microbiome analysis techniques, other scientific fields involving microbiome study may benefit from such work.

Analysis methods used for microbiome study can be separated into three general categories: univariate statistical methods, classifier-based methods, and joint analysis methods focused on microbe-to-microbe relationships. Univariate statistical methods involve the comparison of distributions of a single variable or feature among two or more subpopulations to determine how much they differ. Classifier-based approaches involve models designed to predict disease state based on microbe abundance, which are then analyzed to identify patterns in the data and important microbes. Joint analysis methods are those which specifically target relationships between microbes and provide insight through the comparison of these relationships across disease states. Variations among these microbe-to-microbe connections can then be interpreted to provide insights about the data and any underlying patterns. Despite significant differences in complexity between these categories and the individual methods, each has strengths and weaknesses and plays an important role in microbiome analysis.

This thesis describes and compares methods useful for the study of microbiomes and is structured in three main parts: background pertaining to microbiome study, discussion of

analysis methods, and a case study illustrating the application of a selection of methods. The overview and discussion of analysis methods is partitioned into three chapters addressing the univariate, classifier-based, and joint analysis methods. After these chapters the case study is presented on the study of liver disease and accompanying brain condition, cirrhosis and hepatic encephalopathy.

Chapter 2: Microbiome study background

A human microbiome is a community or ecosystem of bacteria present in a specific part of the body. Microbiomes throughout the body are contributors to various functions including immune response, digestion, and brain function [8], [9]. These communities are constantly changing depending on the state of the host body (e.g. age, medical conditions) as well as environmental factors such as diet and medicine [9]. Microbiomes within a single person are not homogeneous as evidenced by findings that microbial communities from the same physical regions in different people are more similar than those from different regions within a single person [10]. Some commonly studied human microbiomes are the gut, mouth, genitals, skin, airway, placenta, and eye with most recent studies focusing on the gut [11]. This chapter addresses why microbiome study is important, basics of the initial steps for microbiome composition measurement, and key aspects of the data.

2.1 Benefits of microbiome study

The human body contains more microbes than it does human cells, with ratios ranging from 1.3:1 all the way to 2.3:1 [12]. These microbes are spread across various microbiomes throughout the body, including the digestive tract, skin, eyes, and lungs [11]. Although they may not genetically be part of the body, these microbiomes play a critical role in many functions and processes including ones related to illness and disease.

Expanding the study of diseases to include effects on the microbiome can enable a more thorough understanding of how certain diseases work, and eventually how treatments can be improved. Studies of the lung microbiome have suggested that certain microbes in the airway

may contribute to breathing difficulties by causing inflammation [13]. Huang suggests that this type of information allows for the development of new therapies that may treat asthma by altering the composition of related microbiomes. Similarly, the gut microbiome has been identified as a potential factor related to type 2 diabetes, with some changes to microbe abundance already being observed when existing treatment procedures are used [14]. One instance of this discussed by Brunkwall and Orho-Melander is the drug metformin which results in discernable alterations to the microbiome composition, related to both its desired therapeutic effects as well as adverse side effects [14]. Going further, changing the balance between microbes through alternative therapies shows some promise but requires a clear understanding of how the microbiome affects the body's normal functions, disease related functions, and other microbes.

In addition to the development of general therapies, a deeper understanding of the microbes present in humans can enable new techniques and treatments from the field of precision or personalized medicine. Even when considering a single disease or condition, it is generally understood that the same treatment may not be equally successful among all patients. One of the factors which appears to have an effect in determining which treatment options may work for certain patients is the composition of relevant microbiomes. Understanding how microbes interact with the immune system and respond to various therapies is necessary for accurate personalized medicine recommendations. For example, high levels of certain bacteria (*Ruminococcus obeum* and *Roseburia intestinalis*) correspond to patients who do not respond to certain cancer treatments, while the presence of other species (*Bifidobacterium longum*) relates to more positive outcomes [15]. Further delving into the

concepts of precision medicine, the constantly changing nature of microbiomes means that even for a single patient, analyzing the current state of their microbiota can allow for the determination of whether or not a particular treatment option will be likely to succeed at that time [16]. Kuntz and Gilbert assert that these insights may even be realized as patient-specific dosing that is dependent on metabolic processes supported by their microbiome as drug tolerance and effectiveness are some of the aspects which have already been linked to microbe levels [16].

Looking towards existing treatments, the long-term effects of current medicines are not always fully understood or known. Antibiotics are interesting to study from this perspective, as they are expected to have a direct impact on the viability of certain microbes. Studies have shown that the composition of the microbiome is impacted by antibiotics which may have been taken years prior, including increases in the observed levels of antibiotic-resistant genes [17]. Analyzing how the microbiome changes over time after certain treatments can help cultivate a more thorough understanding of how those treatments may alter the long-term health of the patient.

Complementing the treatment related benefits discussed above, improvements to collection and measurement techniques call for the development and expansion of microbiome analysis methods. One of the most commonly used methods is built around the sequencing of 16S genes and allows for the characterization of the entire microbiome from a sample [5]. As measurement techniques continue to evolve and develop, it is important to have analysis

methods which can cope with and comprehend the large number of microbes detected – sometimes up to five times the number of samples present in the data [7].

Lastly, methodologies designed for the analysis of human microbiomes from a medical perspective may be useful in additional fields where microbiomes are present. These additional areas of study include analyses on how the oceans are changing and what impacts we can expect from climate change. As an example, scientists are studying how different coral environments and their microbiomes respond to increases in the temperature and acidity of the oceans [18]. Soil analysis is another area of study where the microbiome is being looked at as a key source of information. Researchers have found that microbe communities in Alaskan soil have certain antibiotic resistance properties [19]. Others are working to analyze the microbiomes present in the atmosphere as well as animals [20], [21]. As techniques are developed to address the difficulties present in analyzing human microbiomes, the sharing of these techniques outside of the medical world can enable advancements in other areas of microbiome investigation.

As discussed in this section, there are many reasons to be excited about the evolution of microbiome analysis methods in different fields. A strong understanding of the human microbiome is critical to gain a more complete picture of how the body reacts to sickness and medicine. This expanded understanding will enable progress in the medical field from improvements in the understanding of diseases and existing treatments to the development of entirely new therapies. It also supports the field of personalized or precision medicine and

makes use of the vast amount of data becoming available through new collection and preliminary analysis methods.

2.2 Data overview

To facilitate analysis, samples are taken from patients using a variety of methods depending on the microbiome being studied and then processed to determine the composition of the microbial community. As discussed by Morgan and Huttenhower, there are multiple processing methods which support the identification of microbes grouped by taxonomy, known as operational taxonomical units (OTUs) [22]. The most commonly used methods to generate OTU-based abundance data from patient samples are based on sequencing of marker genes such as 16S rRNA [23]. These OTUs are considered the features in relative abundance datasets.

Once the OTUs have been identified, the microbiome composition data is structured into a relative abundance representation. Table 1 shows the format of a simple relative abundance table with sample identifiers in the first column, and subsequent columns containing relative abundance values for each OTU. The condensed table corresponds to a hypothetical dataset comprised of N OTUs and M samples. It is important to note that relative abundance values do not directly correspond to measured concentrations of the microbes. Instead they represent the proportion of the total microbes present in the sample which are a part of that taxonomical unit. Due to the values representing a proportion, the sum of relative abundance values from a single sample is one. To support more complex or disease-specific analysis, columns which contain clinical data or disease severity information may be added.

Table 1 Relative Abundance Data Format

	<OTU 1 Name>	…	**<OTU N Name>**
Sample 1	<Value>	…	<Value>
…	…	…	…
Sample M	<Value>	…	<Value>

One challenging aspect of the relative abundance data gathered from microbiomes is the large numbers of features relative to the sample size [7]. The numbers of samples and features (OTUs) for several human microbiome datasets are shown in Table 2. This multiplicity results in increased difficulty when analyzing the data as it becomes more likely for features to be unimportant or redundant [24].

Table 2 Sample and OTU Counts for Multiple Microbiome Datasets [7], [25]

Dataset	**Total Number of Samples**	**Number of OTUs**
Costello Body Habitats	622	2741
Costello Skin Sites	401	2227
Costello Subject	144	1592
Fierer Subject	101	565
Fierer Subject x Hand	101	565
Saboo Cirrhosis/HE	761	149

Although datasets such as those mentioned above contain many unique OTUs, the relative abundance tables often contain many zeros. Many of the OTUs identified appear only in a small subset of the samples [7], which compounds the previously discussed difficulties relating to the sample size and large feature space. This can be seen in Figure 1, generated from data used in the cirrhosis and hepatic encephalopathy case study.

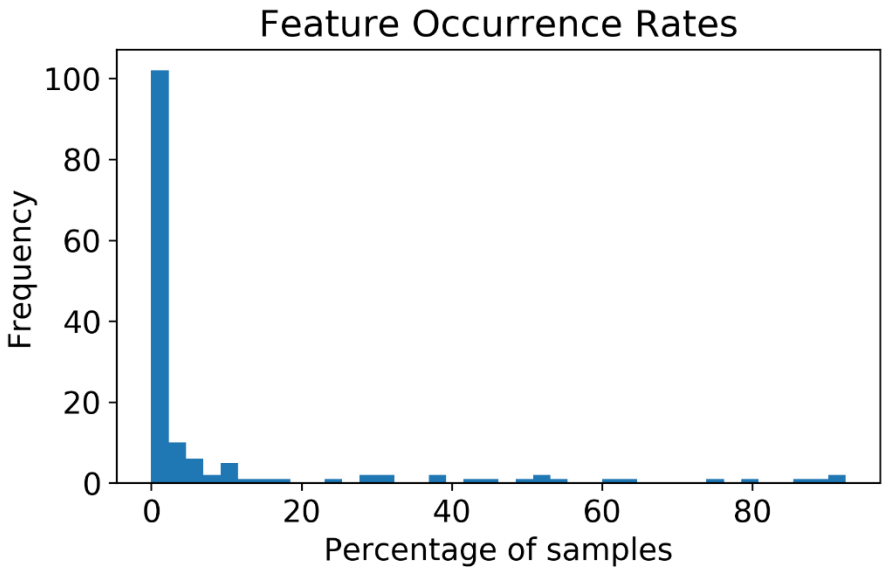

Figure 1 Histogram of nonzero occurrence rates among features

The nature of microbiomes as more than just a collection of organisms significantly complicates this field of study and how data is analyzed. In actuality, the microbiome must be viewed as an entire ecosystem which includes both bacteria and host cells [26]. To effectively study any ecosystem, it is critical to acknowledge the interactions or relationships between components of that ecosystem and how different organisms within the environment affect each other [27]. As different methods for microbiome data analysis are evaluated, these challenges must be taken into consideration when identifying the positive and negative aspects of each approach.

Chapter 3: Univariate analysis

The methods described in this chapter allow for the comparison of sample populations to determine the probability that they have been taken from the same underlying population. In other words, these tests determine if two or more sample populations are significantly different from each other. As univariate methods, they are designed to be used for the comparison of distributions in one variable. When considering commonly used univariate approaches, the two distinctions which appear are whether the test is parametric and whether it is limited to two sample populations. Table 3 shows how the four methods discussed in this chapter compare in these two categories.

Table 3 Comparison of Univariate Methods

	Parametric	**Non-parametric**
Two populations	T-test	Wilcoxon
More than two populations	Analysis of variance (ANOVA)	Kruskal-Wallis

The primary difference between parametric and non-parametric testing methods is that parametric tests require certain assumptions to be true about the distributions being studied. The parameters in a parametric test refer to the parameters of the population distribution, which implies the need to identify the distribution present before comparing samples. Although the selection of parameters adds complexity and requires assumptions, parametric tests tend to provide greater statistical power [28]. As a result, successful selection between parametric and non-parametric tests depends on a good understanding of the data and how well it satisfies the proposed assumptions.

3.1 Univariate methods

The t test is a two-sample parametric test used to determine if there is a significant difference between two populations. The primary assumption required for this test is that the distributions of the variable being compared are normal. Both directional and non-directional variants exist, and the choice between them depends on how much information is known about the variable and what type of testing hypothesis is desired. The generated t score statistic represents the distance between the two sample means in terms of standard deviation, with an additional factor to account for sample size. From this generated statistic a p value can be determined, which represents the probability of obtaining the observed data or something more extreme given that the null hypothesis and assumptions about population normality are true.

Analysis of variance, or ANOVA, testing expands parametric testing methods to studies with more than two sample distributions. Extending the normality assumption from the t test method, ANOVA's normality assumption is that the variable being compared is normally distributed in each sample being compared. The F statistic generated is the analog to the t statistic generated by the t test. When interpreting results from ANOVA testing, this computed F statistic is compared to a critical F value corresponding to a preselected alpha value. Improvements and extensions have been proposed to ANOVA which allow it to be used more generally and address distributions beyond just the normal [29].

The non-parametric Wilcoxon test is used to compare two sample populations like the t test, but without the need for assumptions such as normal distributions within each sample.

