Application of the Microbiome Taxonomic Longitudinal Correlation model to the American Gut Project

Emily Palmer

Spring 2021

Abstract

The Microbiome Taxonomic Longitudinal Correlation model, a recent model for microbiome association studies that incorporates taxonomic hierarchy, was explored in detail and applied to data from the American Gut Project. Using real datasets, which are larger than the case study provided in the initial paper, and which include incomplete taxonomic labels not pre-organized into a taxonomic hierarchy, required code to be developed and workflows defined for intermediate steps. Datasets that include more than a small subset of OTUs run into computational limitations using this model.

1 Introduction

1.1 Microbiome data

The human gut contains millions of microbes. Research on microbiome sequencing data has provided numerous insights about human health. Studies have associated the human gut microbiome to inflammatory bowel disease, cancers, diabetes, and obesity, and have shown the existence of bi-directional interactions between the gut and the central nervous system [3, 7]. These studies show the human microbiome to be a promising field of research.

Microbiome sequencing studies commonly focus on 16S ribosomal ribonucleic acid (16s rRNA). This gene is a useful marker for microbe identification. Microbiome sequencing

data commonly consist of the count of reads on operational taxonomic units (OTUs) for each sample. Operational taxonomic units refer to the bins in which sequences are grouped, and are frequently mapped back to taxonomic labels from existing libraries. Methods for microbiome analysis can either be focused on a single OTU or multiple OTUs simultaneously.

There are several aspects common to microbiome OTU data that present obstacles to analysis using standard statistical approaches. The number of OTUs is often very large compared to the sample size. The value of the count for a given OTU is dependent on the sampling depth for the given sample (the sum of all OTUs across the sample). This sampling depth is not originally constant across samples, so normalization is necessary before counts can be compared across samples. This can be done by rarifying, or randomly sampling OTUs to standardize the sampling depth across samples. Often, OTU counts are converted into a proportion to represent the relative abundance of that OTU in the sample.

Additionally, OTU data are very sparse, with many samples having zero counts across multiple OTUs. Thus microbiome data contains excessive zeros, which must be accounted for in the analysis. Common adjustments are either to use models based on zero-inflated distributions, or two-part models that separately model the presence of an OTU and the abundance (or relative abundance) of present OTUs.

1.1.1 Taxonomic tree for OTUs

Two types of trees can arise in microbial studies: phylogenetic trees and taxonomic trees. Phylogenetic trees are an estimate of the evolutionary history between organisms by splitting of a tree to represent shared ancestors. Taxonomic trees represent the hierarchy of shared labels of different taxonomic levels: Kingdom, Phylum, Class, Order, Family, Genus, and Species. Ideally, these two trees should be equivalent, but this requires explicit naming and organizing of each bacterial species, which is an imposing and incomplete task. The taxonomic labeling system is coarse for microbial taxa, and new discoveries and changing naming systems will change its structure [9]. This project will focus on taxonomic trees, because these are the tree types utilized in the model employed in this report.

When association studies are performed using multiple OTUs, it can be reasonable to use prior knowledge on the given evolutionary or hierarchical organizational relationship between multiple OTUs. Potentially, evolutionally relationships correspond to similar relationships between OTU expression and chosen covariates. If a certain OTU is found to be associated with a covariate, we might expect another OTU that is close on the taxonomic tree to also be associated with a covariate, and associated in a similar way. Broadly, we want to use the knowledge of the placement of a given OTU on the taxonomic tree to provide insight into correlations that exist between different read counts present in samples. Hopefully this correlation can be helpful in modeling overall association with predictors of interest.

1.2 Generalized Estimating Equations

Generalized Estimating Equations (GEEs) were introduced to extend generalized linear models to analysis of longitudinal data [5]. They are a useful tool for analyzing correlated data, such as can arise from longitudinal data, but also more generally for data with correlations from any sort of repeated measure. GEEs estimate population average effects by assuming a within cluster correlation structure.

