Power Analysis of Microbiome Data

Michael Agronah

December 6, 2022

Contents

1 Multivariate analysis of Longitudianal Microbiome data using Generalised Linear Mixed Negative Binomial Model

1.1 Abstract

Human microbiome is dynamic in nature. Understanding the dynamics in longitudinal microbiome data can help explain the mechanisms that underpin human health and disease. However, longitudinal microbiome data analysis, whether 16S rRNA or metagenome shotgun, is difficult. Along with the difficulties presented by microbiome data characteristics such as sparsity, overdispersion, and compositionality, repeated measure introduces correlation between observations from the same individual at different time points. The majority of methods used in the literature to analyse longitudinal microbiome data only model one taxon at a single time point and do not handle the correlation struction between individual taxa across time. In this work, a multivariate Negative Binomial Generalized Mixed Model (MVNBMM) is proposed for anlyzing longitudinal microbiome data. A special feature of MVNBMM is that is can handle the correllation structure of individual taxa over time. Fitting MVNBMM to data is however, very challenging becausure the number of parameter to

be estimated for a negative binomial model is compistationall expensive and almost impossible for microbiome dataset which is typically of dimension in their thousands. In this project, Instead, a reduced rank method is used for fitting the data reduced rank method.

1.2 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.

2 Literature Review

Most of the methods used. Summarixe the methods and the pacakges used in the literature.

3 Research Questions

The goal of this project is to

1. model the relationship between otu counts and covariates (examples; time, age, gender and groups) while capturing the correlation structure within and between individual taxa and temporal patterns of taxa over time.
2. find taxa with differential abundances over time between groups. We consider a simple of where there are only 2 groups but the model can be extended to more than two groups. Model how microbiome abundance changes over time between groups.

4 Methods

4.1 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

More on why a reduced rank model is good.

5 Application

5.1 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:time,d = d),
                                             data = training, family = nbinom2)
                       })

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

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

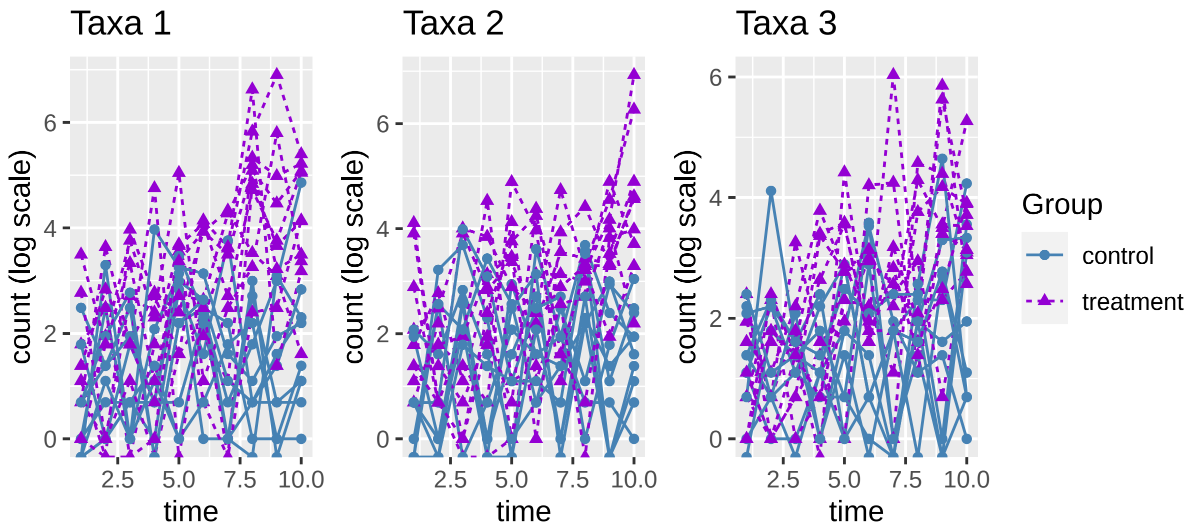

Figure 1: Some simulated taxa profiles with time, group and group * time effects and 0.2 dispersion

6 Results

6.1 Prediction

6.2 Comparing models

In this part, the NBMVM's prediction is compared to that of the univariate Negative Binomial Linear Mixed Model, which has been employed extensively in longitudinal microbiome investigations. Figure.. displays the mean squared errors for the random number of data simulations. It demonstrates that estimation of the handling the convariate structure yields a better fit of the mixed effect for these fixed effect estimates for all the effects.