This allows it to be used more broadly and even in cases where data is not numerical, as long as ordering of the data is still possible [28]. Similarly to the t test, a p value is generated which can then be used to determine if there is a significant difference between the two sample distributions. There are however some concerns with this method, particularly in terms of sensitivity to changes in variance and skew of the data [30]. Even with these concerns, the Wilcoxon test does provide results that compare favorably with other univariate methods in some situations such as data with large errors [31].

The Kruskal-Wallis test extends the Wilcoxon test in much the same way as ANOVA methods extend t testing to multiple variables. Using the same core concept of focusing on rank within the total set of data, this test avoids the assumption that the target variable is normally distributed. One important limitation of the Kruskal-Wallis method is that it detects differences based on the center of the distributions, and does not discern differences in spread or shape very well [32]. Despite these drawbacks, its ability to detect shifts between sample distributions combined with the lack of assumptions make it useful [33], [34].

3.2 Univariate methods in the context of microbiome study

Due to their ease of use and straightforward concepts, univariate methods are used extensively in a number of fields. Often they are the starting point for analysis, as determining if differences or changes to a variable are significant between groups allows for more informed selection of additional analysis or experiments to perform. The closely linked area of genetic studies utilizes them for both preliminary and central analysis [35], [36]. Many microbiome studies rely on these techniques as a core part of data analysis, with a large focus on

determining whether significant differences in the abundance of certain microbes or groups of microbes are present when comparing healthy patients with those who have medical conditions [11], [37], [38].

Despite their extensive use as a part of microbiome study, there are several limitations which qualify their use in this field. The primary concern with univariate statistical methods is that they do not provide much information on interactions between microbes. Similarly, the combined effects of changes to groups of microbes may be missed unless efforts are taken to combine these methods with multivariate techniques. There are also concerns about compatibility between common univariate methods and relative abundance data due to independence assumptions which are violated due to relative abundance values from one sample being normalized to a sum of one [37]. Additionally, the need for statistical testing to be performed across more than one variable introduces issues related to multiple testing, requiring adjustments to significance and error values [39]. Univariate methods are useful for microbiome study, but it is important to consider them alongside more complex methods to ensure that statistical results reflect the full picture and that required assumptions are being satisfied.

Chapter 4: Classifier-based analysis

4.1 Considering classifier-based approaches

As the primary goal when studying human microbiomes is to identify key aspects of the different features in the data, it may seem strange to consider classifiers which are usually used for the purpose of assigning labels or classes to new data based on known information from training data. Although this standard application of classifiers results in a predictive model which does not directly support the analysis of microbes within an ecosystem, interpreting it as a descriptive model allows for a more sophisticated understanding of the microbiome. By analyzing the model implicitly generated by the classifier, scientists can identify connections between the classes (e.g. disease, gender) being assigned and specific features (microbes or OTUs) from the dataset.

Classifiers also may be able to directly address some of the challenges identified earlier with respect to number of features and the sparse nature of microbiome relative abundance data. Utilizing classification methods that provide information on which features have a greater impact on the final label determination enables comparisons between these microbes and identification of the most interesting microbes for further study. Even without specific details about the precise influence each feature or microbe may have, providing a guideline of which aspects of the microbiome to focus on allows for more informed experiment design.

An additional attribute of classifiers which enables deeper insights when applied to microbiome study is that many classifiers allow for some degree of interaction between features. This means that the characteristics of certain microbe-to-microbe relationships may

be encoded in the classification model. While it may not always be possible to extract these characteristics explicitly, any feature importance information that is generated will implicitly consider the relationships between microbes mentioned earlier.

Classifiers have proved useful both within and outside of the microbiome analysis field in the past, which suggests that they are a viable category of methods that should be considered [7]. Within the area of microbiome study, some classifier-based analysis methods have been integrated as a part of the QIIME 2 platform to make the process more streamlined [40]. Looking beyond the scope of microbiomes, they have also been used with microarray data on gene expression levels [41], [42]. The following section describes and assesses popular classification techniques in the context of microbiome study.

4.2 Classification methods

The techniques discussed in this section utilize a supervised learning approach where training data – with class labels – is used to generate a model. The model then predicts the class of unlabeled test data. These methods are considered supervised as they rely on the use of labeled training data as opposed to unsupervised approaches, such as clustering, where labels are not used. An overview of the technique as well as a discussion of how well the method addresses the challenges specific to microbiome analysis are provided for each classifier. Although not every classification approach can be discussed here, the selected ones are commonly used within the data science community.

4.2.1 K nearest neighbor

The k nearest neighbor (KNN) classifier assigns labels to test data based on the labels of the k nearest members of the training dataset. As such, it does not generate a model to represent the system being analyzed – the entirety of the test data can be considered the model. A visual example of the KNN classifier is shown in Figure 2 with colors representing the classes and circles for each point in the training data. The background color corresponds to the predicted class for test data at that location. Although the overall concept is straightforward, complications arise when attempting to select the optimal value for the k parameter [43].

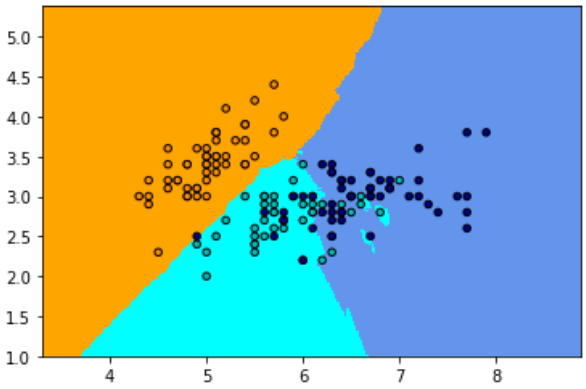

Figure 2 Example of k nearest neighbor classifier

To address the difficulty in determining the ideal value for k, several methods have been studied. One proposed approach is to perform the classification task multiple times with varying k values, and then combine the results to determine a final predicted label [44]. Taking a different approach, the concept of weighting the contributions from each neighbor based on distance has been introduced to reduce the sensitivity of KNN to the selected k value [45]. Attempts to select the k value based on the test data values have also shown improved classification accuracy when compared to the standard approach [46].

The k nearest neighbor method has seen significant success across a variety of fields and is still evolving. Although it often provides a high level of classification accuracy, this method requires a significant amount of care when selecting a k value and distance metric. The large number of zeros present in microbiome data, as discussed in section 2.3, may result in mismatches unless additional feature selection methods are used in conjunction with this approach and the distance metric is selected carefully. Additionally, the lack of a distinct model being generated limits the amount of information that can be extracted about how microbes interact with each other and the health conditions being studied. Even with these limitations, a KNN classifier may be useful as a preliminary step to identify high-level patterns and provide an accuracy benchmark for other classification methods used in microbiome analysis.

4.2.2 Support vector machines

Support vector machines (SVMs) are similar to the k nearest neighbor approach in that the end goal is a division of the feature space into regions corresponding to the classification labels. This is accomplished by attempting to find a hyperplane which separates the training data by class. As real-world data is often not separable, soft margin methods – which penalize but still allow for some misclassification of training data – are used [47]. One key difference between the SVM and KNN techniques is that SVM uses an analytical process to identify smooth boundaries between these regions while KNN does not perform analysis beyond inspecting the neighbors. This difference can be seen when comparing Figure 3 with Figure 2 in terms of edge smoothness and isolated pockets within larger regions. Although the original SVM method is designed for binary classification tasks, it has been extended to work for

multiclass problems via multiple binary classifiers with each class having its own in/out classifier [48], [49].

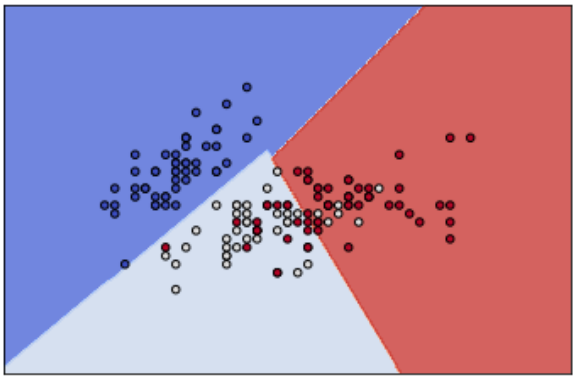

Figure 3 Example of support vector machine

The main decision to be made when using an SVM is kernel choice. Commonly used kernels include linear, polynomial, and radial basis function although other options are sometimes employed [50]. The kernel choice problem is well studied with more automated methods being developed. Utilizing information about the dataset characteristics allows for kernels to be selected on a case by case basis without trial and error, as different kernel functions have been shown to be optimal for different datasets [50]. This process has been enhanced by the combination of kernels, resulting in the SVM method becoming increasingly flexible to different data patterns [51].

When considering SVMs for the task of microbiome analysis, one thing which stands out is the lack of feature selection in the process. Given that unimportant features can inhibit SVM performance, the optimal way to use one may be to pair it with a separate feature reduction method [7]. In the related field of microarray studies, SVMs with automatic kernel selection

have generated encouraging results which suggests that they may be useful for microbiome study [52].

4.2.3 Random forest classifier

Random forest classifiers fall into the category of ensemble methods, where multiple classification models are created and then combined to generate a final model. In this case, the independent classification models are trees which represent a series of decisions based on different thresholds in the features. When implemented normally, the feature and value used to split at each level of the tree are selected using a greedy approach where the most discerning splits are used. When modifying this structure to generate the individual trees in a random forest classifier, the features available for the decision at each level are a randomly selected subset of all the features in the dataset. This allows for different trees to be generated, making it more likely that the model considers features which are sorted out by the greedy regression tree approach.

When assessing the benefits and drawbacks of random forest classifiers in the context of microbiome analysis, it is useful to see how they have been used in adjacent fields. In the field of gene expression study using microarray data, the random forest approach generated similar prediction accuracy and improved feature reduction when evaluated against comparable methods [53]. Two major benefits of the random forest approach when applied to microbiome data are that it allows for multiple features to jointly affect classification and produces feature importance results [54]. Feature importance provides insight into what is otherwise a complicated multi-tree model as it allows for the ranking of features based on their

importance for successful classification. Although these values do not directly represent statistical information like the results from principal component analysis might, they still allow for comparisons between features. As random forests can identify important features and implicitly handle relationships between these features, they are well-suited to the challenges and goals of microbiome analysis.

Chapter 5: Joint analysis

5.1 Considering joint analysis methods

Despite the merits of the univariate and classification-based approaches discussed in the previous two chapters, they do not address some of the challenges identified in the study of microbiome data. In particular, these methods do not provide concrete information on how different microbes interact with and affect one another. This limitation is especially problematic as it has been shown that microbes have very significant effects on each other which can in turn affect the humans they inhabit [26]. Although some of the classification-based approaches allow for features to influence the ways in which other features impact the analysis results, they do not focus on comparing these feature to feature relationships. While the classification and univariate approaches may provide some information on which features are most important or have significant changes between disease groups, not knowing how these features are connected within the microbiome limits the degree of understanding that can be achieved. Gathering information on the connections between features in the data and comparing this information across multiple subpopulations can provide additional insight on how those groups differ and allow for analyses to utilize relationships between microbes.

These concerns can be tackled using joint analysis techniques which focus on interactions between features and provide quantifiable information on these interactions as a part of their results. By viewing the microbiome as a community instead of a collection of organisms which operate independently, researchers in the medical field will have an additional layer of information which can be utilized to design additional studies and experiments. As

microbiome composition plays an important role in the proper function of many parts of the human body, understanding the processes by which these compositions change in connection with different medical conditions and symptoms allows for a more accurate understanding of the underlying disease mechanisms. Relationships between microbes which may be uncovered through the employment of joint analysis techniques include varying degrees of symbiotic and competitive interactions.

Going one step further, certain joint analysis approaches can assist with the identification of joint probability distributions connecting microbe relationships and interactions with disease states in a probabilistic manner. As a result, doctors and researchers may be able to shift away from threshold-based techniques such as searching for elevated or depressed levels of certain key microbes and towards the recognition of patterns or abnormalities among multiple microbes viewed together. The consideration of joint distributions may also assist in the identification of patterns corresponding to groups of patients with distinct underlying characteristics who may have contrasting outcomes from the same treatments. Although not every joint analysis approach leads to a joint probability distribution, they do answer many of the same questions about the data and support progress toward the goal of identifying the joint probability distribution.

5.2 Joint analysis methods

The techniques discussed in this section are centered around understanding how different features within the data influence each other and jointly affect the medical conditions being studied. When considering the approaches described below it is important to consider

what assumptions or requirements are present with each method and how they relate to the challenges described in section 2.3. The methods also provide different types of information, so selecting the most appropriate procedure for a particular investigation or analysis necessitates a clear understanding of both the desired information as well as what results can be obtained using each technique.

5.2.1 Correlation studies

The most basic correlation studies are those which compute the correlation coefficients for pairs of features in the dataset. The primary goal of simple two-feature correlation analysis is to extract information about the relationship between each pair of features, with a positive correlation coefficient indicating that when the value of one feature increases the other feature also tends to increase. In contrast, a negative correlation coefficient indicates that as the value of one feature increases the value of the other feature decreases. The absolute value indicates the strength of this association, with a zero correlation coefficient corresponding to a pair of features with no association. Although they are often considered together, both linear and non-linear correlation coefficients can be used depending on the data characteristics.