First, we introduce some notation. A cluster $\mathbf{y}_k$ is the collection of points from repeated measures $\mathbf{y}_k = (y_{k,1}, \ldots, y_{k,n_j})$, taken from a subject or individual k where $k = 1, \ldots, K$ represents the index of clusters/individual, and $j = 1, \ldots, n_k$ represent the different repeated measures taken, for instance, on different time points, sample sites, or OTU measurements, or any other given repeated measurement taken on a cluster. The concept of a cluster will depend on the data setting, for instance we could even consider a family of related individuals as the cluster designation. For instance, consider the response variable for a family k, measured on two family members (f_1, f_2), on 2 days (t_1, t_2) and 2 OTU measurements (o_1, o_2). Then $\mathbf{y}_k = (y_{k,f_1,t_1,o_1}, y_{k,f_1,t_1,o_2}, y_{k,f_1,t_2,o_1}, y_{k,f_1,t_2,o_2}, y_{k,f_2,t_1,o_1}, y_{k,f_2,t_1,o_2}, y_{k,f_2,t_2,o_1}, y_{k,f_2,t_2,o_2})$.

We additionally have a $p \times 1$ vector of covariates x_{kj} measured for the jth value on each cluster k. The GEE model links the mean of the responses to the covariates through the

regression equation

$$g(\mu_{kj}) = \mathbf{x}_{kj}^T \boldsymbol{\beta}$$

where $E(y_{kj}) = \mu_{kj}$ and $g(\cdot)$ is a known link function. The variance of the responses is $\text{Var}(Y_{kj}) = \phi a_{kj}$, where ϕ is the dispersion parameter and a_{jk} is a known variance function determined by the distribution of the data. Let $\mathbf{R}_k(\boldsymbol{\alpha})$ be the specified working correlation matrix, which is specified by the distinct values of $\boldsymbol{\alpha}$, that describes the assumed within-cluster correlation structure. Define

$$\mathbf{V}_k = \phi \mathbf{A}_k^{1/2} \mathbf{R}_k(\boldsymbol{\alpha}) \mathbf{A}_k^{1/2}$$

as the working covariance matrix of $\mathbf{y}_k$, where $\mathbf{A}_k$ is the diagonal matrix consisting of the values of a_{kj}. Then, $\hat{\boldsymbol{\beta}}$ is the solution to the Generalized Estimating Equation

$$\sum_{k=1}^{K} \frac{\partial \boldsymbol{\mu}_k^T}{\partial \boldsymbol{\beta}} \mathbf{V}_k^{-1} (\mathbf{y}_k - \boldsymbol{\mu}_k) = 0.$$

The above equation depends on the values of $\boldsymbol{\alpha}$ and ϕ, so these must additionally be estimated. The solution to the above equation is found using an iterative formula that switches between estimating $\hat{\boldsymbol{\beta}}$ using given fixed $\hat{\boldsymbol{\alpha}}, \hat{\phi}$ and estimating $\boldsymbol{\alpha}, \phi$ using given fixed $\boldsymbol{\beta}$. GEEs work without specifying the joint distribution of observations, similar to quasi-likelihood approaches.

One useful characteristic of GEEs is that the parameter estimation of $\boldsymbol{\beta}$ is consistent, even if the working correlation matrix is misspecified. The estimate $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance

$$\Big(\sum_{k=1}^{K} \frac{\partial \boldsymbol{\mu}_k^T}{\partial \boldsymbol{\beta}} \mathbf{V}_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}^T}\Big)^{-1} \Bigg(\sum_{k=1}^{K} \frac{\partial \boldsymbol{\mu}_k^T}{\partial \boldsymbol{\beta}} \mathbf{V}_k^{-1} \text{COV}(\mathbf{y}_k) \mathbf{V}_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}^T} \Bigg) \Big(\sum_{k=1}^{K} \frac{\partial \boldsymbol{\mu}_k^T}{\partial \boldsymbol{\beta}} \mathbf{V}_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial \boldsymbol{\beta}^T}\Big)^{-1}$$

where $\text{COV}(\mathbf{y}_k)$ is the true covariance matrix of $\mathbf{y}_k$.

