Longitudinal Microbiome Data Analysis

Michael Agronah

January 28, 2023

Contents

0.1 Longitudinal Microbiome Data Analysis

0.1.1 Literature Review and Background

0.1.1.1 Existing Models

0.1.1.2 Some existing packages

0.1.2 Proposed research

0.1.3 Challenges

0.1.4 Simulation studies

0.1.5 Conclusion

1 Research Goals

The goal of this project is to

1. model the relationship between taxa abundances over time and covariate variables of interest (for instance, clinical and environmental factors such as age, groups or disease status of patients, among others), while accounting for the correlation structure between individual taxa within individual subjects. The unstructured correlation matrix, however, requires too many parameter estimates and is computationally infeasible for a typical ASV or OTU tables with thousands of taxa. Thus, a reduced rank latent variable model's estimation of the structured correlation matrix will provide an understanding of the correlation between taxa on the community level.
2. model how microbiome abundances change over time between groups and determine taxa with differential abundances over time between groups.

2 Methods

2.1 Model Description:

The count data denoted by Y_{ijt} is modeled by a negative binomial model defined as

$$\begin{aligned} Y_{ijt} &\sim NB(mean = \mu_{ijt}, dispersion = \alpha_{it}), \\ \boldsymbol{\mu} &= g^{-1}(E) \\ E &= X\beta + Zb,\ b \sim MVN(0, \Sigma) \end{aligned} \tag{1}$$

where i, j and t denote the i^{th} taxa, j^{th} sample and the t^{th} time, respectively, and $Y_{i,j,t}$ denotes the count data for the i^{th} taxa in the j^{th} sample at time t^{th}. The function g^{-1} is the link function and $\boldsymbol{\mu}$ is a vector of means with entries μ_{ijt}. X and Z are the fixed and the random effect model matrices respectively and β is the fixed effect parameter. The vector b is a multivariate deviate from a gaussian distribution with a variance covariance matrix Σ.

Given an OTU table of dimension n, the parameters required to be estimated for Σ are given by $n(n+1)/2$. Thus, the number of parameter estimates for Σ grows quadratically with n. For instance, a rare OTU table with 10 taxa only will require $55 (= 10 * 11/2)$ parameter estimates for Σ. Consequently, the typical OTU table with thousands of taxa will require millions of parameter estimates, which is computationally impossible. Additionally, the unstructured covariance matrix Σ is often singular due to linear independence of the columns resulting in numerical instability and convergence problems of the fitting algorithms. A remedy is to use a latent variable model. This project applies a reduced rank latent variable model for estimating the structured covariance matrix. The reduced rank model expresses the columns of Σ by a set of $r < n$ latent variables. The number of parameters required to be estimated for the reduced rank model with r latent variables given by $2r + 1$, reducing the number of parameters to be estimated significantly to a linear function of r. In this

project uses the estimation procedure implement in the *glmmTMB* package in R package. Description of the estimation procedure is given in

3 Background

3.1 Introduction

Longitudinal data arises when repeated measures are collected for the same subject. For example; DNA taking the microbiome of pregnant women repeatedly over a multiple time points. The literature of longitudinal microbiome investigations has generally respond mainly to two queries: (1). to determine how microbiome abundances changes over time between groups (for example cases versus controls) (Lewis et al. 2015; Bäckhed et al. 2015), and 2. to investigate the relationship between microbiome abundances and other factors, for instance, environmental factors, clinical outcomes, etc) (Kodikara, Ellul, and Lê Cao 2022; Chen and Li 2016).

Analyzing longitudinal microbiome data is challenging. First, In addition to the special propertes of overdispersion of microbiome data due to the variations in read depth and zero inflation, repeated measurements exhibit correlation between observations taken from the same subject at various times points. Longitudinal data also often tend to have more missing data from a person or data not taken from everyone at all the time period studied (Kodikara, Ellul, and Lê Cao 2022). For instance, a longitudinal study collecting faecal specimens from patients may have missing data due to patients dropping out of the study during the period of the specimen collection.

3.2 Models used in the Literature

Over-Dispersion Models, Zero-Inflated and Zero-hurdle Models, Dirichlet-Multinomial Models, and Multivariate Bayesian Mixed-Effects Model and other Mixed Models have been commonly used in the literature for microbome analysis (Xia et al. 2018). This section gives an overview of some of the common modeled fitted to microbiome data in the literature.