When considering linear correlation measures, the Pearson correlation coefficient is a commonly used approach. The results obtained from linear methods such as the Pearson coefficient must be interpreted carefully to avoid coming to unsupported conclusions. More specifically, a low linear correlation does not guarantee that two features are not correlated as these approaches only analyze the data for linear associations [55]. Thus, it makes the most

sense to use this method alongside other non-linear approaches to ensure that important connections between features are not being missed.

Among non-linear correlation measures, the spearman coefficient is often used. As it is a rank-based method, a linear relationship between feature values is not required for strongly correlated features to be identified. By transforming the data from values which may not be linearly correlated into ranks which will always be linearly correlated if the variables are correlated, this method allows for linear correlation measures to then be used on the transformed data to determine non-linear correlation [56]. A similar approach can be taken which results in a subset of correlation methods based on nonlinear data transformations. These methods involve the transformation of nonlinear data into a space where the relationships are more linear, and then applying the Pearson linear correlation method discussed earlier [57]. As a whole, nonlinear methods allow for a broader application of correlation studies but require additional care when determining the optimal transformations or methods for a particular dataset.

Although the correlation analysis approaches discussed up to this point have been utilized in a variety of fields, they are constrained by the limitation of only being able to compare two features at a time. To allow for correlation studies to provide an understanding of the data based on multiple variables, multiple correlation techniques can be used. The general approach of this method is to perform a regression analysis for each feature in the data with the individual feature as the dependent variable and the remaining features as the independent variables, and compute the correlation between the predicted and actual values of that feature

[58]. From a statistical perspective, the computed multiple correlation corresponds to the proportion of the predicted feature's variance which is explained by the other features [59]. Despite the restriction of having to consider each variable once at a time, multiple correlation allows for the determination of which features are most dependent on or independent from the other features present.

Correlation methods are specifically designed to analyze the relationships between different features in a dataset, but they are limited in the type of information they can provide and require separate analyses for each feature being considered. In the context of microbiome study, the sum-to-one nature of relative abundance data complicates the usage of correlation-based methods as it results in a slight negative bias of the computed correlation coefficients. Even with these concerns and difficulty generalizing results to a joint distribution, correlation techniques may still prove to be beneficial when utilized in conjunction with other analysis methods.

5.2.2 Principal component analysis

Principal component analysis (PCA) is often considered as solely a feature reduction method; however, the compositions of the resulting principal components provide key insights into how the features in the data relate to each other. Studying the variable loadings computed by PCA allows for the identification of features which appear to vary together as well as how the scales of those variations compare. The computed loadings also facilitate enhanced interpretation of subsequent analyses on the dimensionally reduced data. This section

describes and discusses three variants of principal component analysis: the commonly used basic PCA, sparse PCA, and kernel-based PCA.

Basic principal component analysis generates principal components which are linear combinations of the features present in the data. Each principal component is created to explain as much of the remaining variance in the data as possible. This combined with the characteristic of all principal components being uncorrelated with each other allows for PCA to encode information about relationships between the features [60]. Moreover, the principal components are organized in descending order based on the amount of variance explained, which allows for their use as a dimensionality reduction tool. One key drawback of this approach is that principal components usually have nonzero coefficients for most of the variables, making it difficult to relate analysis of the principal components back to individual features in the data.

Methods with varying complexity have been used to limit the number of nonzero coefficients and thus perform implicit feature reduction in addition to the dimensionality reduction already expected from PCA [61], [62]. The added benefit of feature reduction via zero coefficients in this context is that the principal components become easier to interpret and connect to features in the data. One proposed method which accomplishes this task is sparse principal component analysis, where PCA is formatted as a regression problem and a lasso or elastic net is applied to generate a solution with a reduced number of nonzero coefficients [60]. In effect, nonzero values in the principal components are penalized to push the solution toward a sparse coefficient vector. When analyzing the results of sparse PCA, features which contribute

more to the variance can be easily identified by their nonzero coefficients and are candidates for inclusion in a reduced feature set. Additionally, differences between principal components in terms of the features used allow for subsequent analysis to be more directly connected to the original features.

One major limitation of the PCA and sparse PCA methods discussed above is that they are restricted to the computed principal components being linear combinations of the features. Kernel PCA has been developed to allow for the use of principal component analysis in situations where these relationships may not be linear. This method can be considered an extension of standard linear PCA methods via the addition of a transformation step. Kernels which map the original data to a new feature space are used before applying PCA methods to allow for the linearization of nonlinear relationships [63]. Thus, the choice of kernel determines what types of nonlinear relationships are considered. This method of extending a linear analysis method through the use of kernels for transformation before performing the standard approach can also be seen in the area of correlation studies as discussed in section 5.2.1. Commonly used transformations include both Gaussian and polynomial kernels, with de-noising being an area where kernel PCA outperforms other methods [64], [65].

As a whole, principal component analysis is a powerful tool for both dimensionality reduction as well as the identification of influential features within a dataset. The sparse and kernel PCA techniques may be particularly useful when addressing the challenges of irrelevant features and complex relationships common in the context of relative abundance data collected from microbiomes. One major drawback however is the sensitivity of PCA-based methods to

outliers, although less sensitive techniques such as robust PCA have been introduced and used with success on outlier-prone biological data [66]. Despite this challenge, PCA has been used successfully in multiple studies within the field of microbiome analysis [67], [68]. As it does not always provide a complete picture of every aspect within the data and requires careful interpretation (especially for nonlinear analysis), it makes sense to utilize PCA's dimensionality reduction and feature selection capabilities alongside other methods when analyzing microbiome data.

5.2.3 Multivariate kernel density estimation

Kernel density estimation enables the estimation of an underlying distribution based on a set of samples. In contrast to univariate kernel density estimation which estimates the distribution of a single random variable, multivariate kernel density estimation generates an estimated joint probability distribution [69]. This is accomplished through the summation of multiple smoothed kernels with locations specified by the sample data points [70].

The primary concerns with kernel density estimation are centered around the selection of appropriate kernels and kernel bandwidth values for the data [71]. Bandwidth selection is often based on targeting the minimization of mean integrated square error through both single-step and iterative approaches, although alternate error calculations are also used [72]. Despite kernel choice being an aspect which must be considered, it has been shown that bandwidth selection has a much greater effect on the quality of the distribution estimate with larger than desired bandwidths resulting in oversimplification of the structure and bandwidths which are too small highlighting features that may not actually exist [73].

As an evolution of histograms, kernel density estimation allows for better understanding and visualization of a sample distribution without the potential for misleading representations resulting from variation in bin sizing and positioning. Kernel density estimation can be used to expose structural patterns in the data, leading to an enhanced understanding of how features relate to one another [74]. The true multivariate nature of this approach means that information pertaining to the underlying joint distribution of the data can be extracted, as opposed to methods limited to the analysis of two features at a time. Kernel density estimation may also be utilized for classification-based analysis through integration with existing techniques including Bayesian networks and decision trees [75]–[77]. Changes to the structure of the joint distributions can be observed through comparisons between the estimated distributions of different subpopulations within the data, allowing for the identification of changes in the interactions between microbes or other features.

5.2.4 Graph-based approaches

As graph-based analysis techniques allow for the inclusion of information connecting the features present in a dataset, they are prime candidates for utilization as multivariate or joint analysis methods for microbiome data. One key limitation of graph-based methods is that determination of the graph structure often relies on domain knowledge in some way, although graph estimation models are an area of ongoing research [78], [79]. Of the many graph topologies and analysis methods available, this section describes two which have been employed for microbiome analysis: a network-based model and a split graph model.

The network-based graph model as described in Naqvi et al. is comprised of nodes corresponding to each OTU within the dataset [80]. These nodes are then connected with edge weights assigned based on correlations between the features. Individual graphs are created for each subpopulation being studied, with the subsequent analysis focusing on similarities and differences between the generated graphs. Through modification of the edge weight computations such as using the frequency with which two microbes co-occur in nonzero abundance, it may also be possible to orient the graph towards the detection of patterns specific to the data or problem being studied.

Kim et al. identify the split graph model as a way to represent both interactions between microbes as well as the effects of external factors on the microbiome [81]. The graph is comprised of one group of nodes corresponding to all the microbes detected with a second group of nodes representing external factors. As some degree of interaction between the microbiomes in an ecosystem is expected, the first group (with microbe nodes) is fully connected with each edge representing microbe-to-microbe interactions or influence. Connections between the two groups represent the effects of external factors on microbes within the microbiome. The assumption made at this stage is that each external factor only has a direct interaction with a subset of the microbes, thus domain knowledge must be used to determine the appropriate edges between external factor nodes and microbe nodes. The example presented by Kim et al. uses bacterial metabolic pathways encoded as KEGG orthologs for the external factors. Thus, the example uses microbe-pathway relationships contained within the KEGG dataset as the domain knowledge behind the split graph topography [82].

Once graphs for each subpopulation have been created, their structures can be compared by analyzing differences in edge weights and identifying highly connected subgroups of nodes.

As a whole, graph-based methods incorporate microbe-to-microbe relationships very explicitly. By generating graphs individually for each subpopulation in the data and comparing them, structural differences and closely connected feature groups can be identified for additional study. Although it requires more extensive domain knowledge than other techniques, graph-based analysis has the potential to reveal useful patterns in the relationships between microbes.

Chapter 6: Case study: Cirrhosis and hepatic encephalopathy

This chapter demonstrates how some of the microbiome analysis techniques described in previous chapters can be used together to study a specific condition. A brief overview of cirrhosis and hepatic encephalopathy (HE) is provided, followed by details on the methods selected for this problem. Result are then presented, along with discussion and comparison of the methods used.

6.1 Background

Liver cirrhosis is a condition in which a patient's liver is damaged and accounts for over one million deaths each year [83]. This takes the form of scar tissue being present throughout the liver and can result in reduction in liver function, portal hypertension, and liver cancer [84]. The damage is permanent and if severe enough can make a transplant the only treatment option available although there are therapies used to slow disease progression. Common causes of liver cirrhosis include alcoholism, hepatitis C, and hepatitis B [84].

Along with damage to the liver, some cirrhosis patients have an accumulation of toxins in the brain known as hepatic encephalopathy (referred to as HE) which negatively affects brain function but is reversible [85]. The widely accepted connection between cirrhosis and HE is that limited liver function allows certain compounds from the gut to enter the circulatory system, by which they travel to the brain and impair its functionality [86]. Estimates suggest that 30-45% of patients with cirrhosis develop hepatic encephalopathy [87]. Current treatments for HE include the medications lactulose and rifaximin with some patients undergoing surgery to add a shunt to route blood from abdominal organs around the liver [88], [89].

Although cirrhosis and hepatic encephalopathy primarily affect the liver and brain respectively, changes to the gut microbiome have been observed along with corresponding inflammation [90]. As the composition of the gut microbiome changes, conditions such as obesity, diabetes, and cardiovascular disease may occur [91]. With regard to HE specifically, the gut-brain axis has been identified as a potentially important factor as it involves communication and influence between the gut microbiome and the brain [92]. As both cirrhosis and hepatic encephalopathy have shown relation to the gut microbiome, analyzing and understanding changes to the microbiome's composition can contribute to the development of new treatment methods and the improvement of existing ones.

Recent studies have supported this connection between cirrhosis, HE, and the gut microbiome by identifying specific microbes which were observed to have either increased or reduced abundance [93]. These findings are supported by the successful use of certain antibiotics such as Rifaximin in hepatic encephalopathy treatment. Additionally, the gut-brain axis has been studied in the context of other diseases such as schizophrenia. Despite it being thought of as primarily a brain condition, certain bacteria have been identified that have a relation to the severity of a patient's schizophrenia symptoms [94]. Fecal transplantation has been identified as a possible method to positively alter the gut microbiome; however, it is critical that the microbiome's role in the condition being treated is well understood [95].

The goal of this study is to identify microbes that are important in the context of cirrhosis and hepatic encephalopathy. Additionally, a comparison between male and female patients is desired as there are known to be differences in microbiome composition between

men and women [96]. This comparison includes identifying which microbes change the most as well as differences in how specific microbes change for each gender. A secondary goal is to detect groups of microbes which may work together or influence each other.

6.2 Dataset characteristics

The data used in this study is from patients at Virginia Commonwealth University and McGuire VA Medical Center. Patients with infections, non-rifaximin antibiotic use, or probiotic use within the six weeks of the study were excluded, as were illegal drug users and those with alcohol use disorder, primary biliary cholangitis, autoimmune hepatitis, or primary sclerosing cholangitis [25].