If we replace $\boldsymbol{\beta}, \phi, \boldsymbol{\alpha}$ and $\text{COV}(\mathbf{y}_k)$ with $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\phi}$ and $(\mathbf{y}_k - \boldsymbol{\mu}_k)(\mathbf{y}_k - \boldsymbol{\mu}_k)^T$ this gives us

the sandwich estimator, a constant estimator for the covariance of $\boldsymbol{\beta}$.

1.2.1 Implementing GEEs in R

The `geepack` package [2] in R [8] provides a useful way to implement GEEs with a user defined working correlation structure, denoted `zcor`. The function `geeglm()` fits a GEE from the specified regression formula, link function, and correlation structure, similarly to the standard `glm()` function call.

There are four main pre-specified working correlation matrix structures in `geeglm`: independence, exchangeable, AR1, and unstructured. Specifying a user-defined correlation structure is also possible, although seems much less common in applications than using the pre-defined structures. The format of the user-defined correlation structure `zcor` is different than the working correlation matrix defined above the GEE framework. The GEE working correlation matrix $R_k(\alpha)$ is a square matrix that has the same dimension as the length of $\mathbf{y}_k$ for the kth cluster. In contrast, `zcor` uses the upper diagonal correlation parameters of the working correlation matrix $\mathbf{r}_k = (r_{k,1,2}, r_{k,1,3}, \ldots, r_{k,1,n_k}, r_{k,2,3}, \ldots, r_{k,n_{k-1},n_k})$ of $\mathbf{R}_k(\boldsymbol{\alpha})$. The `zcor` $\mathbf{Z}_k$ matrix for cluster k comes from $\mathbf{r}_k = \mathbf{Z}_k\boldsymbol{\alpha}$. The overall `zcor` matrix comes from combining these matrices $(\mathbf{Z}_1^t, \ldots, \mathbf{Z}_K^t)^t$. This matrix will have dimension $\sum_{k=1}^{K} \binom{n_k}{2} \times a$, where a is the length of $\boldsymbol{\alpha}$.

2 Microbiome Taxonomic Longitudinal Correlation (MTLC) model

The Microbiome Taxonomic Longitudinal Correlation (MTLC) model was introduced in 2020 by Chen and Xu[1] as a way to estimate correlations between OTUs as well as associations between OTU responses and covariates, that can be used on longitudinal and repeated measure data. Coefficients are estimated using a two part GEE model that uses a correlation structure based on OTU taxonomic hierarchy as well as longitudinal and repeated measure correlations.

Consider again the cluster $\mathbf{y}_k$ described above, from measurements on two family mem-

bers (f_1, f_2), on two days (t_1, t_2) on two OTUs (o_1, o_2). The observations in $\mathbf{y}_k$ are not independent. There exist possible correlations between the family members, time points, as well as possible correlations between the OTUs measured, and any combination. The following sections will explain how these types of correlations are incorporated into GEE models.

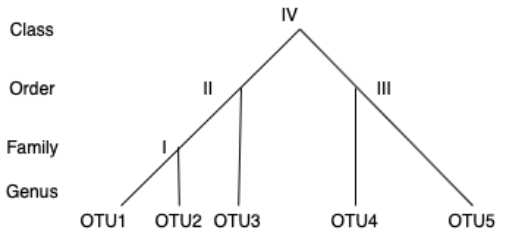
2.1 Correlation structure

2.1.1 Taxonomic correlation

Previous studies considering the correlations between OTUs have used a Dirichlet multinomial distribution [4], which assumes a negative correlation between OTUs. However, the true correlation structure between OTUs can be positive or negative [6], and we want to build models that can account for this flexibility depending on the nature of how given OTUs interact.