3.2.1 Over-dispersion

Poisson and Negative Binomial distributions have been applied extensively for analyzing count data (Zhang et al. 2017). A limitation however of the poisson distributions is that it assumes the mean equals variance and does not account for overdispersion in microbiome data (Anders and Huber 2010). Negative binomial models have been proposed to address the problem of overdispersion (Zhang and Yi 2020). Zhang et al. (2017)

3.2.2 Zero-inflation and Zero-hurdle Models

Zero-Inflated Models and Zero-hurdle Models have been applied to address the problem of sparsity in microbiome data (Xia et al. 2018 ; Zhang and Yi 2020). The hurdle model (also know as a two-part model), considers the data as comprising of zero and nonzero components and model the probability of obtaining a zero by a binomial distribution or a logistic regression model and the non-zero counts by a zero-truncated model such as a zero-truncated Poisson or zero-truncated negative binomial model. Zero inflation models (also known classified as a mixture model), on the other hand, considers the data as consisting of a count of components including rounded zeros and a second component for excess zeros. The excess zeros are modeled by the probability of expecting a zero under the considered distribution.

Examples of zero-inflated and zero-hurdle models used in the literature for count data include zero-inflated Poisson (ZIP) and the zero-inflated negative binomial models (ZINB), zero-hurdle Poisson (ZHP) and the zero-hurdle negative binomial models (ZHNB) (Xia et al. 2018). The zero-inflated beta regression models have been used for proportion data (Xia et al. 2018). Zero-inflated Gaussian models have also been proposed for modelling either count or proposition data (Kodikara et al., 2022, p. 7). Paulson et al. (2013), for instance, developed a zero-inflated Gaussian model using the logtransformation on the read counts to address the problem of excessive zeros. An empirical Bayes method was used to estimate the moderated variances in

the model. Peng, Li, and Liu (2016) and Ospina and Ferrari (2012) also developed the zero-inflated beta regression model and the zero-or-one inflated beta regression model, respectively, for proportion data.

3.2.3 Dirichlet-Multinomial Models

More recently developed model for the analysis of microbiome data are the multinomial and Dirichlet-Multinomial Models (Xia et al. 2018; Holmes, Harris, and Quince 2012). The multinomial and Dirichlet-multinomial distributions are the commonly used parametric probability models (La Rosa et al. 2012; Holmes, Harris, and Quince 2012). Holmes, Harris, and Quince (2012) developed a Dirichlet multinomial mixtures model for classification and clustering otus of microbiome samples. A reparameterized Dirichlet multinomial model was developed by La Rosa et al. (2012) to test the hypothesis of mean abundances as well as scales (variance comparison/dispersion) between groups. This model is implemented in the **HMP** package in R.

3.2.4 Multivariate Bayesian Mixed-Effects Model

Multivariate Bayesian Mixed-Effects Model. Grantham et al. 2019 propose a Bayesian mixedeffects model to analyze microbiome data.

MIMIX performs spike-and-slab variable selection to identify treatment effects on OTUs. Bayesian factor analysis with a Dirichlet-Laplace prior clusters OTUs into different factors. MIMIX is not currently suited for handling data from longitudinal studies

3.2.5 Models for longitudinal data analysis

Mixed models have been used for modeling longitudinal data in the literature.

3.2.5.1 Zero-inflated beta regression Model To test the relationship between microbial abundance and clinical covariates in longitudinal microbiome data, Chen and Li (2016) developed a two-part zero-inflated Beta regression model with random effects (ZIBR). The model consists of a logistic regression component to model the presence/absence of a microbe in the samples, and a Beta regression component to model non-zero microbial abundance, with each component containing a random effect. One of ZIBR's strengths is its ability to account for the sparse nature of the data. However, within-subject correlation structure is not accounted for in the model (Zhang, Guo, and Yi 2020).

3.2.5.2 Negative binomial Mixed Model In order to identify correlations between microbial counts and covariates (such as therapy, age, dietary habits, etc.) while taking into account temporal patterns in microbial abundance within and between patients, the negative binomial mixed model (NBMM) was devised (Zhang et al. 2018). This NBMM is flexible and allows the inclusion of many other covariates limitation. NBMM is able to handle overdispersion and has also accounts for varying read depth by including an offset term. Unlike ZIBR, NBMM accounts for correlation structures among observations from the same subject.