The data is derived from stool samples taken from patients categorized into four groups: healthy controls, cirrhosis without HE, HE being treated with Lactulose, and HE being treated with both Lactulose and Rifaximin. The Lactulose and Rifaximin categories were selected as Rifaximin is prescribed to patients with more severe cases of HE, meaning that the groups can be thought of as controls, cirrhosis without HE, HE, and more severe HE. After initial processing as described in section 2.3, the relative abundance data was generated. The dataset contains 761 samples and 149 features, with each feature corresponding to an operational taxonomical unit (OTU) at the family level. This data is used for the analysis discussed in the remainder of this chapter. Each OTU has a four-part taxonomic identification name containing the phylum, class, order, and family, but they are referred to in this analysis by the family identifier.

6.3 Methods

There are three distinct portions of this analysis: a correlation study to search for relationships between features, a classification-based approach combined with statistical testing to determine the most interesting or important features from the data, and a principal component analysis incorporated with kernel density estimation to identify patterns in how the microbiome changes with disease severity. Each of the methods utilized includes separated analysis for men and women to allow for the identification of gender-specific patterns and differences. Not all approaches discussed in prior chapters are used, but techniques from each of the three categories play a role.

Before applying any of the aforementioned analysis methods, an initial feature selection step was performed to remove features which are not likely to provide useful information and may obscure the results. There are a number of such features due to the zero-inflated nature of microbiome abundance data, with many microbes having zero abundance values for a large proportion of the samples. The threshold used was 10 percent, meaning that only microbes which had non-zero abundance in at least 10 percent of men or 10 percent of women were included. In order to ensure that microbes responsible for key differences between genders were not being eliminated, only those below the 10 percent threshold in both gender subpopulations were removed. All analysis after this point utilizes only the features selected for inclusion through this process. Figure 4 illustrates the decision process for feature inclusion and exclusion.

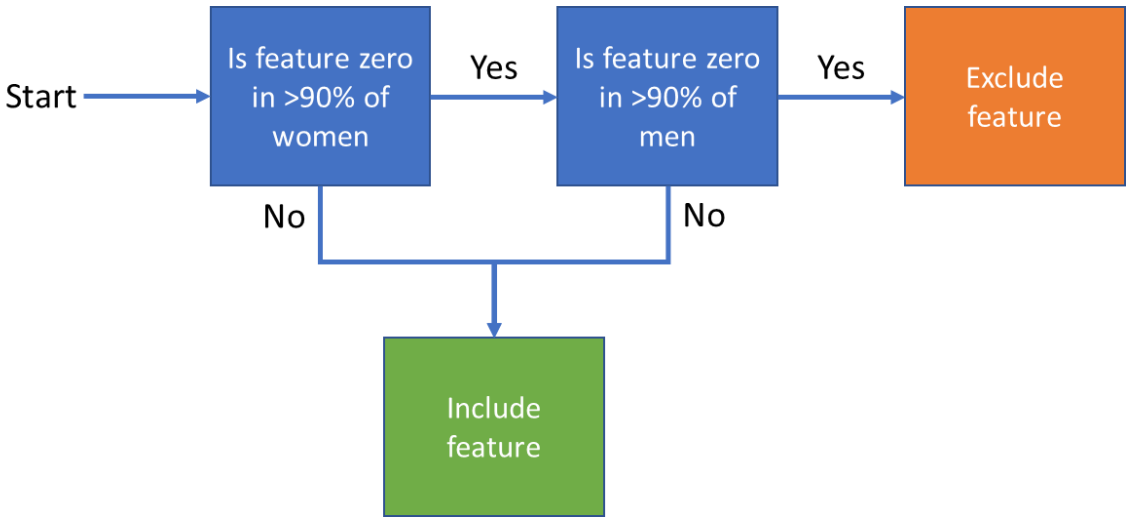

Figure 4 Feature inclusion/exclusion decision diagram

The first analysis technique applied is a correlation study across the features in the dataset. The correlation coefficient for each pair of features is computed using both the Pearson and Spearman rank methods to allow for the identification of some nonlinear correlation. This process is repeated on the male and female subpopulations separately to determine if there are correlation patterns or groupings unique to either gender.

To identify the most important microbes present in the samples, a classification-based approach is taken to allow for the effects of multiple microbes to be considered. The overall classification goal is to predict, based on the microbiome composition data, which of the four groups identified in section 6.2 a patient belongs to. This problem is separated into the nine binary classification tasks shown in Table 4 with class 1 being the group with a less severe disease state.

Table 4 Classification Tasks

Task Name	Class 1	Class 2
Control vs Cirrhosis	Controls	All cirrhosis patients
Control vs Cirrhosis no HE	Controls	Cirrhosis patients without HE
Control vs HE	Controls	HE patients
Control vs HE-Lac	Controls	HE patients taking Lactulose
Control vs HE-Rif	Controls	HE patients taking Rifaximin
Cirrhosis no HE vs HE	Cirrhosis patients without HE	HE patients
Cirrhosis no HE vs HE-Lac	Cirrhosis patients without HE	HE patients taking Lactulose
Cirrhosis no HE vs HE-Rif	Cirrhosis patients without HE	HE patients taking Rifaximin
HE-Lac vs HE-Rif	HE patients taking Lactulose	HE patients taking Rifaximin

Random forest classifiers, as described in section 4.2.3, are used to complete these tasks with a focus on the generated feature importance values. The microbes within each task with the greatest importance values correspond to those which are most useful in discerning between the classes. The random forest classifiers used for this step contain 21 trees with a maximum depth of 4. An 80-20 split of training and testing data is made with 30 iterations of Monte Carlo cross-validation. For each classification task and classifier, area under the curve (AUC) is calculated. These AUC values allow for comparisons to be made between classification tasks to determine which groups are most discernable based on gut microbiome composition. Once the most influential or interesting microbes are detected, Wilcoxon testing is applied to determine if there is a significant difference in the abundance of each of these microbes between disease classes. The originally unsigned feature importance values for each microbe are given signs based on whether the overall relative abundance is increasing or decreasing when going from the less severe to more severe disease class in the task. All of these random forest classification analyses are performed separately for men and women, after which the

lists of influential microbes are compared. At this stage it is possible to identify which microbes are specific to men or women with respect to the progression of cirrhosis and hepatic encephalopathy.

The final process utilized in this study is a combination of principal component analysis and kernel density estimation. First, principal component analysis is used to perform dimensionality reduction of the relative abundance data. The resulting variable loadings in the first two principal components are then employed to provide context for the analysis of the data after it is transformed into the two-dimensional principal component (PC) space. Once data is in the PC space, kernel density estimation with Gaussian kernels is applied to visualize the distributions within each disease class (controls, cirrhosis without HE, HE on Lactulose, and HE on Rifaximin) and identify any differences. This analysis is performed separately on the male and female subpopulations to facilitate comparisons, but the principal component analysis (PCA) is completed only once on the entire dataset. Doing so ensures that PCA and kernel density plots can be compared between genders as the axes represent the same linear combinations of input features. If the PCA computation step is conducted separately for men and women, the generated principal components will not correspond to the same variable loadings and cannot be compared.

6.4 Results

As the initial feature reduction and analysis processes described in the previous section are conducted independently from each other, results will be presented separately for each one. As with section 6.3, the feature reduction will be discussed first, followed by the

correlation study. The random forest classifier and statistical testing will be presented next, with outcomes from the principal component analysis and kernel density estimation methods shared last.

The initial feature reduction step resulted in 29 features selected for inclusion in the analysis. Table 5 provides a list of these features along with the family level of each OTU.

Table 5 Features Included in Analysis

Bifidobacteriaceae	Streptococcaceae	Acidaminococcaceae
Coriobacteriaceae	Clostridiaceae	Veillonellaceae
Bacteroidaceae	Clostridiales cluster IV	Fusobacteriaceae
various Bacteroidales	Clostridiales cluster XI	Sutterellaceae
Porphyromonadaceae	Clostridiales cluster XIII	Desulfovibrionaceae
Prevotellaceae	Lachnospiraceae	Enterobacteriaceae
Rikenellaceae	Peptococcaceae	Pasteurellaceae
Carnobacteriaceae	Peptostreptococcaceae	Synergistaceae
Enterococcaceae	Ruminococcaceae	Verrucomicrobiaceae
Lactobacillaceae	Erysipelotrichaceae	

Heatmaps illustrating the correlation coefficients for each pair of features are provided in Figures 5 and 6. As evidenced by the plots, most feature pairings have correlations close to zero, but there are a few areas of interest on the plots. Clostridiales cluster XI and Peptococcaceae have a high Pearson correlation suggesting a linear relationship between the two. Additionally, several negative correlations can be observed, primarily in the Spearman heatmaps although some negative Pearson coefficients are also shown. Based on these results, the cirrhosis and hepatic encephalopathy data used does not seem to have many features with very direct connections to each other.

