Power Analysis of Microbiome Data

Michael Agronah

December 8, 2022

Contents

1 Multivariate analysis of Longitudinal 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 diseaseS. However, longitudinal microbiome data analysis, whether 16S rRNA or metagenome shotgun, is difficult. Along with the difficulties presented by the characteristics of microbiome data such as sparsity, overdispersion, and compositionality, repeated measures introduce correlation between observations from the same individual at different time points. In an attempt to handle the problem of correlation, the majority of methods used in the literature to analyze longitudinal microbiome data only model one taxon at a single time point and do not account for the correlation structure between individual taxa across time. In this project, a Multivariate Negative Binomial Mixed Model (MNBMM) is proposed for analyzing longitudinal microbiome data. A special feature of MNBMM is that it can handle the correlation structure of individual taxa over time. Fitting MNBMM to data, however, is very challenging because the number of parameters estimates for the unstructured covariance matrix grows quadratically with the dimensionality of the data. For atypical microbiome data with thousands of taxa, fitting this model is computationally expensive and almost impossible. A possible solution to implement a latent variable model for fitting a covariance matrix with fewer parameter estimates. In this project, a reduced rank method is used for fitting the data reduced rank method. The **glmmTMB**

package in R provided a great functionality for fitting the reduced rank model. We demonstrate using from a simulation studies how implementation of the MNBMM works in **glmmTMB**.

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 multiple time points. The literature of longitudinal microbiome investigations has generally responded mainly to two queries: (1). to determine how microbiome abundances change 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).

Special features of microbiome data such as variation in sequencing depth, high dimensionality, overdispersion, excess zeros and compositionality, poses challenges when analyzing microbiome data(Xia et al. 2018). Variations in sequencing depth makes comparison of reads across samples is difficult. The problem of overdispersion renders statistical methods that assume means to be less or equal to variance ineffective. For example, the Poisson model which assumes mean to be equal to variance will underestimate the variance in microbiome data. The high dimensionality of microbiome data leads to numerical instabilities and convergence problems. Due to compositionality of microbiome data, statistical methods with normal distributional assumption cannot be used (Xia et al. 2018).

Longitudinal microbiome data analysis poses additional challenges. First, In addition to the special properties 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 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 fecal specimens from patients may have missing data due to patients dropping out of the study during the period of the specimen collection.

Several models such as the Zero-inflated beta regression Model, Negative binomial mixed model, Block Bootstrap Method, Zero-inflated Gaussian mixed models, Fast zero-inflated negative binomial mixed model and Bayesian semi-parametric generalized linear model have been proposed for the studies of longitudinal microbiome data (Kodikara, Ellul, and Lê Cao 2022). Many researchers have proposed a linear mixed model as a means to handle the correlation exhibited by taxa from repeated measures (Zhang, Guo, and Yi 2020).

In order to overcome the problem of correlations between observations from individuals across time, the majority of longitudinal microbiome research analyzes only one taxa at a time (Chen and Li 2016).

This approach, however, ignores the correlation structure between taxa from the individual subject. In this project, a Multivariate Negative Binomial Mixed Model (MVNBMM) is proposed for the studying longitudinal microbiome data. MVNBMM has the capacity to account for correlation instruction between individual taxa over time.

1.3 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.

1.4 Methods

1.4.1 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. The function g is the link function and $\boldsymbol{\mu}$ is a vector of means with entries μ_{ijt}. X and Z are the fixed and 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 or ASV table of dimension n, the parameter estimates required for Σ are given by $n(n+1)/2$. Thus, the number of parameter estimates grows quadratically with n. For an OTU table with 10 taxa for instance, $55(= 10 * 11/2)$ parameter estimates will be required for Σ. Consequently, the typical microbiome data with thousands of taxa will require millions of parameter estimates, which is computationally infeasible. Additionally, the unstructured covariance matrix Σ is often singular due to linear independence of the columns. This results in numerical instability and convergence problems when fitting the model. A possible remedy is to use a latent variable model, to specify a low dimensional latent variable that describes the full correlation structure. 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 is given by $nr - \binom{r}{2}$, hence, reducing the number of parameters to be estimated significantly to a linear function of n. The *glmmTMB* package (Brooks et al. 2017) in R provides a flexible framework for fitting mixed models with a range of response distributions and also has the capability to fit reduced rank variance structure. The *glmmTMB* package is used in this project for the estimation of the reduced rank model.