3.2.5.3 Zero-inflated Gaussian mixed models ZIGMM is a mixture model involving two components; a logistic regression component and a Gaussian regression component. The model is flexible and can fit both count and proportion data. square root arcsine transformation is used to transform relative abundance and Count data are transformed by the log-ratio transformation log base 2. ... showed from a simulation studies that the ZIGMM outperformed ZIBR and ZINBMM and that the estimation method used is computationally efficient compared to those used in ZIBR and ZINBMM. The method accounts for the correlations within the repeated observations in samples using AR(1) or continuous-time auto-regressive of order 1. The method is implemented in the NBZIMM R package. Just like ... and ... ZIGMM is also univariate and does not account for the complex interactions between taxa , which could provide more insight into the microbiome. *Another problem is that is it is be slow for high dimensional data.-Explore this*

A limitation of all the above-mentioned models is that they are all univariate and do not account for the complex interactions between taxa , which could provide more insight into the microbiome.

4 Objective

The first objective is to study how microbial abundance changes over time between groups of interest (e.g. cases versus controls, disease or treatment groups, Figure 1A, and how the association between microbial abundance and other factors such as clinical outcomes, disease or treatments change over time [7]. In this context, both time and differences between patients or individual groups may be of interest.

Identification of microorganisms with differential abundance over time, between groups and between both group and time

Other multivariate methods have been used in other fields to analyiss high dimensional co

4.1 Packages

The R packages pscl, mgcv, brms, gamlss, GLMMadaptive, and glmmTMB can be used to analyze over-dispersed or sparse count data (Zhang and Yi 2020). The package pscl has a function zeroinfl() for fitting ZIP and ZINB, and the function hurdle() for fitting ZHP and ZHNB (Jackman et al. 2015). The package cannot handle fitting longitudinal data. metagenomeSeq analyzesmicrobiome data presented as proportion using a zero-inflated Gaussian mixture model (Joseph et al. 2013). metagenomeSeq can also be used for longitudinal differential abundance analysis using smoothing splines (Paulson, Talukder, and Bravo 2017).

BhGLM NBMMs

These models, however, do not handle repeatedly measured proportion data such as the longitudinal data considered in this paper. Furthermore, extending these methods to include random effects is not trivial.

5 Challenges

6 Simulation Studies

6.1 Data Simulation

Using the simulate function in *lme4*, we generated count data from a generalised linear mixed model with a negative binomial distribution. The simulation's parameters are displayed in Table 1. The plot of the trajectory for three taxa from the simulated data is shown in Figure 1. The simulated data was splitted into 80% training set and 20% test set. A rank reduced rank random slope model with time nested in subjects as the grouping variable, taxa as the random effect and groups and the interaction between group and time as the fixed effect variable was fitted to the data. The optimal dimension of 3 for the reduced rank was determined by comparing the AICs of the model fitted to dimensions ranging from 2 to 10. Table **??** displays the AIC values and dimensions.

```
fit_list <- lapply(2:10,
                   function(d) {
                     fit.rr <-
                     glmmTMB(count ~ group*time+ rr(-1 + taxa|subject,d = d),
                                                data = training, family = nbinom2)
                       })

# compare fits via AIC
aic_vec <- sapply(fit_list, AIC)
aic_vec - min(aic_vec, na.rm = TRUE)
```

7 What I am studying: The impact of covariates on the data

UNderstand the difference between multivariate and univariate metho

Table 1: Parameter values used in data simulation

Effects	**Other parameters**
Intercept = 1	Number of subjects = 20
Group = 0.2	Number of Taxa = 100
Time =0.1	Length of Time = 10
Group*Time = 0.2	Number of Groups = 2
Dispersion parameter = 2	Random effect parameter (θ) = Identity