The MTLC model uses the taxonomic information hierarchy of each OTU to specify the OTU correlation component, denoted Γ, of the working correlation matrix used in GEE estimation. In order to reduce the complexity from an unspecified correlation matrix, where each OTU has a different correlation (which becomes infeasible to model), we impose the assumption that OTUs have the same correlation if the first shared higher level taxon is the same. Determining the unique correlations will depend on the structure of the taxonomic tree. We will then represent the correlation information from the taxonomic tree into a correlation matrix that keeps track of which correlations are the same.

A taxonomic tree can be represented as a list containing counts of how many OTUs belong each subgroup of successive labels. The highest level should contain all N OTUs, and the lowest level should contain N groups, each with 1 observation. Consider a taxonomic tree that represents the class, order, family, and genus taxonomic labels for five OTUs as illustrated below. The list `list(5, c(3,2), c(2,1,1,1), c(1,1,1,1,1))` represents this tree, as the single class contains all 5 OTUs, the first order group contains 3 OTUs, and the



$$\Rightarrow \Gamma = \begin{pmatrix} 1 & I & II & IV & IV \\ I & 1 & II & IV & IV \\ II & II & 1 & IV & IV \\ IV & IV & IV & 1 & III \\ IV & IV & IV & III & 1 \end{pmatrix}$$

second contains 2, and so following.

The roman numerals represent the distinct correlations we assume for the working correlation matrix. In this example, correlation I represents OTUs in the same family, but different genus, II and III represent same orders, but different families, and IV represents different orders but the same class.

2.1.2 Longitudinal and repeated measure correlation

Additional correlations in the data occur from repeated measures. A common form of repeated measure comes from measurements at different discrete time points. Auto-regressive, Toeplitz, unstructured, exchangeable, and independent are all common correlation structures that can be can be used on longitudinal data. Each of these structures gives a working correlation matrix that indicates the distinct correlation between points. Any of these can be used for longitudinal repeated measures in this model. If there are additional repeated measures in the data besides those from longitudinal measures, the repeated measure working correlation matrix is found by taking all combinations of time points and repeated samples into one repeated measure working correlation matrix. If we have T time points and S different additional repeated measures, the repeated measure working correlation matrix Ω has dimension $(T \cdot S) \times (T \cdot S)$

2.1.3 Integrative correlation matrix

Specifying the working correlation matrix to use in GEE estimation combines both the taxonomic correlation matrix and repeated measure correlation matrix. The integrative correlation matrix R is a block matrix containing entries $\Omega(\Gamma_{ab})$, where $a, b : 1, \ldots, N$. If we have N different OTUs, and L different repeated measures, the integrative correlation

matrix R has dimension $(N \cdot L) \times (N \cdot L)$, where every sub-matrix in R represents the distinct correlations representing the given time/repeated measure/OTU correlation.

For example, if we have data with the OTU structure above, and two time points, which we assume are correlated, the integrative correlation matrix would be

$$R = \begin{pmatrix} 1 & 2 & 3 & 5 & 5 & 6 & 7 & 8 & 10 & 10 \\ 2 & 1 & 3 & 5 & 5 & 7 & 6 & 8 & 10 & 10 \\ 3 & 3 & 1 & 5 & 5 & 8 & 8 & 6 & 10 & 10 \\ 5 & 5 & 5 & 1 & 4 & 10 & 10 & 10 & 6 & 9 \\ 5 & 5 & 5 & 4 & 1 & 10 & 10 & 10 & 9 & 6 \\ 6 & 7 & 8 & 10 & 10 & 1 & 2 & 3 & 5 & 5 \\ 7 & 6 & 8 & 10 & 10 & 2 & 1 & 3 & 5 & 5 \\ 8 & 8 & 6 & 10 & 10 & 3 & 3 & 1 & 5 & 5 \\ 10 & 10 & 10 & 6 & 9 & 5 & 5 & 5 & 1 & 4 \\ 10 & 10 & 10 & 9 & 6 & 5 & 5 & 5 & 4 & 1 \end{pmatrix},$$

where each distinct integer represents a unique correlation. We can write $R = R(\alpha)$, where $\alpha = 1 : 10$. For instance, correlation 6 represents the correlation between OTU_1 and OTU_2 between the two time points sampled.