1.5 Application

1.5.1 Data Simulation

Using the simulate function in **lme4**, we generated count data from a generalized 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. A rank reduced rank random slope model with subjects as the grouping variable, taxa as the random effect variable, 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.

```
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)
```

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

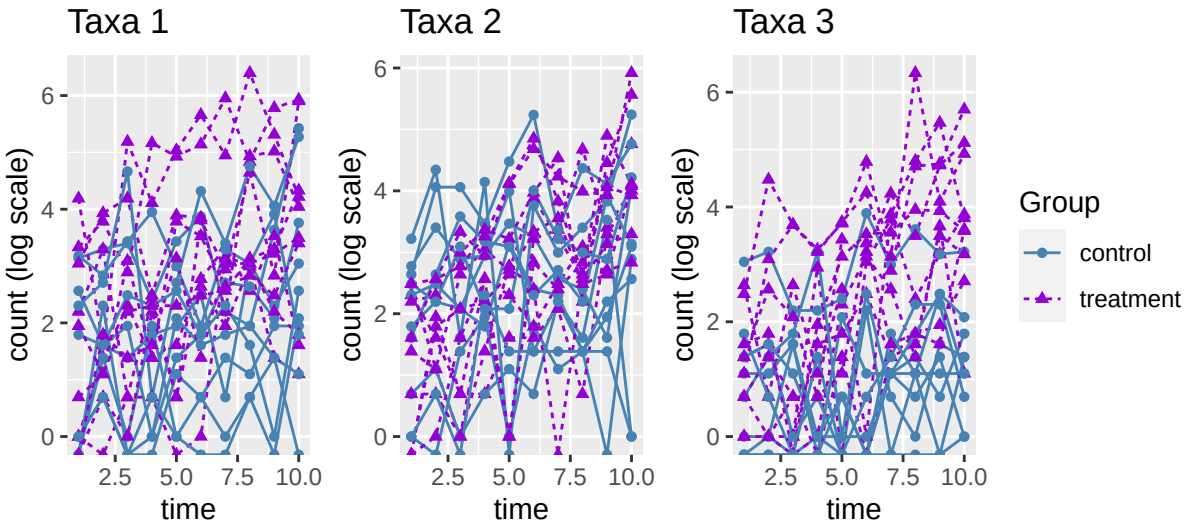


Figure 1: Some simulated taxa profiles with time, group and group * time effects and 0.2 dispersion

2 Results

2.1 Comparison with the Univariate Negative Binomial Mixed Model

The univariate negative binomial mixed model (UVNBMM) has often been used in the literature of longitudinal microbiome studies (Kodikara, Ellul, and Lê Cao 2022) . In this section we compare the effectiveness of the univariate NBMM and MVNBMM in predicting the fixed effect parameter. The univariate NBMM was fitted using the glmm.nb() package developed by which is commonly used for longitudinal microbiome studies in the literature (Zhang and Yi 2020). Figure 2 shows the mean square errors of the fixed effect parameter estimated from both models for 50 simulations. For simplicity, 10 taxa are used for this comparison study. The plot shows lower errors from the MVNBMM model. Consequently, accounting for correlations between the individual taxa provides better estimates of the fixed effect parameters.