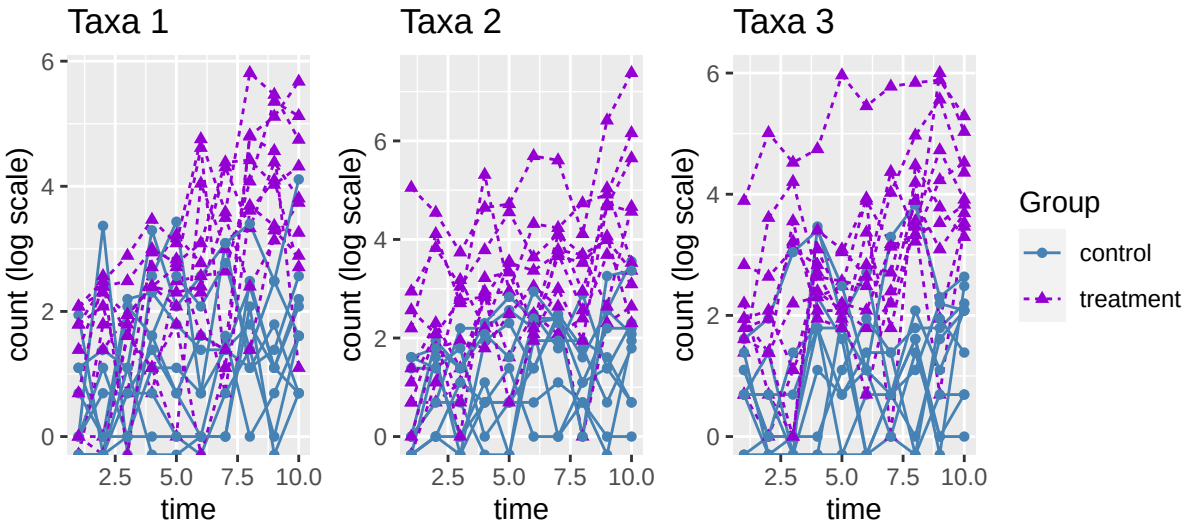

Figure 1: Some simulated taxa profiles with time, group and group * time effects and 0.2 dispersion

What am L doing with ny model, what does my model be used for what will my model be used for. Speak to Dr Stearn. What would my model be used for, what would my model It is What would the model be useed for?

7.1 What I will be doing

1. Fittingthe model itself.
2. I will be doing goodness of fit test
3. I will be shrinking the estimates and comparring the results with NZZBIM and other existing models
4. I will be doing model inferenceings
5. Find a way to do ordination plots and how to interprete it

8 Ideas for solving the 3rd (sample size calculation) question

Give a specified range for effect sizes and parameters for simulating control abundances, effect sizes and control abundances with be simulated from a truncated Cauchy distribution and a skewed normal restively using different number of sample sizes (example 10, 20 and 30) per group. Power estimates will be computed for each of sample size to determine the range of power in each case. We speculate that as numbers of sample sizes increases will lead to wider ranges power accross abundances.

I will be doing the that have been used in the analysis of microbiome data

1. The gllvm packge What are some of the advanges of the pack

Severeal modelds ha Niku et al. (2017) p. 498 applied a generalized linear latent variable model to fitting microbiome data. The

variable to account for covariation between species and correlation and applied it to the model metagenomic dataset. This generalized linear latent variable model is implemented in the gllvm package in R and can be used for fitting longitudinal data.

The algorithm used in the model-fitting is fast and can handle much larger datasets ()MLE and avoids MC estimations. Thus it can handle large datasets relatively quicker than other existing packages. Secondly, it uses a maximum likelihood framework, which allows the use of likelihood-based tools for inferencing. A limitation however, of gllvm is that it does assume observations are independent and thus does not account for correlation. Also, it does not include random slopes for predictors, which could enhance the account for the interspective variations in the responses.

The gllvm package implements generalised linear latetnt variable for modelling microbiome gene data. The latent variable component is used to capture the covariation between species not accounted for by the predictors. A key advantage of this package is that is computational speed and its ability to handle large datasets.

gllvm has been recently been update with the capacity to include random slope component to the GLLVM models, which was not there previously. to capture variation in environmental response not captured by the trait model." (Niku et al., 2019, p. 2180)

The packege has various functionality for statistical inference, model visualization and model selection (AICs). Give examples The used what is known as maximises the log-likelihood using the Gaussian variational approximations (VA, Hui et al., 2017) for overdispersed counts, binary and ordinal responses, or using Laplace approximations (LA, Niku et al., 2017) for other exponential family distributions when a fully closed form variational" (Niku et al., 2019, p. 2175). It has functionality for modeling poision, NZ and ZIP for count data. It uses automatic differentiation in C++ and has carfully chosen startign values

The corresponding ordination plot then provides a graphical representation of which sites are similar in terms of their species composition." (Niku et al., 2019, p. 2176) # Special features of microbiome data that makes it difficult to analyse Microbiome data have several features. Microbiome count data, i.e., OTU counts, taxa abundance, are naturally constrained, high dimensional, sparse with containing a large proportion of zero counts in the OTU(Taxa) table, complex covariance and correlation structures among different OTUs(taxa), and over-dispersed with large within-group heterogeneities.

9 Possible Extentions

10 Timeline

10.1 Time Line

1. March 1st - May 30th Goal: Write up results from the Power project and submit work to a journal for publication.

Detailed Steps a. Answer question 3: March 20th - April 10th b. Write up results for question 3 and the entire work April 10th - April 12th c. Submit write-up to a journal for publication April 12th - April 20th

2. May 20th - July 20th.

a. Get all my results for longitudinal project part 1 and write it up and hand it in for publication in
b.
c.

