Impacts of Diet Intervention on the Gut Microbiome

Simranjeet Bilkhu

August 22, 2025

1 Introduction

Recent diet innovations on social media, such as the carnivore diet, can be a cause for concern for how they might impact the human microbiome over time. Diet-induced changes to the microbiome are in fact suspected of contributing to chronic health issues such as irritable bowel syndrome and obesity[1]. Indeed, diet may significantly shape the structure and function of the gut microbiome[2]. There has been increased research on using diet interventions as treatments to long term health issues. For example, plant-based diet interventions for the prevention and treatment of type 2 diabetes[3]. It's also been proposed that the biological mechanisms of action that diet impacts health is may in part be through the gut microbiome[4]. To further investigate these relationships, we are using the dataset from David et al.'s[5] study where they randomly assigned diets (either animal or diet) and collected daily microbiome samples from subjects. Based on this dataset, our goals are to:

1. First, apply principal component analysis to reduce the dimension of our data (originally, there were over 1400 features in the processed data used in the analysis), and explore the relationships between the components and the original features.

2. Use model based clustering to see how much the clusters trained on the samples from the plant based group differ in comparison to the clusters from the animal based group

[1] Lawrence A David, et al., "Diet rapidly and reproducibly alters the human gut microbiome", *Nature* vol. 505, no. 7484, 2014, pp. 559–63.

[2] Gary D Wu, et al., "Linking long-term dietary patterns with gut microbial enterotypes", *Science* vol. 334, no. 6052, 2011, pp. 105–08.

[3] Michelle McMacken and Sapana Shah, "A plant-based diet for the prevention and treatment of type 2 diabetes", *Journal of geriatric cardiology: JGC* vol. 14, no. 5, 2017, p. 342.

[4] Justin L Sonnenburg and Fredrik Bäckhed, "Diet–microbiota interactions as moderators of human metabolism", *Nature* vol. 535, no. 7610, 2016, pp. 56–64.

[5] David et al.

3. Predict the type of diet someone has been consuming through random forest, and explore the feature importances.

Achieving the final goal would allow researchers to determine which microbes are over or under-represented under certain diet protocols, and thus the health benefits through diet interventions may be achieved through supplementation of the desired microbes. The rest of this paper will discuss the data and the preprocessing steps used to prepare the data, describe the models used in the analysis, discuss the results, and finally go over potential pitfalls of the analysis. A brief background on microbiome data analysis is provided in the appendix.

1.1 Data Description

	Sample$_1$	**Sample**$_2$	$\cdots$	**Sample**$_j$	**Sample**$_N$
Taxa$_1$	K_{11}	K_{12}	$\cdots$	K_{1j}	K_{1N}
Taxa$_2$	K_{21}	K_{22}	$\cdots$	K_{2j}	K_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Taxa$_m$	K_{m1}	K_{m2}	$\cdots$	K_{mj}	K_{mN}

Table 1: Example Table of Taxa vs. Specimens

The data used for this analysis is from: *Diet rapidly and reproducibly alters the human gut microbiome*, and is available here[6]. In the paper, the researchers randomly assigned diets at random times, and collected gut microbiome samples from the subjects on a daily basis. The resulting data was split into three key parts:

- An OTU table containing the counts of each taxonomic unit (which represent different species of microbes) per sample. See table 1 for an example.
- A Taxa table, which containes the phylogenetic information for each taxa.
- Sample metadata, which contains subject id's, the diet assigned, and the intervention times

For the purposes of this analysis, we will ignore the temporal aspect of the data, and focus only on the taxa counts, and the diet assigned for each sample. Overall, after preprocessing was completed there were 236 observations (one observation for each sample), one response

[6] David et al.

variable containing (being the diet assigned to each subject at the time the sample was drawn) and 1415 features, each feature representing their own separate taxa (a proxy for the species of microbe).