We compare the performance of our model estimates to a negivebive biomian dit which fits on model at a time

1. What is the difference between the 2 models.

–>

7 Conclusion

8 Future Work

I would like to be able to say that using the full model or even the reduced rank model

1. Fix the warning
2. Study on what the reduced rank correllation tells me about the data but first you have to do the diagnoses on the model fit first.
3. SHow that the fits for is either the same or better than that of the single fit models by chceking the standard errors, etc the
4. Fix the table of values
5. Check this L <- gt_load("vignette_data/sherman.rda") again and see if it does what they say it will do
6. Understand Ben's code on fiting the confidence intervals on prediction and write on it. Model diagnostics

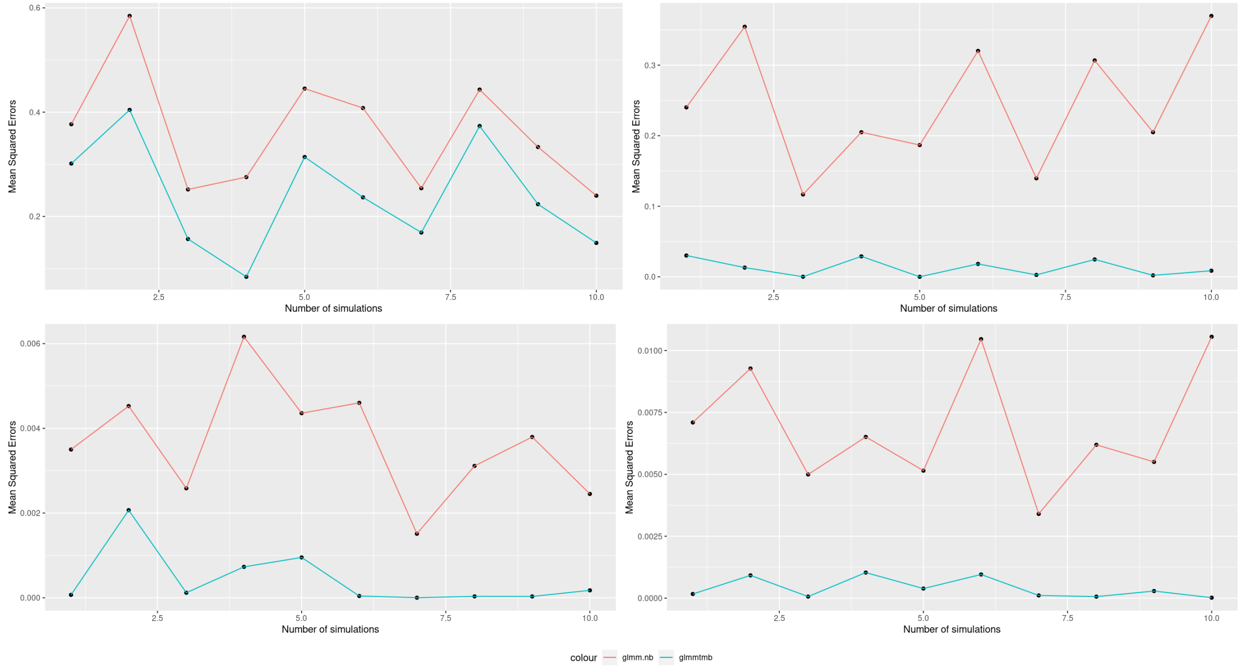

Figure 2: Heatmap showing the power and total number of taxa in grids of control and effect sizes

7. Intepreting low rank correlation matrices
8. What I can use my low rank matrix for. What the low rank tells me about the community

How otus are changing over time between groups while still capturing the correlation between otus in an individual over time 1. Understand the AR.

I need a way of assessing my model to see how it is doing on simulated data even before I use it on test data. How good is my reduced

Recent microbiome studies have employed the longitudinal study design to investigate the dynamic changes of microbial abundance over time and the associations between the microbiome and host environmental/clinical factors [11–15]. microorganisms work in concert to modulate and influence their environment [37]. We need to consider the multivariate relationship among microorganisms at a single time point, as well as across time points in the case of longitudinal studies. However, most existing modelling approaches consider one microorganism at a time and fail to capture the multivariate nature of the data [7, 30, 67, 86]. Moreover, in a longitudinal setting, representing the interactions and similarities or connectivity between microorganisms becomes complex when the data are high-dimensional, resulting in high computational cost and low prediction accuracy of most statistical methods.

profile of microbiome with host and environment interactions.The advantage of longitudinal analysis is also suitable for microbiome data. It will enhance our understanding of short-and long-term trends of microbiome by intervention, such as diet, and the development and persistence of chronic diseases caused by microbiome These microbiomes work together to produce the joint effects they porduce. Therefore we ne How to assess my

Goal: study the other methods, do the prediction, understand the syntax, study the direiclet multinomial,etc

So the whole point is that I cannot fit the full model so I fit the reduced rank model Noe w

I want to do prediction for individual taxon. Modelling

Predict function is predicting the mean of the 2 groups???

1. Listen to the

9 References

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.

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

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.