3. August - October Research on part 2 of longitudinal project and get all my results down and then write up and hand it in for publication

4. Design package. . . d. Write an R package for the power analysis project April 20th - May 30th

5. November to December, writing up the R packages for chapters 1 and 2 and is to apply of jobs attend conferences and present in conferences and to write up all my thesis. My thesis must be done and handed in by Febuary 20th.

Conclusion

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Nature Precedings*, 1–1.

Bäckhed, Fredrik, Josefine Roswall, Yangqing Peng, Qiang Feng, Huijue Jia, Petia Kovatcheva-Datchary, Yin Li, et al. 2015. "Dynamics and Stabilization of the Human Gut Microbiome During the First Year of Life." *Cell Host & Microbe* 17 (5): 690–703.

Chen, Eric Z, and Hongzhe Li. 2016. "A Two-Part Mixed-Effects Model for Analyzing Longitudinal Microbiome Compositional Data." *Bioinformatics* 32 (17): 2611–17.

Holmes, Ian, Keith Harris, and Christopher Quince. 2012. "Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics." *PloS One* 7 (2): e30126.

Jackman, Simon, Alex Tahk, Achim Zeileis, Christina Maimone, Jim Fearon, Zoe Meers, Maintainer Simon Jackman, and MASS Imports. 2015. "Package 'Pscl'." *Political Science Computational Laboratory* 18 (04.2017).

Joseph, N, C Paulson, H Corrada Bravo, and M Pop. 2013. "Robust Methods for Differential Abundance Analysis in Marker Gene Surveys." *Nat. Methods* 10: 1200–1202.

Kodikara, Saritha, Susan Ellul, and Kim-Anh Lê Cao. 2022. "Statistical Challenges in Longitudinal Microbiome Data Analysis." *Briefings in Bioinformatics* 23 (4): bbac273.

La Rosa, Patricio S, J Paul Brooks, Elena Deych, Edward L Boone, David J Edwards, Qin Wang, Erica Sodergren, George Weinstock, and William D Shannon. 2012. "Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data." *PloS One* 7 (12): e52078.

Lewis, James D, Eric Z Chen, Robert N Baldassano, Anthony R Otley, Anne M Griffiths, Dale Lee, Kyle Bittinger, et al. 2015. "Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease." *Cell Host & Microbe* 18 (4): 489–500.

Niku, Jenni, David I Warton, Francis KC Hui, and Sara Taskinen. 2017. "Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology." *Journal of Agricultural, Biological and Environmental Statistics* 22 (4): 498–522.

Ospina, Raydonal, and Silvia LP Ferrari. 2012. "A General Class of Zero-or-One Inflated Beta Regression Models." *Computational Statistics & Data Analysis* 56 (6): 1609–23.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. "Differential Abundance Analysis for Microbial Marker-Gene Surveys." *Nature Methods* 10 (12): 1200–1202.

Paulson, Joseph N, Hisham Talukder, and Héctor Corrada Bravo. 2017. "Longitudinal Differential Abundance Analysis of Microbial Marker-Gene Surveys Using Smoothing Splines." *BioRxiv*, 099457.

Peng, Xiaoling, Gang Li, and Zhenqiu Liu. 2016. "Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data." *Journal of Computational Biology* 23 (2): 102–10.

Xia, Yinglin, Jun Sun, Ding-Geng Chen, et al. 2018. *Statistical Analysis of Microbiome Data with r*. Vol. 847. Springer.

Zhang, Xinyan, Boyi Guo, and Nengjun Yi. 2020. "Zero-Inflated Gaussian Mixed Models for Analyzing Longitudinal Microbiome Data." *PloS One* 15 (11): e0242073.

Zhang, Xinyan, Himel Mallick, Zaixiang Tang, Lei Zhang, Xiangqin Cui, Andrew K Benson, and Nengjun Yi. 2017. "Negative Binomial Mixed Models for Analyzing Microbiome Count Data." *BMC Bioinformatics* 18 (1): 1–10.

Zhang, Xinyan, Yu-Fang Pei, Lei Zhang, Boyi Guo, Amanda H Pendegraft, Wenzhuo Zhuang, and Nengjun Yi. 2018. "Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data." *Frontiers in Microbiology* 9: 1683.

Zhang, Xinyan, and Nengjun Yi. 2020. "NBZIMM: Negative Binomial and Zero-Inflated Mixed Models, with Application to Microbiome/Metagenomics Data Analysis." *BMC Bioinformatics* 21 (1): 1–19.