1.2 Data Preparation

1.2.1 Filtering

We first ensured that the taxa table only contains taxa that are present in the OTU table. Since measurement errors can introduce unidentifiable taxa, we removed any taxa that couldn't be identified through the taxa table. This step eliminated taxa for which we couldn't determine what species of microbe they represent. Subsequently, we filtered out taxa that appeared in fewer than 5 percent of the samples, as these likely represent sequencing errors rather than actual microbes.

1.2.2 Accounting For Differing Library Sizes

A key challenge when it comes to dealing with microbiome data is that different samples will often have different library sizes (see the appendix for details), meaning that the length of the sequences generated are different. This intales that we cannot directly use counts, nor relative abundances in our analysis, as this could bias results in the analysis. Mathematically, we can express this problem as follows:

$$E(K_{ij}) = \mu_{ij} d_j$$

Here, μ_{ij} is the true mean of the inderlying distribution generating the counts, d_j is the size factor induced by the uneven sequncing depth. In short, to account for this, we use the median of ratio's method[7] to estimate the $d_j's$, and normalize the data by the size factors.

1.2.3 Variance Stabilization

Moreover, since we are working with count data, we needed to stabilize the counts to achieve homogeneity of variance. We applied a variance stabilization to the normalized counts using an arc sine transformation.

[7] Simon Anders, et al., "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor", *Nature protocols* vol. 8, no. 9, 2013, pp. 1765–86.

2 Model Description

2.1 Principal Component Analysis

The objective of PCA is to transform a large set of features to a smaller set of features while still retaining as much information as possible. The principal components may be derived via spectral decomposition, as follows:

Let $\mathbf{x}$ be a $p \times 1$ random standardized vector such that $\mathbf{x} \sim \mathbf{N}(\mathbf{0}, \mathbf{\Sigma})$, where Σ is a correlation matrix. Since Σ is a correlation matrix, it's positive semi-definite, and thus admits a spectral decomposition $\mathbf{\Sigma} = \mathbf{CDC^T}$. Then $\mathbf{z} = \mathbf{C^T x}$ are the principal components of $\mathbf{x}$

Note that $\mathbf{z} = \mathbf{C^T x} \Rightarrow \mathbf{x} = \mathbf{Cz}$. This equation is re-parameterized to a factor model in the following:

$$\begin{aligned}
\mathbf{x} &= \mathbf{Cz} \\
&= \mathbf{CD}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}\,\mathbf{z} \\
&= \underbrace{\mathbf{CD}^{\frac{1}{2}}}_{k\times k}\ \underbrace{\mathbf{z}_2}_{k\times 1} \quad \text{(Normalize the principal components)} \\
&= (\underbrace{\mathbf{\Lambda}}_{k\times t} \mid \underbrace{\mathbf{M}}_{k\times(k-t)}) \left(\begin{array}{c} \mathbf{f} \\ \hline \mathbf{g} \end{array}\right) \begin{matrix} \leftarrow t \times 1 \\ \leftarrow (k-t) \times 1 \end{matrix} \quad \text{(Take the first } t \text{ components and separate out the rest)} \\
&= \mathbf{\Lambda f} + \mathbf{Mg} \\
&= \mathbf{\Lambda f} + \mathbf{e} \quad \text{(The omitted components are absorbed in the error term)}
\end{aligned}$$

Because of the above re-parameterization, for the rest of the paper, when needed, we will be referring to the entries of $\mathbf{C}$ as the loadings for the principal components.

2.2 Parsimonious Gaussian Mixture Models

We start with first considering the following Gaussian mixture model [8]:

$$p\left(\mathbf{x}|\theta\right) = \sum_{k=1}^{K} \pi_k \phi\left(\mathbf{x} \mid \mu_{\mathbf{k}}, \mathbf{\Sigma_k}\right).$$

The above model may be used to cluster the data. However, in high dimensional settings the number of covariance parameters is so large that it becomes unfeasible for large data

[8] From the STATS 790 Statistical Learning Slides

sets. One way to alleviate this issue is to constrain the covariance matrices to that of a factor analysis model. The parameters are then estimated using an EM algorithm.