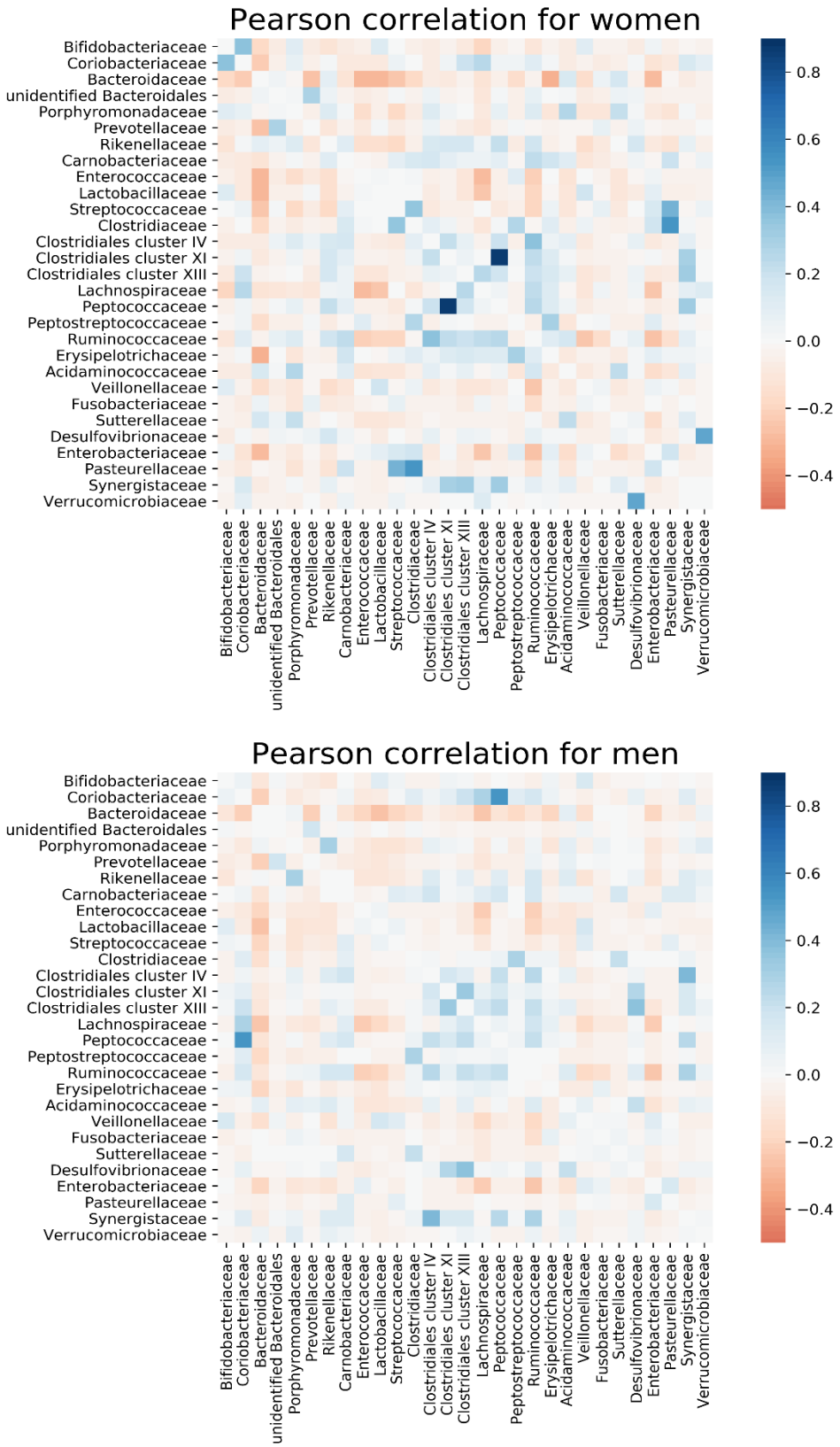

Figure 5 Pearson correlation heatmaps for women and men

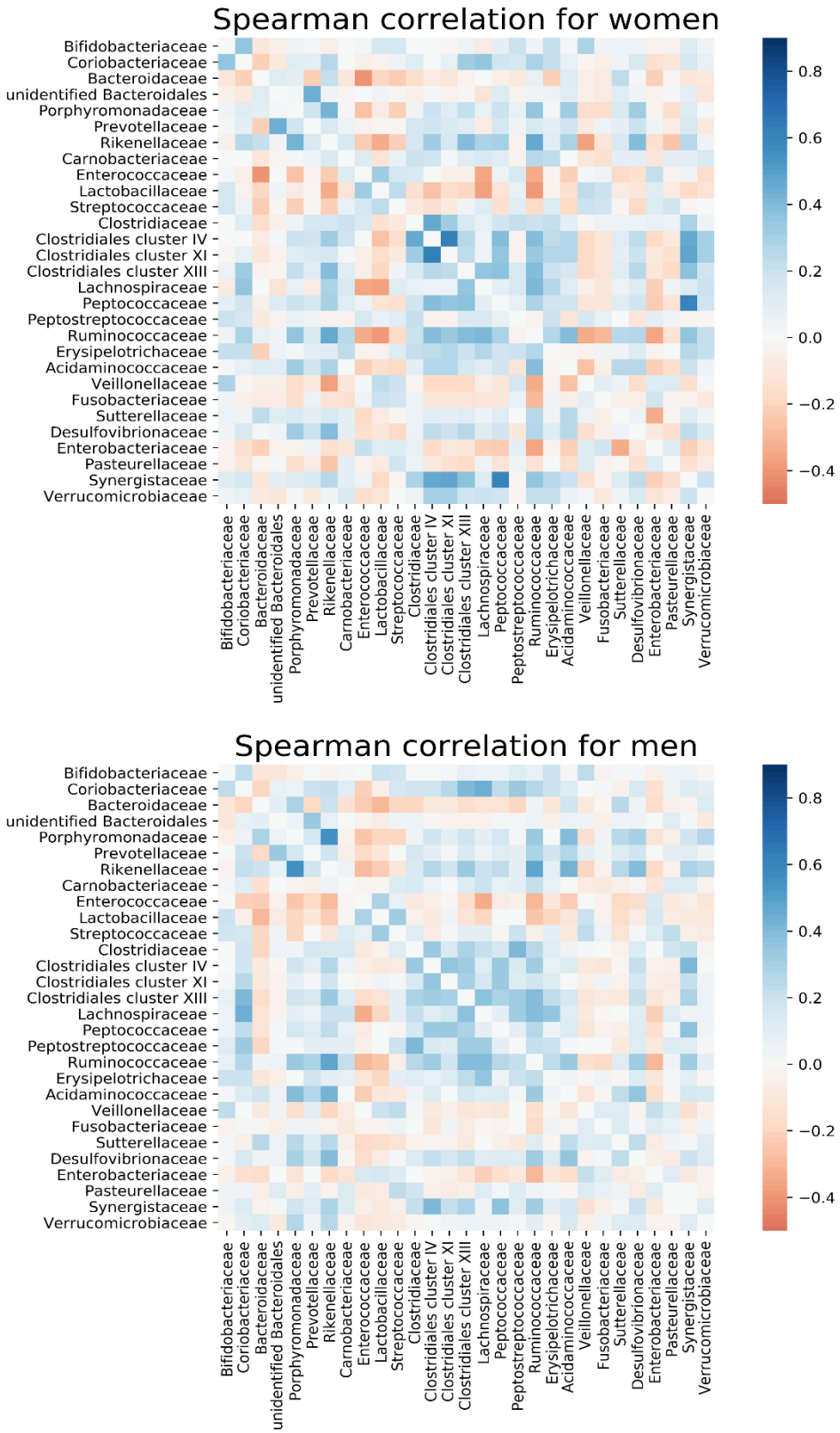

Figure 6 Spearman correlation heatmaps for women and men

The classification performance results (area under the curve) for the random forest classifiers across the nine tasks in Table 4 for both men and women are presented in Table 6. As evidenced by the variation in AUC values, the classifier performs better for certain tasks than others. The classifier appears to perform better when predicting between control and patients with HE, as may be expected due to the drastic difference in disease severity. The classification performance results also suggest that there are differences in how the gut microbiomes of men and women are affected by cirrhosis and HE as 'Control vs Cirrhosis,' 'Control vs Cirrhosis no HE,' 'Cirrhosis no HE vs HE,' and 'Cirrhosis no HE vs HE-Rif' show strong gender differences.

Table 6 AUC Values for Classification Tasks

Task Name	**AUC (std. dev.) for Women**	**AUC (std. dev.) for Men**
Control vs Cirrhosis	0.85 (0.06)	0.78 (0.05)
Control vs Cirrhosis no HE	0.89 (0.06)	0.73 (0.06)
Control vs HE	0.87 (0.08)	0.88 (0.03)
Control vs HE-Lac	0.84 (0.09)	0.86 (0.07)
Control vs HE-Rif	0.92 (0.08)	0.89 (0.02)
Cirrhosis no HE vs HE	0.72 (0.12)	0.80 (0.04)
Cirrhosis no HE vs HE-Lac	0.71 (0.10)	0.74 (0.06)
Cirrhosis no HE vs HE-Rif	0.74 (0.10)	0.84 (0.05)
HE-Lac vs HE-Rif	0.68 (0.12)	0.67 (0.10)

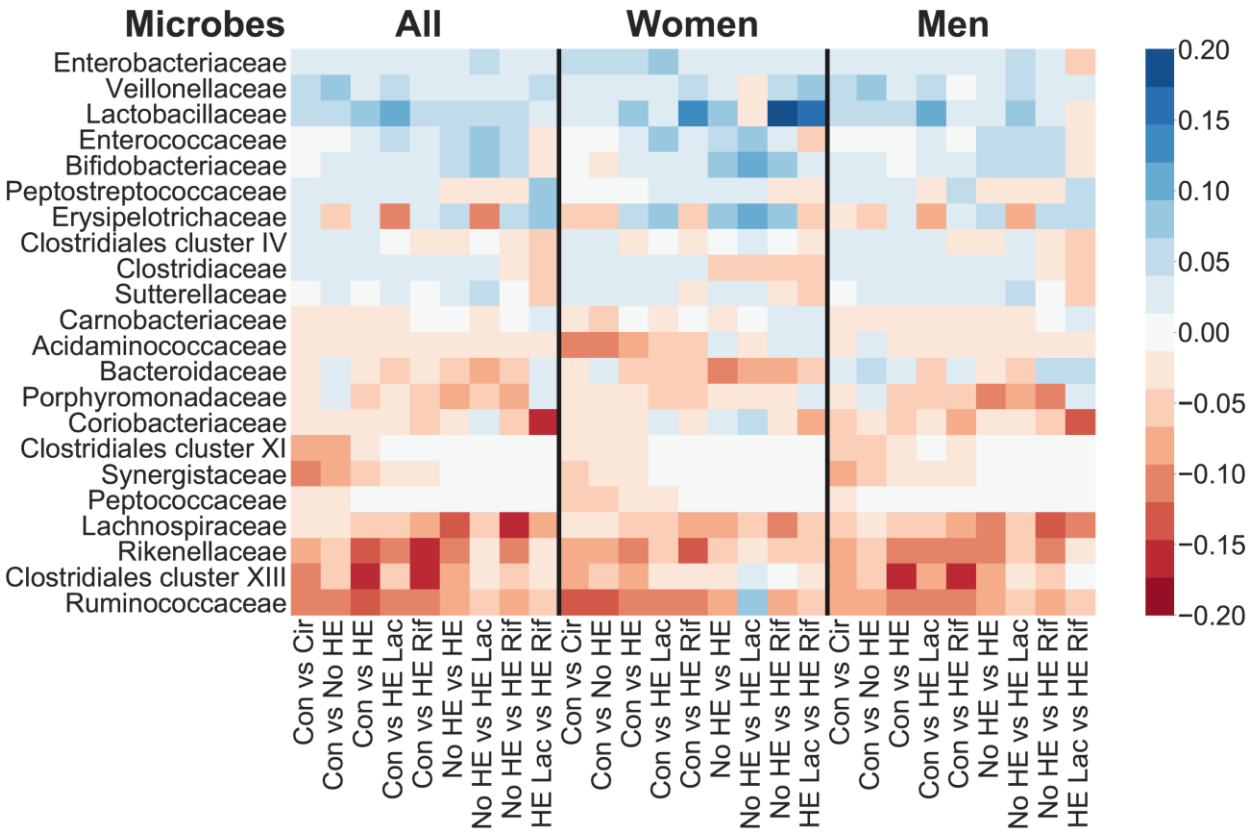

Figure 7 Signed feature importance heatmap for all classification tasks

The signed feature importance results for microbes which placed high in the feature importance list for any of the classification tasks are shown in Figure 7 with positive values indicating that the relative abundance of that microbe is generally greater in the more severely affected subpopulation and negative values indicating that the relative abundance of that microbe is generally lower in the more severely affected subpopulation. The primary purpose of this segmented heatmap is to facilitate the comparison of feature importance patterns between men and women. Some microbes which exhibit such differences include Lactobacillaceae, Acidaminococcaceae, and Bacteroidaceae having a greater importance for classification among women, Clostridiales cluster XIII and Porphyromonadaceae having a greater importance among men, and Erysipelotrichaceae showing differences in sign between

the genders. The classification tasks on the horizontal axis allow for the identification of which disease stages are most relevant for each microbe through a logical analysis of which tasks generated the greatest feature importance values. As an example, in men Rikenellaceae appears to be maximally relevant in tasks involving control or 'no HE' patients being compared to HE patients taking Rifaximin. Coriobacteriaceae is most important when classifying between HE patients on Lactulose and those on Rifaximin, indicating that it may be related to HE severity or could be an aspect of the microbiome being altered due to Rifaximin. These findings are corroborated by the relative abundance plots for Bacteroidaceae, Coriobacteriaceae, and Porphyromonadaceae in Figure 8.

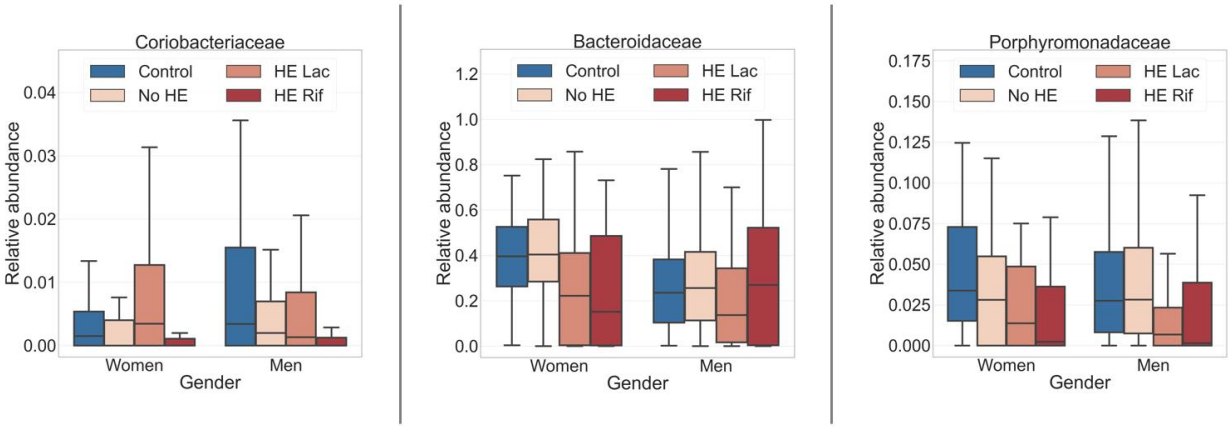

Figure 8 Relative abundances for select microbes

As a major goal of this project is exploration of the relationship between the gut microbiome and the brain, further statistical study is performed on this set of important microbes through the use of Wilcoxon tests between the healthy controls and patients with hepatic encephalopathy. Table 7 presents these computed p-values, separated by gender to allow for comparison. Although only microbes identified as important by the random forest

classifier were tested in this way, not all of them had significant differences with a threshold of 5.00E-02. P-values which were above this threshold are colored orange to allow for easy identification. Some were significantly different in only one gender, supporting the idea that men and women are affected differently by cirrhosis and HE.

Table 7 P-Values for HE vs Controls

Microbes	Women	Men
Enterobacteriaceae	8.00E-04	1.85E-03
Veillonellaceae	3.86E-04	2.16E-05
Lactobacillaceae	3.81E-06	9.20E-09
Enterococcaceae	2.83E-03	4.42E-04
Bifidobacteriaceae	1.53E-01	1.09E-01
Peptostreptococcacceae	7.48E-01	1.33E-05
Erysipelotrichaceae	1.6E-03	7.60E-04
Clostridiaceae	1.82E-02	2.37E-04
Streptoccaceae	4.79E-03	7.22E-01
Sutterellaceae	8.56E-02	5.59E-01
Carnobacteriaceae	5.14E-02	1.31E-03
Acidaminococcceae	2.18E-05	4.49E-04
Bacteroidaceae	2.14E-03	2.66E-01
Porphyromonadaceae	3.02E-03	3.72E-07
Coriobacteriaceae	3.94E-01	3.93E-08
Clostridiales cluster IV	7.31E-03	1.92E-03
Synergistaceae	5.48E-03	1.92E-04
Peptococcaceae	1.75E-02	6.44E-02
Lachnospiraceae	1.46E-04	1.12E-09
Rikenellaceae	3.85E-08	3.64E-14
Clostidiales cluster XIII	5.81E-05	1.56E-11
Ruminococcaceae	1.33E-08	1.74E-15

Figure 9 provides an overview of the variance contributions from each principal component generated via PCA and the resulting transformed data. The left scatter plot gives an

overview of the data colored by gender while the right scatter plot shows the data colored based on the four disease classes described in section 6.2. These plots show a triangular pattern, with the spread of principal component 1 being larger at lower values for principal component 0. The scatter plot colored by disease class indicates that samples from patients with HE tend to have lower values for principal component 1, with some concentration in the bottom left corner with low values for both principal components.