2.2 Two-part GEE model

The association between OTUs and covariates is modeled in two parts. Since OTU data contains excessive zeros, the presence/absence of an OTU ($y_{kj}^{(0)}$)is modeled separately from the logarithm of the non-zero relative abundance of the OTU ($y_{kj}^{(+)}$). We assume each OTU observation y_{kj} follows a mixture of Bernoulli and log-normal distributions. $P(y_{jk}^{(0)} = 1) = \mu_{kj}^{(0)}$, and $y_{kj}^{(+)} \sim N(\mu_{kj}^{(+)}, \sigma^2)$

To model these two responses separately, we specify the logit link function for the present/absent outcomes, and the identity link for the non-zero log transformed relative abundances, where the parameter estimates $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(+)}$ are found using the GEEs described above.

$$\log \frac{\mu_{kj}^{(0)}}{1 - \mu_{kj}^{(0)}} = \mathbf{x}_{kj}^t \boldsymbol{\beta}^{(0)} \quad \text{and} \quad \mu_{kj}^{(+)} = \mathbf{x}_{kj}^t \boldsymbol{\beta}^{(+)}$$

3 Application to the American Gut Project

Applying the MTLC to the American Gut Project data made evident some missing intermediate steps needed for the application of this model. The MTLC paper described above [1] provided code for their simulation study and calculating the integrative correlation matrix from a given list representing the taxonomic structure (https://github.com/chenbo4/GEE). However, no code was provided for applying this method to actual datasets, or how to generate the taxonomic list for a larger dataset where creating this list cannot be done by hand. Part of the work done in this project has been writing general code for using this model for new datasets, and filling out functionality for necessary intermediate steps to transform the data into an appropriate form for use of this model.

All of the examples introduced in this report, as well as the examples, simulation, and application worked through in the original paper, have a limited set of OTUs included. However, most microbiome datasets have measurements on the scale of thousands of different OTUs. The question then arises, how will the MTLC model scale to a more realistic dataset?

To explore the scalability of the MTLC model, we examined a subset of the American Gut Project [7]. The American Gut Project data consists of citizen-science self-selected individuals mailing in fecal samples for microbiome sequencing. Along with microbiome data, it contains sample information from a broad health survey, containing self-reported demographic, dietary, medical, and overall health information. Only observations at one time point were considered, so the working integrative correlation matrix will only consist of the taxonomic correlations (no longitudinal/repeated measures).

Before attempting to apply the MTLC model on the American Gut data, some initial filtering must be done for quality of the data. This consisted in filtering out samples with low sampling depth (total reads < 500), and only including samples from single sample site (fecal). Additionally OTUs were filtered include those in the kingdom Bacteria. OTUs were combined to the genus level, and only OTUs with genus present in 10% of samples were

included.

3.1 Adjusting the correlation matrix for incomplete clusters

The above explanation of how to build the working correlation matrix assumes that all clusters have equal size/length. In other words, each individual/cluster needs to have the same number of OTUs measured and the same number/type of repeated samples. However, in practice this is rarely the case; each individual may not be measured the same number of times, and each OTU is rarely present each sample. Even if we did have balanced clusters in the presence/absence model, due to the sparse nature of OTUs, it is likely that the relative abundance model will have essentially 'missing' data for OTUs that are not present in a given sample, resulting in unequal cluster sizes.

The function `geeglm()` can account for missing (`NA`) data values in the response or covariate columns. However, both the integrative correlation matrix and the `zcor` matrix must be adjusted to remove the corresponding entries. For each cluster $k = 1, \ldots, N$, we make this adjustment by finding the expected OTU/repeated measure/time combinations that are missing in the given cluster, replacing all entries of the corresponding row and column of the working integrative correlation matrix with a replacement dummy value (e.g. -2). The corresponding $\texttt{zcor}_k$ matrix was created using the upper triangular values of R_k, and then filtered to remove all rows that contained the dummy value. This was done for each cluster, and then the overall `zcor` matrix was made up of the resulting adjusted cluster matrices.