2.3 Random Forests

To understand how random forests work, we must begin with ensemble learning. Let's first recall a key weakness of decision trees: their instability. As noted in *Probabilistic Machine Learning: An Introduction* (Murphy), decision trees can vary significantly when trained on different subsets of data, indicating high variance. The core idea of ensemble learning addresses is to address this limitation by training multiple trees and aggregating their predictions (typically through averaging or similar methods) to create a model with reduced variance. This can be expressed as[9]:

$$f\left(y|\mathbf{x}\right) = \frac{1}{M}\sum_{n=1}^{M} f_n\left(y|\mathbf{x}\right).$$

Were f_n are decision trees.

2.3.1 Precursor to Random Forests: Bagging

To begin building a model that follows the ensemble approach, we first explore bagging (Bootstrap Aggregating). Note that the above equation only works when each of the trees $\hat{f}_n$ are different, which bagging addresses by first creating M bootstrap samples, and then training a tree on each of those samples and then aggregate them. Specifically, let $\hat{f}_n$ be the tree trained on the n^{th} bootstrap sample. Then, averaging the M trees gives[10]:

$$\hat{f}_{bag} = \frac{1}{M}\sum_{n=1}^{M} \hat{f}_n.$$

2.4 Random Forests

One problem with the above approach is that the trees themselves are correlated with each other. To address this issue, we train each tree in a different way: Each time a split is due in a tree, choose a subset of the predictor variables (chosen at random), and then choose the variable that gives the best split. To choose the variable to split on, we can use different

[9] Kevin P. Murphy, *Probabilistic Machine Learning: An introduction* (MIT P, 2022).

[10] Murphy.

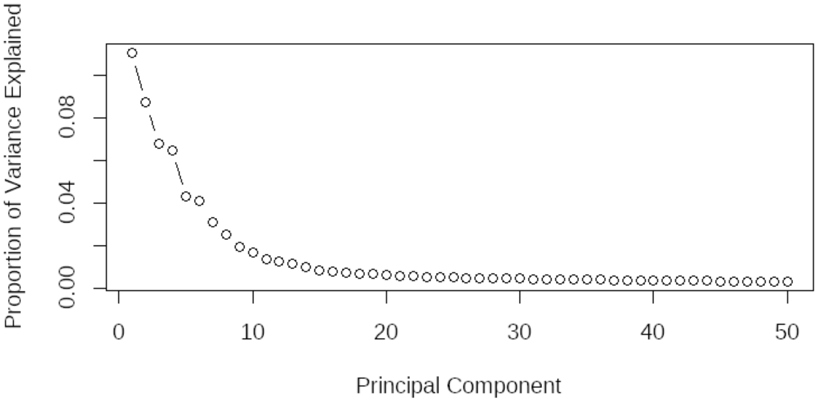

Figure 1: Scree plot of the first 50 out of 236 principal components

measures such as missclassification error, the gini index, or cross entropy. Aggregating trees trained in this manner gives a random forest model.[11]

3 Exploratory Data Analysis

3.1 Principal Component Analysis

To begin the exploratory stage of the analysis, we applied principal component analysis to the data. As shown in Figure 1, we determined that 20 components were sufficient. We selected more components than typically suggested by the traditional elbow method to ensure that at least 60% of the variance in the original features is explained by the chosen principal components.