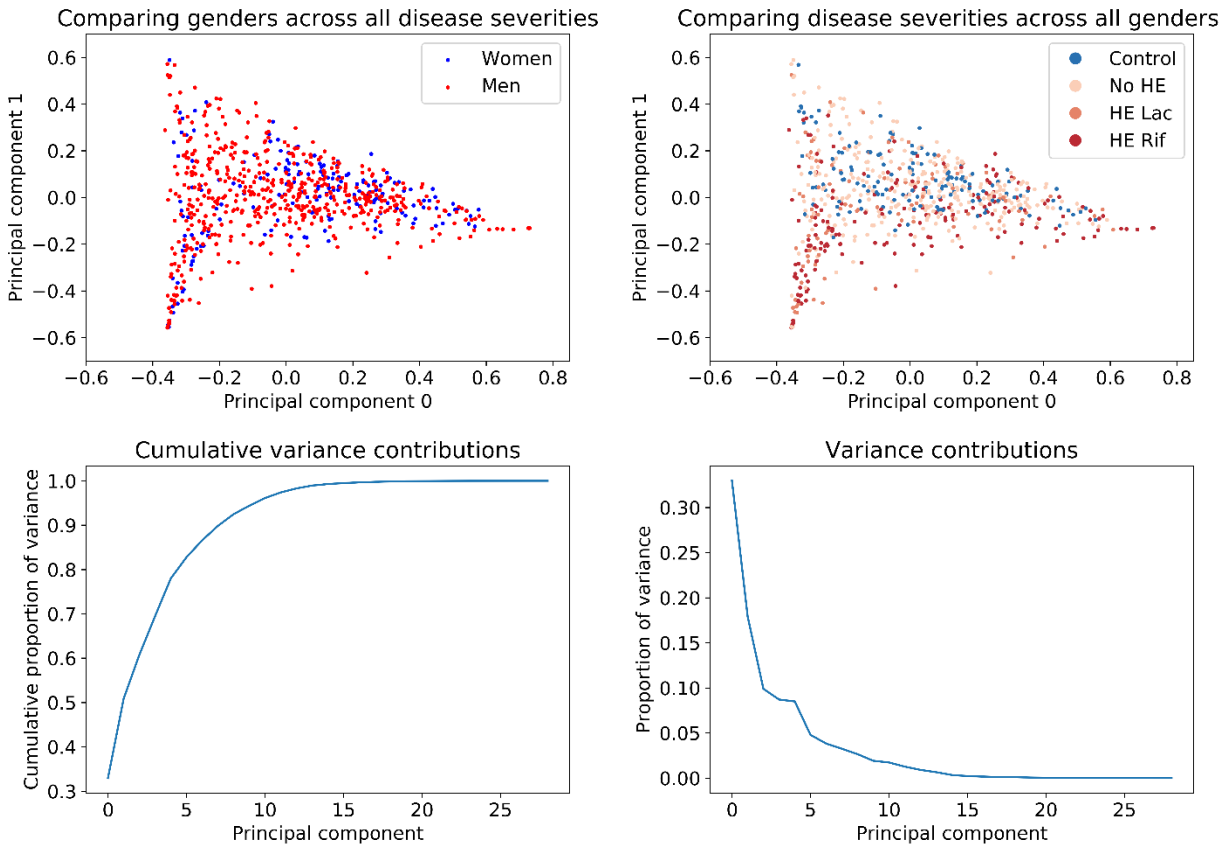

Figure 9 PCA results

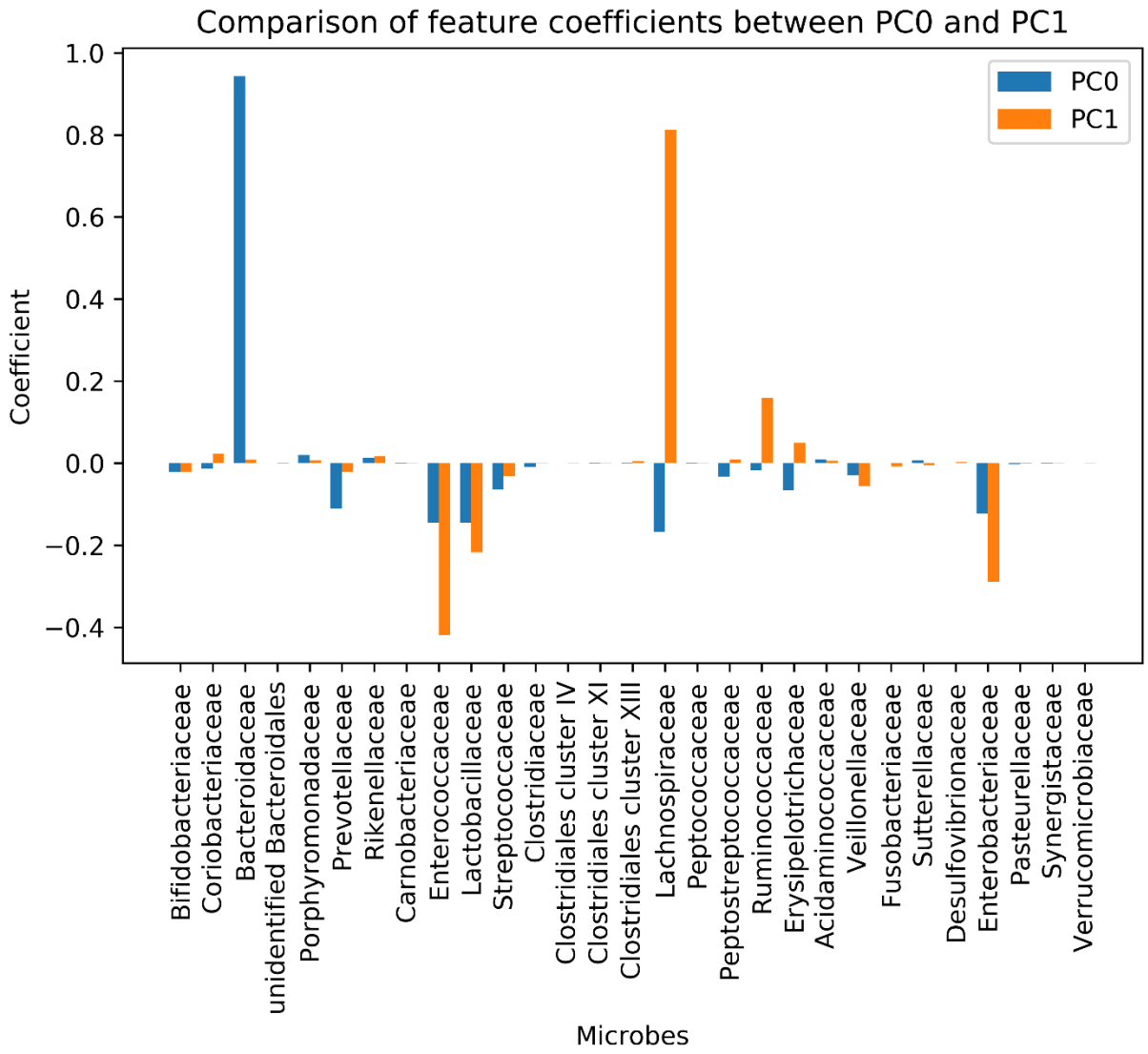

Figure 10 PC0 and PC1 variable loadings

Variable loadings for the first two principal components are shown in Figure 10. This plot by itself does not provide much information, but it outlines the composition of the principal components which allows for interpretation of subsequent analysis performed on the transformed data.

Once principal component analysis has been performed and the data has been transformed into the new two-dimensional space, kernel density estimation is applied to each of the eight disease-gender subpopulations as shown in Figure 11. It is apparent for both men and women that the distribution shifts towards negative values for both principal components as disease condition worsens – particularly once HE appears, as was suggested by the initial PCA scatter plots. While there appear to be men in the control subpopulation toward the top left corner, the only women who appear in this area are a small proportion of those with HE. Men also show higher densities when principal component zero values are low and principal component one values are near zero. Such distinctions between genders support the idea that cirrhosis and hepatic encephalopathy affect men and women in different ways. These changes can be put into the context of the microbes by combining the patterns apparent in the kernel density plots in Figure 11 with the variable loadings from Figure 10. Of particular interest is that the only way to significantly alter PC0 without affecting PC1 is through the reduction of Bacteroidaceae. This combined with the tendency for PC1 to be very low only when PC0 is also very low suggests that increases in abundance of the microbes with negative variable loadings for both principal components (e.g. Enterococcaceae, Lactobacillaceae, Enterobacteriaceae) may be somewhat dependent on decreases in Bacteroidaceae abundance. As evidenced by the earlier correlation analysis, this is not a straightforward relationship but may indicate that the presence of one microbe beyond a certain abundance level inhibits the ability of another microbe to replicate. Although the underlying biological processes cannot be explicitly identified, such results provide a starting point for targeted studies and experiments in the future.

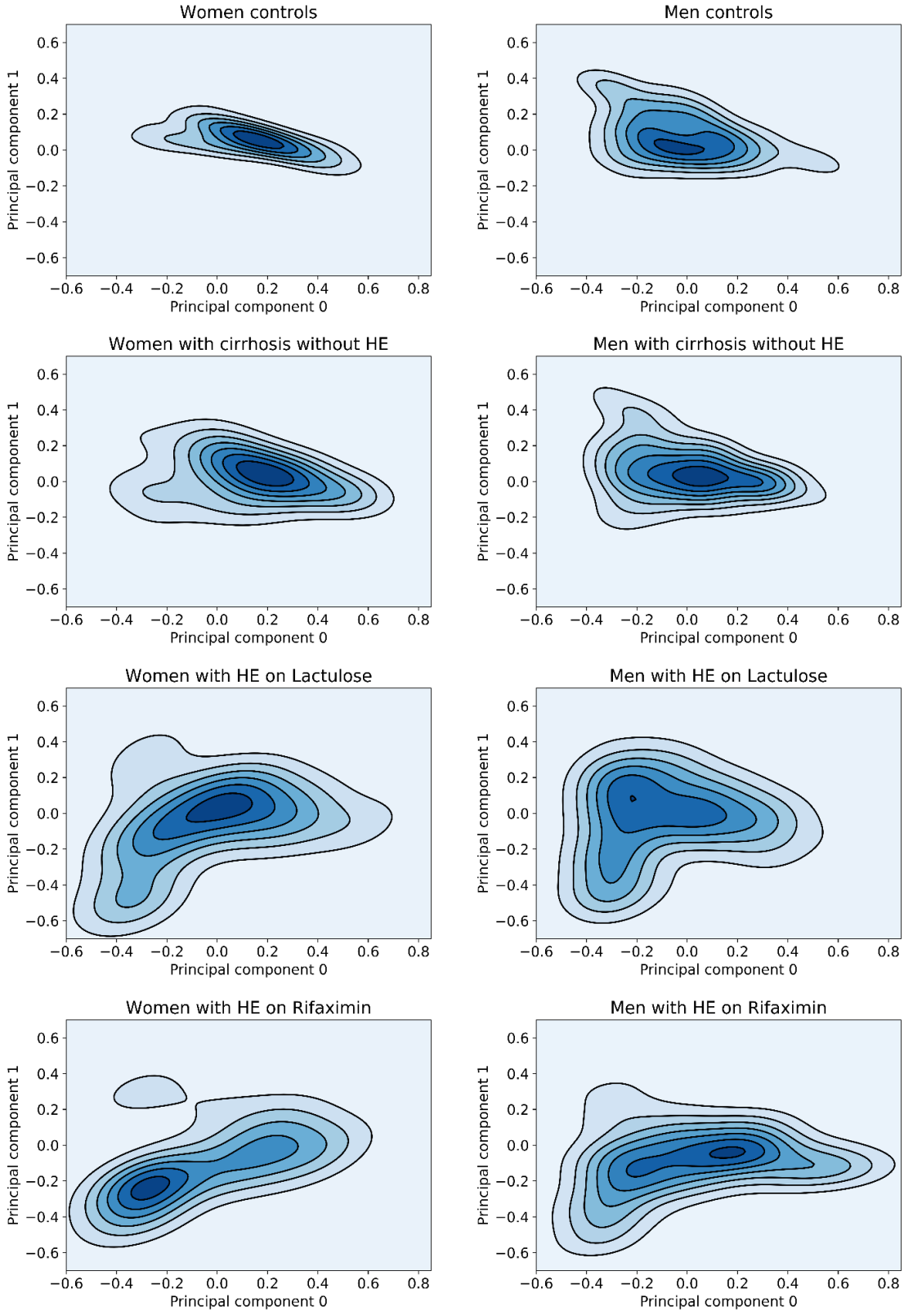

Figure 11 Kernel density for all disease classes across men and women

6.5 Methods discussion

A variety of methods from all three groups discussed in earlier chapters were used in this analysis. Univariate statistical methods (Chapter 3) were used to compute the p-values which indicate how significant the differences between disease classes were for the microbes identified by the classifier as important. The random forest classifier utilized to determine feature importance and identify influential microbes for men and women separately is a classification method (Chapter 4) at work. Multivariate joint analysis methods (Chapter 5) present in the cirrhosis and HE analysis are correlation study, principal component analysis, and kernel density estimation.