3.2 Correlations for OTUs without full taxonomic level labeling

Moving beyond small examples, there are many instances when the complete set of taxa labels are not available for OTU for each level. The existing set of taxonomic labels for the microbiome may not perfectly describe the entire set of microbial organisms down to the species level. The taxonomy associated with OTU data may therefore only have labels for the Kingdom, Phylum, Class or Order, or fewer.

We then must decide how and when to treat OTU values with any number of missing taxonomic level labeling distinct, and how to categorize the correlations. If the values are not distinct, we should combine the OTUs into one observation. We additionally need to be able to compute the taxonomic tree structure, and compute the distinct taxonomic working correlations based on this half-labeled information.

For example, consider four OTUs with measured counts, shown on the left table below. Which (or any) of these OTUs should be consolidated into one observation? What distinct correlations should we estimate?

OTU	class	order	family	value
OTU_1	A	B	C	v_1
OTU_2	A	B		v_2
OTU_3	A			v_3
OTU_4	A			v_4

$\Rightarrow$

OTU	class	order	family	value
OTU_1	A	B	C	v_1
OTU_2	A	B	B	v_2
$OTU_3 + OTU_4$	A	A	A	$v_3 + v_4$

To calculate the taxonomic tree when there are missing taxa levels, we essentially grow down the tree. If an OTU is missing a label for a certain level, it will be given a placeholder label of the lowest taxon. If two or more OTUs share the same label at the lowest level, they will be consolidated into one observation.

In the above example, there are three distinct taxa labels at the lowest level (family). Thus OTU_3 and OTU_4 will be collapsed into one observation, and there will be two correlation parameters; the correlation between OTU_1 and OTU_2 and that between OTU_1 or OTU_2 and $OTU_3 + OTU_4$.

3.3 Generating taxonomic correlation from microbiome data

A detailed tree showing the hierarchy of taxonomic labels is not generally provided with OTU data. Instead, there is often a column listing the given taxonomic labels. To specify the working correlation matrix, a tree, or equivalently the number of taxa in each grouping, must be recreated from the list of labels.

This is done iteratively in the function `tax2cor` written for this analysis. This function

first grows down the tree to fill in missing taxa labels, and then adds up non-unique lowest labels for a chosen lowest level. For this project, OTUs were aggregated to the genus level. Then, starting at the top taxa level (Kingdom), the number of OTUs in each taxa-level group is counted. This relies on sorting the taxa alphabetically, grouping successively, so clusters in the dataset should share this same ordering. This results in a list detailing the counts of OTUs belonging to each level. This list can be used to generate the taxonomic correlation working matrix.

3.4 Computational challenges

When using `geeglm`, the user-defined working correlation structure is stored as a Matrix data structure. In R, matrices are indexed using integers, and the maximum length of a vector is $2^{31} - 1$. This means that the total number entries in the `zcor` matrix cannot exceed this limit. If we have K individuals, measured on t time points for N OTUs, and α has length a, this must satisfy the inequality $K\binom{tN}{2} \times a < 2^{31} - 1$.

Consider the case when we have 200 subjects, and 200 OTUs measured, on 4 time points. Suppose there are 50 distinct correlation arising from the integrative correlation matrix. The `zcor` matrix in this seemingly small example would have $3.2 \times 10^9 > 2^{31} - 1$ entries, and would exceed this limit. This is only the very first initial step of initializing the working correlation matrix. The computation involved in actually solving the GEE for situations where the working correlation matrix becomes large is limited by memory of the computer performing the calculation. The full dataset from the American Gut Project will result in a `zcor` matrix too large to even initialize.

