Masters Project Report

Emily Palmer

Spring 2021

Abstract

The Microbiome Taxonomic Longitudinal Correlation, a recent model for microbiome association studies, was explored in detail and code was developed for application in general datasets beyond the case study provided in the initial paper. Applications of this method provided computational challenges when the scale of the data is beyond a small number of selected OTUs.

Contents

1 Introduction

1.1 Microbiome data

The human gut contains millions of microbes. Genomic research on microbiome sequencing data has provided numerous insights to human health. Studies have associated the human gut microbiome to irritable bowel syndrome, inflammatory bowel disease, cancers, diabetes, and obesity, and shown the existence of bi-directional interactions between the gut and the central nervous system [3], [5]. These studies show the human microbiome to be a research rich and important area of of focus. Characteristics of this data make this data difficult to analyze using standard statistical methods, so there is also needed work to develop or alter the statistical approaches used for this type of data.

Microbiome data commonly arises from 16S ribosomal ribonucleic acid RNA (16s rRna) sequencing studies. This gene is a useful marker for microbe identification. Microbiome sample can be collected from fecal samples, but oral, skin, vaginal, etc are also common areas (CITE). Microbiome sequencing data commonly consists of the count of reads on operational taxonomic units (OTUs) for each sample. Operational taxonomic units refer the bins in which sequences are grouped, but can frequently be mapped back to taxonomic labels from existing libraries. Methods for microbiome analysis can either be focused on a single OTU or multiple OTUs simultaneously.

1.1.1 Characteristics of microbiome data

OTU data present many obstacles to analysis using standard statistical methods.There are several aspects common to microbiome data that violate the assumptions of common statistical tests and regression models.

Frequently, the sampling depth (FINISH) is not consistent across samples. One technique to adjust for this is rarifying (CITE, FINISH)

The count for a given OTU does not necessarily represent the overall count. In other words, these counts represent relative abundance, not absolute abundance. OTU data is thus compositional in nature. Often, OTU counts are converted into a proportion to represent the relative abundance of that OTU in the sample.

Additionally, OTU data is very sparse. The library size of different samples can vary. For many OTUs, the total reads is zero. Thus microbiome data contains excessive zeros, which much be accounted for in the analysis. Common adjustments, or ways to take into account this, are using zero inflated models, or two-part models that separately model the presence of an OTU, and then next the abundance of present OTUs.

SECTION ON LONGITUDINAL STUDIES

1.1.2 Taxonomic tree for OTUs

From sequencing information, two types of trees that can arise in microbial studies are phylogenetic trees and taxonomic trees. Phylogenetic trees estimate the evolutionary history between organisms by splitting of a tree to represent shared ancestors. Taxonomic trees represent the hierarchy of shared labels of different classified taxa levels (Kingdom, Phylum, Class, Order, Family, Genus, Species). Ideally, these two trees should be equivalent, but this requires explicit naming and organizing of each bacterial species, which is a imposing and incomplete task. The taxonomic labeling system is coarse for microbial species, and new discoveries and changing naming systems will change its structure. Taxonomic trees will be focused on in this project.

[7]

When association studies are performed using multiple OTUs, it could be reasonable to use prior knowledge on the given evolutionary or hierarchical organizational relationship between multiple OTUs. Potentially, evolutionally relationships might correspond to similar relationships between the OTU expression, and chosen covariates. In other words, 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) are a useful tool for correlated data, such as those that arise from longitudinal data. GEEs were introduced to extend generalized linear models to longitudinal data [4]. While initially aimed for correlations arising from longitudinal data, GEEs can be useful for data with any sort of correlations, longitudinal, repeated measure, etc (FINISH).

First, we introduce some notation. Consider a response variable y_{it}, and a $p \times 1$ vector of covariates x_{it}, where $i = 1, \ldots, K$ represents the index of clusters, and $t = 1, \ldots, n_i$ represent the number of time points, or more generally, different values, or repeated measures in an individual cluster. GEEs work without specifying the joint distribution of observations, similarly to quasi-likelihood approaches.

Consider a response $\mathbf{y}$ consisting of K different clusters, where clusters consist of repeated measures (perhaps time points, or different OTUs) on the same individual. Then, we write

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

where μ_k is the mean of the ys, and $g()$ is a known link function that relates the mean to the response.

$\boldsymbol{\beta}$ is then the solution to the Generalized Estimating Equation

$$U(\boldsymbol{\beta}) = \sum_{k=1}^{K} \dots$$

finish typeseting...

Iterative formula switching between estimating $\hat{\beta}$ and $\hat{\alpha}, \hat{\phi}$....

One useful characteristic of GEEs is that the parameter estimation of β is consistent, even if the working correlation matrix is misspecified.

1.3 GEEs in R

The `geepack` package[2] in R [6] provides a useful way to implement GEEs with a user defined working correlation structure. The function `geeglm()` fits the GEE from the specified regression formula, link function, and correlation structure.

The format of the correlation structure that needs to be specified is different than the working correlation matrix.

Workign with Yuan, Incorporating taxonomic correlation structure in generalized estimating equations

for association studies.

2 Microbiome Taxonomic Longitudinal Correlation (MLTC) model

The Microbiome Taxonomic Longitudinal Correlation (MLTC) model was introduced in 2020 by Chen and Xu[1] as a way to estimate correlations between OTUs and associations between a predictor an, 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 correlations and longitudinal and repeated measure correlations.