Although some techniques provide more information than others, each has benefits and drawbacks which must be considered when deciding which ones are most appropriate for a given problem and its corresponding data. Consider kernel density estimation; it provides a great visualization of the different distributions present in the data; however, converting data with a large number of features into a two-dimensional space requires the use of additional methods. In this case principal component analysis was used, but alternate dimensionality reduction techniques may also be applicable. Consideration of the variable loadings generated by PCA also allowed for more in-depth interpretation of the kernel density plots. Similarly, the feature importance results from the random forest classifier are effective for the identification of features which are likely to be important to the condition being studied, but the Wilcoxon tests were needed to determine statistical significance. Careful selection of methods which complement each other enables a more complete analysis along with a broader understanding

of the information encoded within the data, which consequently generates more beneficial insights about the microbiome being studied.

Chapter 7: Conclusion

This thesis reviews multiple analysis methods across three categories which can be applied to microbiome data. Univariate methods provide concrete statistical information about individual features in the data but are not able to capture the important interactions between microbes. The classification-based approaches allow for a more complete understanding of the influence of each of the features present, by generating a predictive model and interpreting it as a descriptive model. The broad category of multivariate or joint analysis techniques facilitates a deeper understanding of how the microbes influence each other and their joint relationship with the diseases or conditions being studied. A selection of these techniques is employed as part of a study on cirrhosis and hepatic encephalopathy with additional analysis performed on the differences between men and women.

Selection of the most appropriate techniques for a given problem is highly dependent on the nature of the data as well as the study's objectives. Most analyses will warrant the application of multiple methods as each has its own advantages and limitations. As such, a thorough understanding of the domain being investigated is required. Collaboration between data scientists performing the analysis and domain experts is beneficial at each step in the process to ensure that results are reasonable and ensuing steps are designed to address the desired questions. The final component necessary for any microbiome analysis is an in-depth consultation with domain experts to assess the overall validity of results and determine how

the insights from the data should be translated into information relevant to the conditions and processes being studied.

The primary area of advancement in this space is the development of new methods and evolution of existing methods to more directly determine the joint distribution relating the abundances of the most influential microbes to the disease states identified for the condition being studied. The high likelihood of complex interactions between microbes combined with the unique nature of each microbiome and condition motivate the creation of flexible methods which are capable of identifying the aforementioned joint distributions. A secondary opportunity for extension of the work presented here is the sharing of techniques between researchers studying human microbiomes in a medical context and scientists studying microbiomes in nature.

References

[1] I. Solt, “The human microbiome and the great obstetrical syndromes: A new frontier in maternal–fetal medicine,” *Best Pract. Res. Clin. Obstet. Gynaecol.*, vol. 29, no. 2, pp. 165–175, Feb. 2015, doi: 10.1016/j.bpobgyn.2014.04.024.

[2] J. P. Kiley and E. V. Caler, “The lung microbiome. A new frontier in pulmonary medicine,” *Ann. Am. Thorac. Soc.*, vol. 11, no. Supplement 1, pp. S66–S70, Jan. 2014, doi: 10.1513/AnnalsATS.201308-285MG.

[3] P. C. Kashyap, N. Chia, H. Nelson, E. Segal, and E. Elinav, “Microbiome at the frontier of personalized medicine,” *Mayo Clin. Proc.*, vol. 92, no. 12, pp. 1855–1864, Dec. 2017, doi: 10.1016/j.mayocp.2017.10.004.

[4] M. J. Bonder *et al.*, “The effect of host genetics on the gut microbiome,” *Nat. Genet.*, vol. 48, no. 11, Art. no. 11, Nov. 2016, doi: 10.1038/ng.3663.

[5] C. J. Robinson, B. J. M. Bohannan, and V. B. Young, “From structure to function: The ecology of host-associated microbial communities,” *Microbiol. Mol. Biol. Rev. MMBR*, vol. 74, no. 3, pp. 453–476, Sep. 2010, doi: 10.1128/MMBR.00014-10.

[6] M. Hattori and T. D. Taylor, “The human intestinal microbiome: A new frontier of human biology,” *DNA Res.*, vol. 16, no. 1, pp. 1–12, Feb. 2009, doi: 10.1093/dnares/dsn033.

[7] D. Knights, E. K. Costello, and R. Knight, “Supervised classification of human microbiota,” *FEMS Microbiol. Rev.*, vol. 35, no. 2, pp. 343–359, Mar. 2011, doi: 10.1111/j.1574-6976.2010.00251.x.

[8] N. Hasan and H. Yang, “Factors affecting the composition of the gut microbiota, and its modulation,” *PeerJ*, vol. 7, Aug. 2019, doi: 10.7717/peerj.7502.

[9] D. Rothschild *et al.*, “Environment dominates over host genetics in shaping human gut microbiota,” *Nature*, vol. 555, no. 7695, Art. no. 7695, Mar. 2018, doi: 10.1038/nature25973.

[10] C. Huttenhower *et al.*, “Structure, function and diversity of the healthy human microbiome,” *Nature*, vol. 486, no. 7402, Art. no. 7402, Jun. 2012, doi: 10.1038/nature11234.

[11] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower, “The healthy human microbiome,” *Genome Med.*, vol. 8, Apr. 2016, doi: 10.1186/s13073-016-0307-y.

[12] R. Sender, S. Fuchs, and R. Milo, “Revised estimates for the number of human and bacteria cells in the body,” *PLOS Biol.*, vol. 14, no. 8, p. e1002533, Aug. 2016, doi: 10.1371/journal.pbio.1002533.

[13] Y. J. Huang, “Asthma microbiome studies and the potential for new therapeutic strategies,” *Curr. Allergy Asthma Rep.*, vol. 13, no. 5, pp. 453–461, Oct. 2013, doi: 10.1007/s11882-013-0355-y.

[14] L. Brunkwall and M. Orho-Melander, “The gut microbiome as a target for prevention and treatment of hyperglycaemia in type 2 diabetes: From current human evidence to future possibilities,” *Diabetologia*, vol. 60, no. 6, pp. 943–951, Jun. 2017, doi: 10.1007/s00125-017-4278-3.

[15] A. Behrouzi, A. H. Nafari, and S. D. Siadat, “The significance of microbiome in personalized medicine,” *Clin. Transl. Med.*, vol. 8, May 2019, doi: 10.1186/s40169-019-0232-y.

[16] T. M. Kuntz and J. A. Gilbert, “Introducing the microbiome into precision medicine,” *Trends Pharmacol. Sci.*, vol. 38, no. 1, pp. 81–91, Jan. 2017, doi: 10.1016/j.tips.2016.10.001.

[17] H. E. Jakobsson, C. Jernberg, A. F. Andersson, M. Sjölund-Karlsson, J. K. Jansson, and L. Engstrand, “Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome,” *PLoS ONE*, vol. 5, no. 3, Mar. 2010, doi: 10.1371/journal.pone.0009836.

[18] A. G. Grottoli *et al.*, “Coral physiology and microbiome dynamics under combined warming and ocean acidification,” *PLOS ONE*, vol. 13, no. 1, p. e0191156, Jan. 2018, doi: 10.1371/journal.pone.0191156.

[19] H. K. Allen, L. A. Moe, J. Rodbumrer, A. Gaarder, and J. Handelsman, “Functional metagenomics reveals diverse β-lactamases in a remote Alaskan soil,” *ISME J.*, vol. 3, no. 2, Art. no. 2, Feb. 2009, doi: 10.1038/ismej.2008.86.

[20] D. Gat, Y. Mazar, E. Cytryn, and Y. Rudich, “Origin-dependent variations in the atmospheric microbiome community in eastern mediterranean dust storms,” *Environ. Sci. Technol.*, vol. 51, no. 12, pp. 6709–6718, Jun. 2017, doi: 10.1021/acs.est.7b00362.

[21] H. E. Vuong, J. M. Yano, T. C. Fung, and E. Y. Hsiao, “The microbiome and host behavior,” *Annu. Rev. Neurosci.*, vol. 40, no. 1, pp. 21–49, 2017, doi: 10.1146/annurev-neuro-072116-031347.

[22] X. C. Morgan and C. Huttenhower, “Chapter 12: Human microbiome analysis,” *PLOS Comput. Biol.*, vol. 8, no. 12, p. e1002808, Dec. 2012, doi: 10.1371/journal.pcbi.1002808.

[23] J. Pollock, L. Glendinning, T. Wisedchanwet, and M. Watson, “The madness of microbiome: Attempting to find consensus ‘best practice’ for 16S microbiome studies,” *Appl. Environ. Microbiol.*, vol. 84, no. 7, pp. e02627-17, /aem/84/7/e02627-17.atom, Feb. 2018, doi: 10.1128/AEM.02627-17.

[24] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” presented at the ICML-2003, Washington DC, 2003.

[25] K. Saboo *et al.*, “Gender-related alterations in gut microbial composition and function in hepatic encephalopathy,” *J. Hepatol.*, In print.

[26] K. R. Foster, J. Schluter, K. Z. Coyte, and S. Rakoff-Nahoum, “The evolution of the host microbiome as an ecosystem on a leash,” *Nature*, vol. 548, no. 7665, pp. 43–51, Aug. 2017, doi: 10.1038/nature23292.

[27] E. R. Deyle, R. M. May, S. B. Munch, and G. Sugihara, “Tracking and forecasting ecosystem interactions in real time,” *Proc. R. Soc. B Biol. Sci.*, vol. 283, no. 1822, p. 20152258, Jan. 2016, doi: 10.1098/rspb.2015.2258.

[28] S. Siegel, “Nonparametric statistics,” *Am. Stat.*, vol. 11, no. 3, pp. 13–19, Jun. 1957, doi: 10.1080/00031305.1957.10501091.

[29] A. Gelman, “Analysis of variance--why it is more important than ever,” *Ann. Stat.*, vol. 33, no. 1, pp. 1–53, Feb. 2005, doi: 10.1214/009053604000001048.

[30] M. W. Fagerland and L. Sandvik, “The Wilcoxon–Mann–Whitney test under scrutiny,” *Stat. Med.*, vol. 28, no. 10, pp. 1487–1497, 2009, doi: 10.1002/sim.3561.

[31] M. P. Fay and M. A. Proschan, “Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules,” *Stat. Surv.*, vol. 4, pp. 1–39, 2010, doi: 10.1214/09-SS051.

[32] J. H. McDonald, “Kruskal–Wallis test,” in *Handbook of Biological Statistics*, 3rd ed., Baltimore, Maryland: Sparky House Publishing.

[33] S. Cafazzo, P. Valsecchi, R. Bonanni, and E. Natoli, “Dominance in relation to age, sex, and competitive contexts in a group of free-ranging domestic dogs,” *Behav. Ecol.*, vol. 21, no. 3, pp. 443–455, May 2010, doi: 10.1093/beheco/arq001.

[34] T. Sueyoshi and S. Aoki, “A use of a nonparametric statistic for DEA frontier shift: The Kruskal and Wallis rank test,” *Omega*, vol. 29, no. 1, pp. 1–18, Feb. 2001, doi: 10.1016/S0305-0483(00)00024-4.

[35] P. Pavlidis, “Using ANOVA for gene selection from microarray studies of the nervous system,” *Methods*, vol. 31, no. 4, pp. 282–289, Dec. 2003, doi: 10.1016/S1046-2023(03)00157-9.

[36] J. Graessler *et al.*, “Metagenomic sequencing of the human gut microbiome before and after bariatric surgery in obese patients with type 2 diabetes: correlation with inflammatory and metabolic parameters,” *Pharmacogenomics J.*, vol. 13, no. 6, Art. no. 6, Dec. 2013, doi: 10.1038/tpj.2012.43.

[37] Y. Xia and J. Sun, “Hypothesis testing and statistical analysis of microbiome,” *Genes Dis.*, vol. 4, no. 3, pp. 138–148, Sep. 2017, doi: 10.1016/j.gendis.2017.06.001.

[38] G. Falony *et al.*, “Population-level analysis of gut microbiome variation,” *Science*, vol. 352, no. 6285, pp. 560–564, Apr. 2016, doi: 10.1126/science.aad3503.

[39] W. S. Noble, “How does multiple testing correction work?” *Nat. Biotechnol.*, vol. 27, no. 12, Art. no. 12, Dec. 2009, doi: 10.1038/nbt1209-1135.