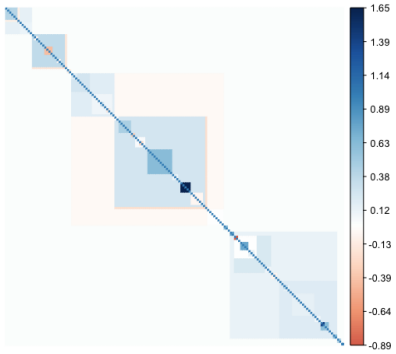
3.5 Results on a limited subset of the American Gut project

One attempt to limit the computational challenge was to limit the number of samples. Selecting only 100 of the original 4827 samples, results in a corresponding 164 distinct OTUs. The integrative correlation matrix, based only on taxonomic information - no longitudinal measures here, results in 55 distinct taxonomic correlations to estimate. Fitting both model

parts is time consuming and requires overnight computation. The relative abundance normal model takes over three hours to run, and the presence/absence model failed to complete after two days of running.

This model was fit with antibiotic usage as the single covariate, as we might expect antibiotic usage to have a significant effect on the microbiome. The antibiotic usage variable is one if the subject had used antibiotics in the past year and zero otherwise. The relative abundance model found no significant effect of antibiotic use in the past year on positive log-transformed relative abundance of OTUs, $\beta^{(+)} = 0.068, p = 0.35$.

The following heat map shows the estimated correlations. Notice the majority of the correlations are positive. For example, the correlation block at the very top left represent correlation between OTUs with order *Actinomycetales*, but different families. Correlations closest to the diagonal are higher, which represent correlations between OTUs closer on the taxonomic tree. However, it is concerning to see two estimated correlations above one. It is unclear if this is an error in `geepack` code or somehow from unusual artifacts in the data.



3.6 Results on individual classes of the American Gut project

Instead of focusing on the entire taxonomic hierarchy, and a subset of samples, we alternatively focus on a subset of the taxonomic hierarchy and all samples. From the heat map above, we saw that correlations between different classes were near-zero. Thus instead of building the taxonomic tree starting from the Kingdom level, and then building the working correlation matrix on the entire microbiome, we separately model each class of OTUs, looping through the 19 different classes included. This process took less than an hour for both models. Another benefit of this is approach is interpretability. Instead of results for

antibiotic use across the entire microbiome, we now see the population average effect of antibiotics on a single class. This gives us a more specific focus for any biological interpretation or future studies. The correlation matrices from this approach will depend on the number of OTUs included in each class, which ranges from 1 to 45.

The following table shows the estimated effects of antibiotic use as well as the p-values adjusted using FDR for multiple comparisons for the presence/absence model and the relative abundance model. We see in particular that in classes where the association with antibiotic use is significant, the estimated slope coefficient is negative, indicating a decrease in overall presence of the OTUs in this class. This corresponds to what we would expect from antibiotic.

Class	$\hat{\beta}^{(0)}$	p-value	$\hat{\beta}^{(+)}$	p-value+
Actinobacteria	-0.0290	0.4221	0.0129	0.5940
Coriobacteriia	-0.1251	0.0002	-0.0119	0.5090
Bacteroidia	-0.1170	0.0000	0.0306	0.3555
Bacilli	-0.0171	0.5602	-0.0420	0.0217
Clostridia	-0.0845	0.0000	-0.0223	0.0673
Erysipelotrichi	0.0347	0.2119	-0.0186	0.2888
Fusobacteriia	0.0302	0.6265	-0.0531	0.1656
Alphaproteobacteria	-0.3106	0.0000	0.2277	0.0000
Betaproteobacteria	-0.1281	0.0000	0.1570	0.8935
Deltaproteobacteria	-0.3463	0.0000	-0.0023	0.9170
Gammaproteobacteria	0.0147	0.6265	-0.0624	0.0009
Flavobacteriia	-0.2183	0.0790	-0.0261	0.0831
4C0d-2	-0.6903	0.0000	-0.1466	0.0000
Chloroplast	-0.1454	0.0860	-0.0353	0.1008
Lentisphaeria	-0.4641	0.0000	-0.0728	0.0000
Epsilonproteobacteria	-0.1138	0.2119	-0.0242	0.2276
Mollicutes	-0.4588	0.0000	-0.1104	0.0000
RF3	-0.1415	0.1603	-0.0253	0.1656
Verrucomicrobiae	-0.4137	0.0000	-0.0639	0.0001