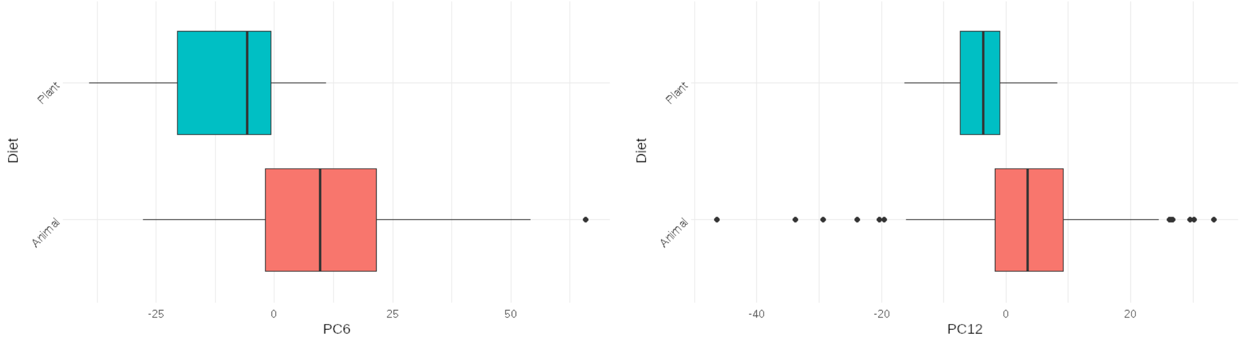

Figure 2: (left) Boxplot of the $6'th$ principal component by diet, (right) Boxplot of the $12'th$ principal component by diet

[11] Gareth James, et al., *An introduction to statistical learning*, vol. 112 (Springer, 2013).

	(Animal) Cluster 1	(Animal) Cluster 2
(Plant) Cluster 1	17	31
(Plant) Cluster 2	17	18

Table 2: Class agreements between the *GPCM* Model trained on the plant based samples vs the *GPCM* Model trained on the animal based samples

Next, we explore the relationships between the principal components the response. The clearest relationships between the principal components and diet are shown in figure 2.

3.2 Clustering

We used model-based clustering as a tool for exploratory analysis, specifically to examine whether clusters trained on plant-based samples differ from those trained on animal-based samples. Using the `gpcm` function from the `mixture` package[12], we found that training a GPCM model on plant-based samples yielded a fit with 2 components and a VVI covariance structure. Similarly, the model trained on animal-based samples produced a comparable fit; 2 components with a VVI covariance structure. As shown in Table 2, the cluster models demonstrate poor agreement with each other, further confirmed by the adjusted Rand index of just 0.01. This suggests that the microbiome profiles for individuals following a plant-based diet may differ from those following an animal-based diet.

4 Results

4.1 Random Forests

	Predicted: Animal	Predicted: Animal
True Label: Animal	36	2
True Label: Plant	4	28

Table 3: Class agreements between the *GPCM* Model trained on the plant based samples vs the *GPCM* Model trained on the animal based samples

We now perform classification using random forests. To evaluate the models, we use a

[12] Pocuca, Nik, et al. mixture: Mixture Models for Clustering and Classification. 2024, R package version 2.1.1. CRAN.R-project.org/package=mixture.

70-30 train test split. The random forest model was built using 4000 trees, with 3 variables used at for each decision. We see from table 3 that the random forest model performed well in classifying which samples were gathered from subjects who were on a plant based diet from the those who from an animal based diet. The prediction accuracy as 91 %, and the adjusted rand index as 0.68, which is evidence that the random forest model performs well in this problem.

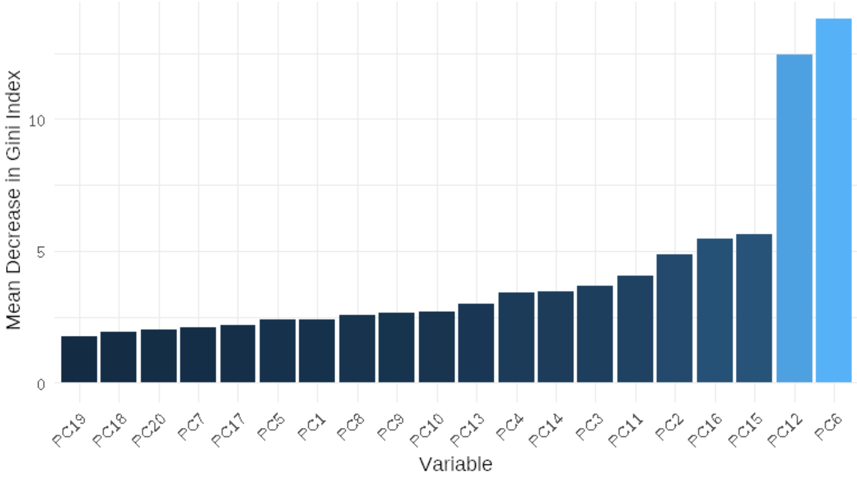

Figure 3: Variable importance plot from the random forest model

Figure 3, we see that the $6'th$ and $12'th$ principal components were the most important for predicting the subject's diet.