2.1 Model Data structure

Commonly, OTU data is stored in two tables, an OTU table and a metadata table. The rows of the OTU table are the OTU ids, and the columns are the samples. In the metadata table, the rows are samples.

Consider a dataset with K individuals. Each individual k, $k = 1, \dots K$ has n_k different samples taken from it, perhaps at different points in time. In total there are $M = \sum_{k=1}^{K} n_k$ samples taken. The collection of observations taken on an individual can be referred to a cluster.

In this model, the response is the OTU value (either presence absence, or transformed relative abundance). Both the response, and predictors, must be transformed into a ”long” format. If there are N OTUs recorded,

The response $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_K)^t$, where for a given individual/cluster $\mathbf{y}_k = (y_{k,1}, \dots, y_{k,n_k})$. The vector of covariates for a given k, j has length p.

EXAMPLE. Consider a small example, where there is only one subject, measured at two time points, for two OTUs. The OTU table would have the form:

	S_1, T_1	S_1, T_2
OTU_1	y_{S_1,T_1,OTU_1}	$y_{S1,T2,OTU1}$
OTU_2	$y_{S1,T1,OTU2}$	$y_{S1,T2,OTU2}$

Then the transformed data would have the form:

$$\mathbf{y} = \begin{pmatrix} y_{S1,T1,OTU1} \\ y_{S1,T1,OTU2} \\ y_{S1,T2,OTU1} \\ y_{S1,T1,OTU2} \end{pmatrix} \qquad \mathbf{x} = \begin{pmatrix} x_{S1,T1} \\ x_{S1,T1} \\ x_{S1,T2} \\ x_{S1,T2} \end{pmatrix}$$

The observations in $\mathbf{y}$ are now not independent. There exist possible correlations between the time points, as well as possible correlations between the OTUs measured. The following sections will explain how these two types of correlations are accounted for.

2.2 Accounting for Correlation

2.2.1 Taxa correlation

Previous studies considering the correlations between OTUs have used a Dirichlet multinomial distribution. However, this assumes a negative correlation between OTUs (CITE La Rosa, etc), which might not be the case.

In order to model, the correlation structure between OTUs, the MLTC uses the taxonomic information of each OTU to specify the OTU correlation aspect of the working correlation matrix used in GEE estimation.

In order to reduce 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 they belong to the same taxonomic level at some higher. 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.

Consider a taxonomic tree that represents the following 5 OTUs.

EXAMPLE. (use paper's example?)

2.2.2 Longitudinal and repeated measure correlation

Additional correlations in the data occur from repeated measures. The most simple form of repeated measure comes from measurements at different discrete time points.

There are many different correlation structures that can be assumed from longitudinal data. For instance, one can assume that only adjacent time points are correlated (AR1 structure), or . This model can account for any specified repeated measure correlation structure.

2.2.3 Interative 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 ....

2.2.4 Two-part GEE model

The association between OTUs and the predictors is modeled in two parts. Since OTU data is space, with excessive zeros, the presence/absence of an OTU is modeled separately from the relative abundance of the OTU. The first models the presence/absence of the OTU. Since this is a binary response, a logistic link function in the GEE framework would be appropriate.

3 Extensions to the method

The paper provided code for their simulation study and calculating the integrative correlation matrix. However, no code was provided for applying this method. Much of the work I have done in this project is writing general code for use in new datasets.

3.1 "Missing data"

The above explanation of how to build the working correlation matrix assumes that all clusters have equal size/length. However, in practice this is rarely the case. Even if this was the case for the presence/absence modeling part, it is likely that the relative abundance modeling part will have essentially "missing" data for OTUs that are not present in a given sample.

Both the integrative correlation matrix and the `zcor` matrix must be adjusted to remove the corresponding entries

3.2 How to deal with unlabeled taxa

If we move beyond toy examples, there are many instances when the complete set of taxa labels are not available for each level. This can happen because... (find source).

The question of missing labels is whether OTUs with missing labels should be treated as distinct or not. If not, they should be combined into one observation.

For example, if we have measured counts on OTU_1 and OTU_2, and OTU_1 has labeled class "A" and order "B", and OTU_2 has labeled class "A", but no label for order. Should OTU_1 and OTU_2 be consolidated into one observation? Are these distinct? Consider a third OTU_3 with class "A" and missing order label. Is this distinct from OTU_2? Additionally what correlations should these OTUs have?

FINISH

3.3 Generating the taxa correlation from microbiome data

A detailed tree showing the hierarchy of taxa labels is not generally provided with OTU data. Instead, there is often a column listing the given taxa 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 `insert name here`. Starting at the top taxa level, the number of OTUs in each group is counted. This results in a list detailing the counts of OTUs belonging to each successive level.

INSERT EXAMPLE

3.4 Workflow of code

3.5 Application to American gut data

Currently, 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 this model scale?

3.5.1 About the American gut dataset

To explore the scalability of the MLTC, we use the American Gut dataset. For now, consider only observations at one timepoint, so the only correlations that occur will be taxonomic.

The American Gut Project....

3.5.2 Computational challenges

In `geeglm`, the user-defined working correlation structure is a Matrix. 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 individuals, and α has length a, this must satisfy the following 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.

Note that 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. On my computer, an overnight calculation of a small subset of the American gut, when the entries dataset fails to complete.

3.5.3 Filtering the American gut dataset

4 Conclusion

4.1 Future work

- GEE implementation that works with larger data
- Phylogenetic information
- Alternative to $-log_{10}$ relative abundance.

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

[5] 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.

[6] 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/`.

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