4 Discussion and Conclusion

A new model for microbiome association studies that estimates predictor effects on OTUs and correlations between OTUs based on taxonomic structure was explored and applied to real data. This MTLC method differs from other methods that aim to identify the OTUs that are associated with a covariate of interest. Instead it has a two-part goal: quantify

the overall association between a covariate of interest and the entire sampled microbiome, and also estimate correlations between the OTUs present. If an association is present, this method does not identify the taxa driving this association. Functionality of the original paper was improved and added to, including calculating the taxonomic tree from a list taxa labels, which might be blank for some levels.

If we build this model using the full taxonomic hierarchy, the MTLC model assumes there will only be one regression parameter for each covariate, giving the impact on the microbiome as a whole. This is likely too restrictive an assumption, and severely limits interpretations. If we have a true significant effect of antibiotics on a subset of the microbiome, it is possible it might not be significance in the overall model, and we can't tell which subsets of the microbiome are impacted.

Additionally, the current implementation of fitting GEEs is limited by the size of the working correlation matrix, which becomes too large when we use the entire taxonomic hierarchy. If we want to continue to use the model this way, perhaps to focus on modeling the correlation structure instead of predictor effects, we would need to find or create an alternative to the current use of the user-defined `zcor` matrix for the working correlation matrix in the `geeglm` package.

This model appears to work best when we instead focus on a subset of the taxonomic hierarchy. The computation time is reasonable, even for datasets with thousands of samples. We additionally are able to identify specific classes where antibiotic use is associated with a decrease in the abundance of OTUs in that class. Future work should go explore how or if choice of taxonomic level to use, for example phylum or order instead of class, influences results.

The performance of this model should also be re-evaluated in cases when we have more than two repeated measures and two OTUs. The MTLC model was previously shown to outperform other models in simulation studies in this two repeated measure and two OTU setting. It would be helpful to know if the MTLC would continue to outperform other models

before attempting to re-implement GEE code in R with large working correlation matrices.

Additionally, future work could find a way to incorporate phylogenetic information instead of taxonomic information. The phylogenetic tree contains an estimate of how close OTUs are evolutionarily, which is not directly present in taxonomic trees. This might generate a correlation structure closer to the true correlations between OTUs.

References

[1] Bo Chen and Wei Xu. "Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures". In: *PLoS computational biology* 16.9 (2020), e1008108.

[2] Ulrich Halekoh, Søren Højsgaard, and Jun Yan. "The R Package geepack for Generalized Estimating Equations". In: *Journal of Statistical Software* 15/2 (2006), pp. 1–11.

[3] James M Kinross et al. "The human gut microbiome: implications for future health care". In: *Current gastroenterology reports* 10.4 (2008), pp. 396–403.

[4] Patricio S La Rosa et al. "Hypothesis testing and power calculations for taxonomic-based human microbiome data". In: *PloS one* 7.12 (2012), e52078.

[5] Kung-Yee Liang and Scott L Zeger. "Longitudinal data analysis using generalized linear models". In: *Biometrika* 73.1 (1986), pp. 13–22.

[6] Siddhartha Mandal et al. "Analysis of composition of microbiomes: a novel method for studying microbial composition". In: *Microbial ecology in health and disease* 26.1 (2015), p. 27663.

[7] Emeran A Mayer, Kirsten Tillisch, Arpana Gupta, et al. "Gut/brain axis and the microbiota". In: *The Journal of clinical investigation* 125.3 (2015), pp. 926–938.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org/`.

[9] Alex D Washburne et al. "Methods for phylogenetic analysis of microbiome data". In: *Nature microbiology* 3.6 (2018), pp. 652–661.