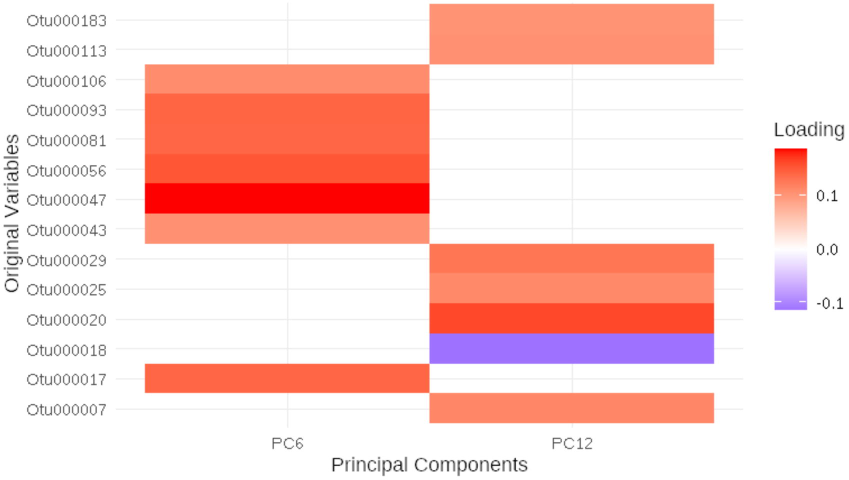

Figure 4: Heatmap of the loadings for the principal components that were at least 0.1 in magnitude

Since principal components are merely a linear combination of the original variables, we interpreted the results by identifying variables with sufficiently high loadings on each component. That is to say, we interpreted the principal components by observing the original set of variables that had the strongest relationships with the principal components. From figure 4, we see that we are able to observe a fairly parsimonious representation of the 6'th and 12'th principal components, each corresponding to their own group of 7 different taxa.

5 Discussion

So far we have used principal component analysis to reduce the dimension of our data, and observed potential relationships between the principal components and the response in figure 2. We then used parsimonious Gaussian mixture models to compare the microbiome profiles between subjects that were given a plant based diets and animal based diets. Finally we were able to get fairly accurate predictions for whether or not someone was on a plant based diet versus an animal based diet based on their microbiome profiles. We were also able to tell which of the specific microbes were most important by observing figure 4. The key drawback to this analysis is the fact that we did not draw any directional conclusions between the relationships between the taxa and the subjects diet as this would require generalized linear models. The reason why generalized linear models were out of the question of this analysis is due to the temporal aspect of the data. In order to feasably perform the analysis, we had to use principal component analysis to reduce the dimension of the data, in doing so, we had to use all the observations in the data set, thus the use of principal component analysis may be unsuitable since we are not taking into account the potential autocorrletion between the observations. Further analysis would require the use of dynamic factor models which take into account the autocorrelation present in the data, while reducing the dimension of the features present.

6 Appendix: Background on Micro-biome data

Historically, microbiology studies were almost entirely in vitro, and cultures were grown and study in controlled environments[13]. It was only up until 2005 wherein advances in DNA-sequencing technologies allowed researchers to study microbes directly extracted from a sample. The current gold standard approach in DNA-sequenceing is 16S rRNA gene sequencing, which works as follows[14]:

[13] Yinglin Xia, et al., *Statistical analysis of microbiome data with R*, vol. 847 (Springer, 2018).

