Masters Project Report

Emily Palmer

Spring 2021

Abstract

Application of a recent method for microbiome association studies to larger datasets provided computational challenges. Several aspects of the method to more general data were developed for use beyond the case study provided in the initial paper.

Contents

1 Introduction **2**

1.1 About microbiome data 2

1.1.1 Issues with microbiome data 3

1.1.2 Organizing OTUs onto a tree 3

1.2 Generalized Estimating Equations 4

2 About the method () **4**

2.1 Accounting for Correlation 4

2.1.1 Model Data structure 4

2.1.2 Taxa correlation 6

2.1.3 Longitudinal and repeated measure correlation 6

2.1.4 Interative correlation matrix 6

2.1.5 Two-part GEE model 6

3 Extensions to the method **6**

3.1 "Missing data" 6

3.2 Generating the taxa correlation from microbiome data 6

3.3 How to deal with unlabeled taxa 6

3.4 Application to American gut data 6

3.4.1 About the American gut dataset 6

3.4.2 Filtering the American gut dataset 6

3.4.3 Computational challenges 6

4 Conclusion **6**

4.1 Future work 6

1 Introduction

1.1 About microbiome data

The human gut contains millions of microbes. Genomic research on microbiome sequencing data has provided numerous insights to ...*** Studies have associated the human gut microbiome to irritable bowel syndrome, inflammatory bowel disease, cancers, diabetes, and obesity, with more studies coming out with more interesting results. [2] Bidirectional interactions have been found between the gut and the central nervous system, colloquially referred to as the "gut-brain-axis". [4]

These studies show the microbiome to be a research rich and important area of research for advances in medical, and our understanding of ourselves. (FIX)

One technique for microbiome analysis is sequencing the 16S ribosomal rRNA gene. This gene is a useful marker for

Microbiome data commonly is sampled from fecal samples, but oral, skin, vaginal, etc are also common areas (CITE)

Microbiome data commonly arises from 16S ribosomal ribonucleic acid RNA (rRna) sequencing studies. Microbiome sequencing data commonly consists of the count of reads on operational taxonomic units (OTUs) for each sample. Operational taxonomic units are.... Methods for microbiome analysis can either be focused on a single OTU or multiple OTUs simultaneously.

1.1.1 Issues with 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)

First, 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.

Secondly OTU data is very sparse. The library size of different samples can vary a lot. For many OTUs, the total reads is zero. Thus microbiome data contains excessive zeros, which much be accounted for in the analysis. Adjustments can be made using zero inflated models, or two-part models that separately model the presence of an OTU, and then next the abundance of present OTUs.

1.1.2 Organizing OTUs onto a tree

(FIND CITATIONS)

All life can be linked back to a biologic tree that describes how each organism is related to each other.

Two types of trees that can arise in microbial studies are phylogenetic trees and taxonomic

trees.

Using knowledge of a given OTU (FIX), it can provide insight into correlations that exist between different read counts present in samples, and 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. [3] 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.

2 About the method ()

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 Accounting for Correlation

2.1.1 Model Data structure

Commonly, OTU data is stored in two tables. Consider a dataset with k individuals. Each individual i has n_i different samples taken from it. In total there are $M = \sum_{i=1}^{k} n_j$ samples

taken. Additionally, each sample has N OTUs measured, and p predictors. The first table contains counts for OTU values, with OTU labels as rows and samples as columns. This table will have overall dimension $N \times M$ The second table contains metadata (predictor) information, with samples as rows and predictor values as columns, having dimension $M \times p$.

In this model, the response is the OTU value (either presence absence, or transformed relative abundance). The reponse $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_k)^t$

2.1.2 Taxa correlation

2.1.3 Longitudinal and repeated measure correlation

2.1.4 Interative correlation matrix

2.1.5 Two-part GEE model

3 Extensions to the method

3.1 "Missing data"

3.2 Generating the taxa correlation from microbiome data

3.3 How to deal with unlabeled taxa

3.4 Application to American gut data

3.4.1 About the American gut dataset

3.4.2 Filtering the American gut dataset

3.4.3 Computational challenges

4 Conclusion

4.1 Future work

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

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

[4] 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.