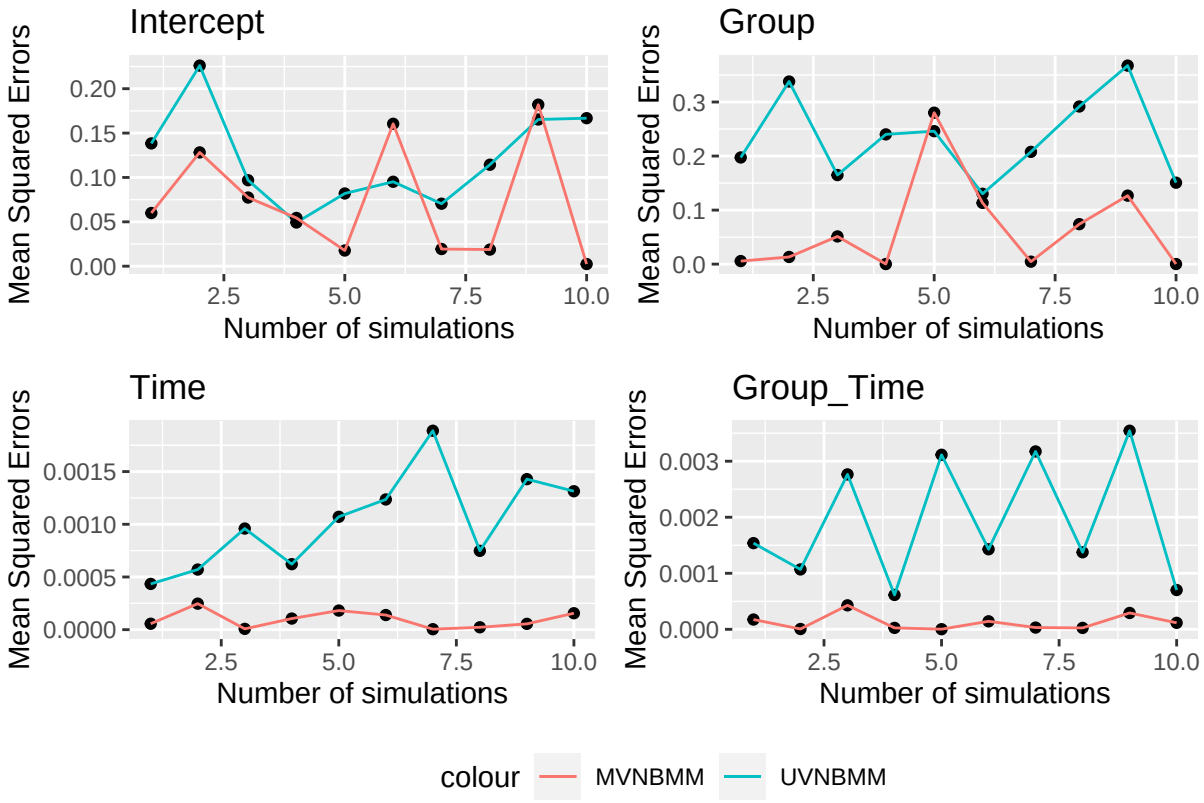
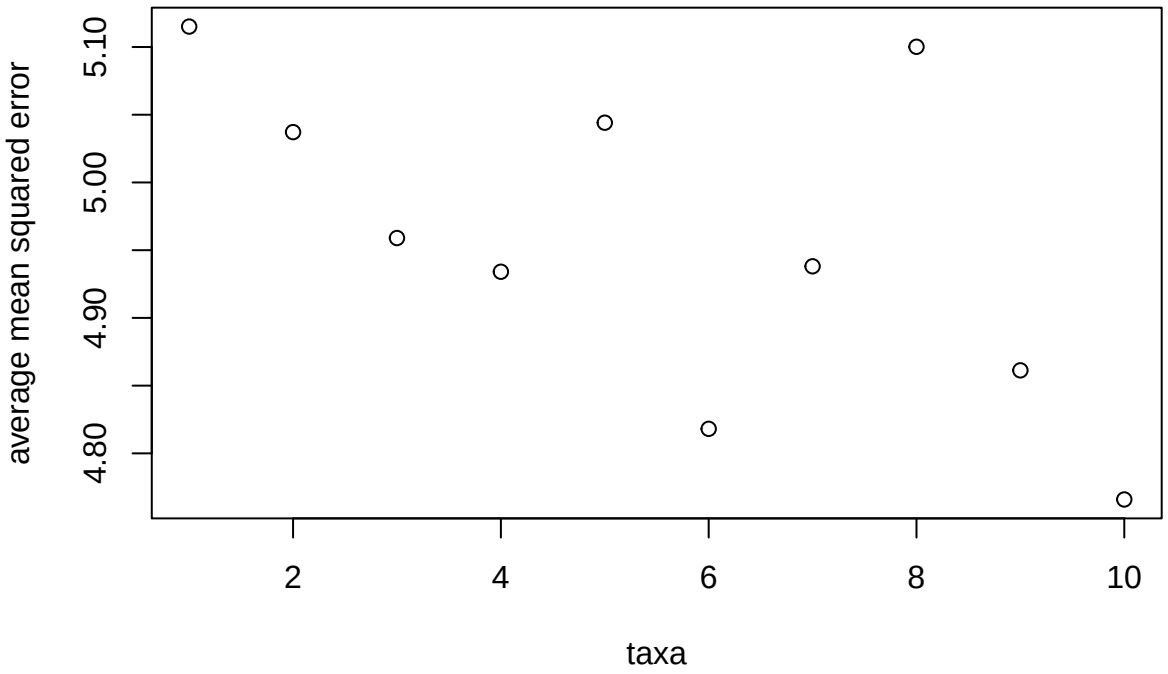


Figure 2: Comparison of the mean squared error fixed effect parameters from the Univariate and Multivariate Negative Mixed Models

2.1.1 Model Prediction



This section shows the mean squared error of our model prediction for individual taxa for 50 simulations. The points on the plot represent that average mean squared errors for individual taxa across time. Here, we consider 10 taxa On average the mean squared error for these over taxa is

2.2 Conclusion

In this project, Multivariate Negative Binomial Mixed Model (MNBMM) was proposed for analyzing longitudinal microbiome data. The model is capable of handling correlations between taxa from individuals over time. The reduced rank component of MNBMM allows fitting of this model to microbiome data of large dimensions. MNBMM also has the capability to handle overdispersion in microbiome data. This project demonstrates how to fit the model in the **glmmTMB** package in R using simulation studies. We show using the simulation studies that the models provide better estimates of the fixed effect parameters compared to the traditional univariate Negative Binomial Mixed Model applied in the literature of longitudinal microbiome studies.

3 Further development

A limitation of MNBMM is that it does not handle zero inflation found in microbiome data. Additionally, MNBMM cannot handle microbiome data presented in propositions and does not handle compositionality in microbiome data. Further work on this project will extend this model to include zero inflation component so as to account for the zero inflation seen in microbiome data. A new model with a framework similar to MNBMM, for example a Multivariate Gaussian Mixed Model will be developed to handle microbiome data presented as proportions.

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.

Brooks, Mollie E, Kasper Kristensen, Koen J Van Benthem, Arni Magnusson, Casper W Berg, Anders Nielsen, Hans J Skaug, Martin Machler, and Benjamin M Bolker. 2017. "glmmTMB Balances Speed and Flexibility Among Packages for Zero-Inflated Generalized Linear Mixed Modeling." *The R Journal* 9 (2): 378–400.

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.

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, 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.