[14] Tamang, Sanju. Amplicon sequencing: Principle, steps, types, uses, diagram. *Microbe Notes*, Dec. 2024. microbenotes.com/amplicon-sequencing/.

- Amplify: Replicate certain important regions (the 16S rRNA gene specifically) of the DNA strands
- Construct Libraries: collect DNA fragments cloned into vectors and store them within host organisms.
- Sequence: Use the libraries to construct DNA Sequences

Since there may be sequencing errors, we cluster similar sequences into taxonomic units and generate a table of counts showing how many of these taxonomic units (called taxa) appear in each sample. Thus, taxa can be thought of as a proxy for the different species present in a sample. Unfortunately, several statistical challenges arise when dealing with this data. These include having many more taxonomic units than samples, uneven sequencing depth (resulting from the sequencing technologies used rather than the sample itself), measurement error, non-homogeneity of varianec, zero inflation, correlated taxa, and more[15]. Many of the preprocessing steps used in microbiome analysis are specifically designed to address these challenges.

[15] Xia et al.

Works Cited

Anders, Simon, et al. “Count-based differential expression analysis of RNA sequencing data using R and Bioconductor”. *Nature protocols*, vol. 8, no. 9, 2013, pp. 1765–86.

Barrett, Tyson, et al. data.table: Extension of ‘data.frame‘. R package version 1.17.0, 2025.

Callahan, BJ, et al. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. F1000Res. 2016; 5: 1492. 2020.

David, Lawrence A, et al. “Diet rapidly and reproducibly alters the human gut microbiome”. *Nature*, vol. 505, no. 7484, 2014, pp. 559–63.

James, Gareth, et al. *An introduction to statistical learning*. Vol. 112, Springer, 2013.

Liaw, Andy, and Matthew Wiener. “Classification and Regression by randomForest”. *R News*, vol. 2, no. 3, 2002, pp. 18–22. CRAN.R-project.org/doc/Rnews/.

Love, Michael I., et al. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. *Genome Biology*, vol. 15, 12 2014, p. 550. https://doi.org/10.1186/s13059-014-0550-8.

McMacken, Michelle, and Sapana Shah. “A plant-based diet for the prevention and treatment of type 2 diabetes”. *Journal of geriatric cardiology: JGC*, vol. 14, no. 5, 2017, p. 342.

McMurdie, Paul J., and Susan Holmes. “phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data”. *PLoS ONE*, vol. 8, no. 4, 2013, e61217. dx.plos.org/10.1371/journal.pone.0061217.

Meyer, David, et al. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-16, 2024, CRAN.R-project.org/package=e1071.

Murphy, Kevin P. *Probabilistic Machine Learning: An introduction*. MIT P, 2022, probml.github.io/book1.

Pocuca, Nik, et al. mixture: Mixture Models for Clustering and Classification. R package version 2.1.1, 2024, CRAN.R-project.org/package=mixture.

Schloerke, Barret, et al. GGally: Extension to ’ggplot2’. R package version 2.2.1, 2024, CRAN.R-project.org/package=GGally.

Sonnenburg, Justin L, and Fredrik Bäckhed. “Diet–microbiota interactions as moderators of human metabolism”. *Nature*, vol. 535, no. 7610, 2016, pp. 56–64.

Tamang, Sanju. Amplicon sequencing: Principle, steps, types, uses, diagram. *Microbe Notes*, Dec. 2024. microbenotes.com/amplicon-sequencing/.

Wickham, Hadley, et al. “Welcome to the tidyverse”. *Journal of Open Source Software*, vol. 4, no. 43, 2019, p. 1686. https://doi.org/10.21105/joss.01686.

Wu, Gary D, et al. “Linking long-term dietary patterns with gut microbial enterotypes”. *Science*, vol. 334, no. 6052, 2011, pp. 105–08.

Xia, Yinglin, et al. *Statistical analysis of microbiome data with R*. Vol. 847, Springer, 2018.