[40] N. A. Bokulich, M. R. Dillon, E. Bolyen, B. D. Kaehler, G. A. Huttley, and J. G. Caporaso, “q2-sample-classifier: Machine-learning tools for microbiome classification and regression,” *J. Open Res. Softw.*, vol. 3, no. 30, 2018, doi: 10.21105/joss.00934.

[41] S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *J. Am. Stat. Assoc.*, vol. 97, no. 457, pp. 77–87, Mar. 2002, doi: 10.1198/016214502753479248.

[42] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, “An extensive comparison of recent classification tools applied to microarray data,” *Comput. Stat. Data Anal.*, vol. 48, no. 4, pp. 869–885, Apr. 2005, doi: 10.1016/j.csda.2004.03.017.

[43] A. Kataria and M. D. Singh, “A review of data classification using K-nearest neighbour algorithm,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 6, Jun. 2013.

[44] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, “Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach,” *ArXiv14090919 Cs*, Sep. 2014, Accessed: Jun. 10, 2020. [Online]. Available: http://arxiv.org/abs/1409.0919.

[45] J. Gou, L. Du, Y. Zhang, and T. Xiong, “A new distance-weighted k-nearest neighbor classifier,” *J. Inf. Comput. Sci.*, vol. 9, no. 6, pp. 1429–1436, 2012.

[46] S. Zhang, D. Cheng, Z. Deng, M. Zong, and X. Deng, “A novel kNN algorithm with data-driven k parameter computation,” *Pattern Recognit. Lett.*, vol. 109, pp. 44–54, Jul. 2018, doi: 10.1016/j.patrec.2017.09.036.

[47] D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou, “Support vector machine soft margin classifiers: Error analysis,” *J. Inf. Comput. Sci.*, vol. 5, pp. 1143–1175, 2004.

[48] K. Crammer and Y. Singer, “On the algorithmic implementation of multiclass kernel-based vector machines,” *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, 2001.

[49] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002, doi: 10.1109/72.991427.

[50] S. Ali and K. A. Smith-Miles, “A meta-learning approach to automatic kernel selection for support vector machines,” *Neurocomputing*, vol. 70, no. 1, pp. 173–186, Dec. 2006, doi: 10.1016/j.neucom.2006.03.004.

[51] T. Jebara, “Multi-task feature and kernel selection for SVMs,” in *Proceedings of the Twenty-First International Conference on Machine Learning*, Banff, Alberta, Canada, Jul. 2004, p. 55, doi: 10.1145/1015330.1015426.

[52] J. Nahar, S. Ali, and Y.-P. P. Chen, “Microarray data classification using automatic SVM kernel selection,” *DNA Cell Biol.*, vol. 26, no. 10, pp. 707–712, Oct. 2007, doi: 10.1089/dna.2007.0590.

[53] R. Díaz-Uriarte and S. Alvarez de Andrés, “Gene selection and classification of microarray data using random forest,” *BMC Bioinformatics*, vol. 7, no. 1, p. 3, Jan. 2006, doi: 10.1186/1471-2105-7-3.

[54] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[55] J. Benesty, J. Chen, Y. Huang, and I. Cohen, “Pearson correlation coefficient,” in *Noise Reduction in Speech Processing*, vol. 2, Berlin, Heidelberg: Springer, 2009, pp. 1–4.

[56] M. Kotlyar, S. Fuhrman, A. Ableson, and R. Somogyi, “Spearman correlation identifies statistically significant gene expression clusters in spinal cord development and injury,” *Neurochem. Res.*, vol. 27, no. 10, pp. 1133–1140, Oct. 2002, doi: 10.1023/A:1020969208033.

[57] A. J. Bishara and J. B. Hittner, “Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches,” *Psychol. Methods*, vol. 17, no. 3, p. 399, 20120507, doi: 10.1037/a0028087.

[58] A. G. Asuero, A. Sayago, and A. G. González, “The correlation coefficient: An overview,” *Crit. Rev. Anal. Chem.*, vol. 36, no. 1, pp. 41–59, Jan. 2006, doi: 10.1080/10408340500526766.

[59] H. Abdi, “Multiple correlation coefficient,” in *Encyclopedia of Measurement and Statistics*, N. Salkind, Ed. Thousand Oaks, CA: Sage, 2007.

[60] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, Jun. 2006, doi: 10.1198/106186006X113430.

[61] J. Cadima and I. T. Jolliffe, “Loading and correlations in the interpretation of principle compenents,” *J. Appl. Stat.*, vol. 22, no. 2, pp. 203–214, Jan. 1995, doi: 10.1080/757584614.

[62] I. T. Jolliffe, “Rotation of principal components: Choice of normalization constraints,” *J. Appl. Stat.*, vol. 22, no. 1, pp. 29–35, Jan. 1995, doi: 10.1080/757584395.

[63] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998, doi: 10.1162/089976698300017467.

[64] R. Rosipal, M. Girolami, L. J. Trejo, and A. Cichocki, “Kernel PCA for feature extraction and de-noising in nonlinear regression,” *Neural Comput. Appl.*, vol. 10, no. 3, pp. 231–243, Dec. 2001, doi: 10.1007/s521-001-8051-z.

[65] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, “Kernel PCA and de-noising in feature spaces,” *Adv. Neural Inf. Process. Syst.*, pp. 536–542, 1999.

[66] M. Hubert and S. Engelen, “Robust PCA and classification in biosciences,” *Bioinformatics*, vol. 20, no. 11, pp. 1728–1736, Jul. 2004, doi: 10.1093/bioinformatics/bth158.

[67] K. A. Capone, S. E. Dowd, G. N. Stamatas, and J. Nikolovski, “Diversity of the human skin microbiome early in life,” *J. Invest. Dermatol.*, vol. 131, no. 10, pp. 2026–2032, Oct. 2011, doi: 10.1038/jid.2011.168.

[68] B. Vitali, F. Cruciani, G. Picone, C. Parolin, G. Donders, and L. Laghi, “Vaginal microbiome and metabolome highlight specific signatures of bacterial vaginosis,” *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 34, no. 12, pp. 2367–2376, Dec. 2015, doi: 10.1007/s10096-015-2490-y.

[69] R. Liu and L. Yang, “Kernel estimation of multivariate cumulative distribution function,” *J. Nonparametric Stat.*, vol. 20, no. 8, pp. 661–677, Nov. 2008, doi: 10.1080/10485250802326391.

[70] M. Rudemo, “Empirical choice of histograms and kernel density estimators,” *Scand. J. Stat.*, vol. 9, no. 2, pp. 65–78, 1982.

[71] X. Zhang, M. L. King, and R. J. Hyndman, “A Bayesian approach to bandwidth selection for multivariate kernel density estimation,” *Comput. Stat. Data Anal.*, vol. 50, no. 11, pp. 3009–3031, Jul. 2006, doi: 10.1016/j.csda.2005.06.019.

[72] S.-T. Chiu, “Bandwidth selection for kernel density estimation,” *Ann. Stat.*, vol. 19, no. 4, pp. 1883–1905, 1991.

[73] B. A. Turlach, “Bandwidth selection in kernel density estimation: A review,” presented at the CORE and Institut de Statistique, 1993.

[74] T. Duong, A. Cowling, I. Koch, and M. P. Wand, “Feature significance for multivariate kernel density estimation,” *Comput. Stat. Data Anal.*, vol. 52, no. 9, pp. 4225–4242, May 2008, doi: 10.1016/j.csda.2008.02.035.

[75] A. Pérez, P. Larrañaga, and I. Inza, “Bayesian classifiers based on kernel density estimation: Flexible classifiers,” *Int. J. Approx. Reason.*, vol. 50, no. 2, pp. 341–362, Feb. 2009, doi: 10.1016/j.ijar.2008.08.008.

[76] C. M. Shafer and C. A. Doswell III, “Using kernel density estimation to identify, rank, and classify severe weather outbreak events,” *Electron. J Sev. Storms Meteor*, vol. 6, no. 2, pp. 1–28, 2011.

[77] P. Smyth, A. Gray, and U. M. Fayyad, “Retrofitting decision tree classifiers using kernel density estimation,” in *Machine Learning Proceedings 1995*, A. Prieditis and S. Russell, Eds. San Francisco (CA): Morgan Kaufmann, 1995, pp. 506–514.

[78] M. Kolar, H. Liu, and E. Xing, “Markov network estimation from multi-attribute data,” in *International Conference on Machine Learning*, Feb. 2013, pp. 73–81, Accessed: Jun. 25, 2020. [Online]. Available: http://proceedings.mlr.press/v28/kolar13a.html.

[79] M. Kolar, H. Liu, and E. P. Xing, “Graph estimation from multi-attribute data,” *J. Mach. Learn. Res. JMLR*, vol. 15, no. May, pp. 1713–1750, May 2014.

[80] A. Naqvi, H. Rangwala, A. Keshavarzian, and P. Gillevet, “Network-based modeling of the human gut microbiome,” *Chem. Biodivers.*, vol. 7, no. 5, pp. 1040–1050, May 2010, doi: 10.1002/cbdv.200900324.

[81] S. Kim, I. Thapa, L. Zhang, and H. Ali, “A novel graph theoretical approach for modeling microbiomes and inferring microbial ecological relationships,” *BMC Genomics*, vol. 20, no. 11, p. 945, Dec. 2019, doi: 10.1186/s12864-019-6288-7.

[82] M. Kanehisa and S. Goto, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.

[83] E. A. Tsochatzis, J. Bosch, and A. K. Burroughs, “Liver cirrhosis,” *The Lancet*, vol. 383, no. 9930, pp. 1749–1761, May 2014, doi: 10.1016/S0140-6736(14)60121-5.

[84] D. Schuppan and N. H. Afdhal, “Liver cirrhosis,” *The Lancet*, vol. 371, no. 9615, pp. 838–851, Mar. 2008, doi: 10.1016/S0140-6736(08)60383-9.

[85] P. Ferenci, “Hepatic encephalopathy,” *Gastroenterol. Rep.*, vol. 5, no. 2, pp. 138–147, May 2017, doi: 10.1093/gastro/gox013.

[86] A. T. Blei and J. Córdoba, “Hepatic encephalopathy,” *Am. J. Gastroenterol.*, vol. 96, no. 7, pp. 1968–1976, Jul. 2001, doi: 10.1016/S0002-9270(01)02527-8.

[87] F. F. Poordad, “Review article: The burden of hepatic encephalopathy,” *Aliment. Pharmacol. Ther.*, vol. 25, no. s1, pp. 3–9, 2007, doi: 10.1111/j.1746-6342.2006.03215.x.

[88] N. M. Bass, G. Neff, T. Frederick, and A. Shaw, “Rifaximin treatment in hepatic encephalopathy,” *N Engl J Med*, vol. 362, no. 12, pp. 1071–1081, 2010.

[89] S. M. Riordan and R. Williams, “Treatment of hepatic encephalopathy,” *N. Engl. J. Med.*, vol. 337, no. 7, pp. 473–479, Aug. 1997, doi: 10.1056/NEJM199708143370707.

[90] C. Acharya and J. S. Bajaj, “Altered microbiome in patients with cirrhosis and complications,” *Clin. Gastroenterol. Hepatol.*, vol. 17, no. 2, pp. 307–321, Jan. 2019, doi: 10.1016/j.cgh.2018.08.008.

[91] N. S. Betrapally, P. M. Gillevet, and J. S. Bajaj, “Gut microbiome and liver disease,” *Transl. Res.*, vol. 179, pp. 49–59, Jan. 2017, doi: 10.1016/j.trsl.2016.07.005.

[92] C. R. Martin, V. Osadchiy, A. Kalani, and E. A. Mayer, “The brain-gut-microbiome axis,” *Cell. Mol. Gastroenterol. Hepatol.*, vol. 6, no. 2, pp. 133–148, Apr. 2018, doi: 10.1016/j.jcmgh.2018.04.003.

[93] R. Rai, V. A. Saraswat, and R. K. Dhiman, “Gut microbiota: Its role in hepatic encephalopathy,” *J. Clin. Exp. Hepatol.*, vol. 5, no. Suppl 1, pp. S29–S36, Mar. 2015, doi: 10.1016/j.jceh.2014.12.003.

[94] P. Zheng *et al.*, “The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice,” *Sci. Adv.*, vol. 5, no. 2, p. eaau8317, Feb. 2019, doi: 10.1126/sciadv.aau8317.

[95] K. Nemani, R. Hosseini Ghomi, B. McCormick, and X. Fan, “Schizophrenia and the gut–brain axis,” *Prog. Neuropsychopharmacol. Biol. Psychiatry*, vol. 56, pp. 155–160, Jan. 2015, doi: 10.1016/j.pnpbp.2014.08.018.

[96] V. Taneja, “Chapter 39 - Microbiome: Impact of gender on function & characteristics of gut microbiome,” in *Principles of Gender-Specific Medicine (Third Edition)*, M. J. Legato, Ed. San Diego: Academic Press, 2017, pp. 569–583.