NBZIMM: Negative Binomial and Zero-Inflated Mixed Models, with Application to Microbiome Data Analysis

Nengjun Yi

Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA

ABSTRACT

Summary: `NBZIMM` is a freely available R package that provides functions for setting up and fitting negative binomial mixed models, zero-inflated negative binomial mixed models and zero-inflated Gaussian mixed models. It also provides functions to summarize the results from fitted models, both numerically and graphically. `NBZIMM` is built on top of the commonly used R packages `nlme` and `MASS`, allowing us to incorporate the well-developed analytic procedures into the framework for analyzing over-dispersed and zero-inflated count data with multilevel structures (e.g., longitudinal studies). The statistical methods and their implements in `NBZIMM` particularly address the data characteristics and the complex designs in microbiome/metagenomic studies. Thus, the `NBZIMM` package provides useful tools for complex microbiome data analysis. This note describes the models, algorithms and associated features implemented in `NBZIMM`.

Availability: The package is freely available from the public GitHub repository https://github.com/nyiuab/NBZIMM.

Contact: nyi@uab.edu

1 INTRODUCTION

The advent of next-generation sequencing technology, including the 16S ribosome RNA gene sequencing and the shotgun sequencing, enables the generation of large volume of microbiome/metagenomic data, which provides valuable resources for investigating interactions between the microbiome and host clinical/environmental factors (Gilbert, et al., 2016). High-throughput microbiome data have characteristics that require tailored analytic tools(Zhang, et al., 2017; Zhang, et al., 2016). First, the total sequence reads largely vary over the samples, making the observed counts for a certain taxon incomparable directly. Secondly, the observed counts are over-dispersed, due to both biological and technical variabilities. Thirdly, the observed counts for many taxa are zero-inflated, because the organism can be simply absent (true zeros) or sufficiently rare such that it does not appear in the sequence collection (sampling zeros). Lastly, the number of measured taxa is usually large, thus requiring computationally efficient methods for detecting significant taxa. In addition to these special data features, many microbiome studies collect samples with hierarchical and temporal structures (Gilbert, et al., 2016; La Rosa, et al., 2014; Romero, et al., 2014); for instance, many recent studies employ logitudinal designs, which measure the same subject at multiple time points. The microbiome data within the same subject may be correlated and time-dependent. Thus it is needed to account for correlation and to characterize the time trends within and between subjects.

This note introduces the freely available R package `NBZIMM` for analyzing complex microbiome data, which is an implementation of our published methods (Zhang, et al., 2017; Zhang, et al., 2018) and our new methodological developments. The methods in `NBZIMM` address the data characteristics and the complex designs described above. Negative binomial and zero-inflated negative binomial distributions are commonly used for analyzing over-dispersed and zero-inflated count data, and mixed models are the standard approaches for dealing with multilevel data structures (Pinheiro and Bates, 2000; Venables and Ripley, 2002). `NBZIMM` provides functions for setting up and fitting negative binomial mixed models (NBMMs), zero-inflated negative binomial mixed models (ZINBMMs) and zero-inflated Gaussian (linear) mixed models (ZIGMMs). It also has functions for numerically summarizing the fitted models and graphically visualizing the results. The main functions in `NBZIMM` are developed based on the commonly used R packages, `MASS` and `nlme`, for analyzing negative binomial models and

linear mixed models (LMMs)(Pinheiro and Bates, 2000; Venables and Ripley, 2002), and inherit powerful features of these standard tools. Thus, the `NBZIMM` package provides efficient and flexible tools for analyzing complex microbiome data.

2 MODELS AND ALGORITHMS

We describe our mixed models and algorithms with a two-level design where samples are grouped in subjects, whereas our methods can deal with more complicated multilevel designs. Assume that a microbiome study collects n subjects (individuals) and n_i samples for the i-th subject. For each sample, we measure the counts for m taxa (OTU, species, genus, classes, etc.), C_{ijh}; $h = 1, \cdots, m$, the total sequence read T_{ij}, and some relevant covariates X_{ij}. The goal is to detect significant associations between taxa and covariates.

2.1 Negative Binomial Mixed Models (NBMMs)

The function `glmm.nb` in the `NBZIMM` package allows us to analyze the data for taxon h with NBMMs:

$$C_{ijh} \sim \mathrm{NB}(C_{ijh} \mid \mu_{ijh}, \theta_h),\ \log \mu_{ijh} = \log(T_{ij}) + X_{ij}\beta_h + G_{ij}b_{ih},\ b_{ih} \sim N(0, \Psi_h) \qquad (1)$$

where the dispersion θ_h determines the over-dispersion, the offset $\log(T_{ij})$ accounts for the varying total sequence reads, β_h is a vector of fixed effects, and b_{ih} are random effects. Inclusion of the random effects accounts for subject-specific effects and avoids biased inference on the fixed effects(Pinheiro and Bates, 2000). The `glmm.nb` function iteratively approximates the NBMM by a linear mixed model, which in turn is fitted using the function `lme` in the package `nlme`. The dispersion θ_h is updated using Newton-Raphson algorithm as in the function `glm.nb` of `MASS`. This framework allows us to incorporate the powerful features of `lme` into NBMMs.

2.2 Zero-Inflated Negative Binomial Mixed Models (ZINBMMs)

The function `glmm.zinb` implements ZINBMMs that directly model true zeros and thus can be more efficient for analyzing taxa with excessive zeros than `glmm.nb`:

$$C_{ijh} \sim \begin{cases} 0 & \text{with probability } p_{ijh} \\ \text{NB}(C_{ijh} \mid \mu_{ijh}, \theta_h) & \text{with probability } 1\text{- } p_{ijh} \end{cases} \quad (2)$$

Here, the means μ_{ijk} are modeled as above, and the zero-inflation probabilities p_{ijh} are assumed to depend on some covariates and be modeled via a logistic regression logit(p_{ijh}) = $Z_{ij}\alpha_h$ or logistic mixed model logit(p_{ijh}) = $Z_{ij}\alpha_h$ + $G_{ij}a_{ih}$, where α_h is a vector of fixed effects and the random effects $a_{ih} \sim$ N(0, Φ_h). The `glmm.zinb` function employs an EM algorithm to fit the above ZINBMMs, also taking advantage of the standard function `lme`.

2.3 Zero-Inflated Linear Mixed Models

Rather than directly analyzing the observed counts, some methods analyze transformed data(Paulson, et al., 2013), for example, y_{ijh} = $\log_2(C_{ijh} + 1)$. The transformed data can be analyzed using the zero-inflated linear (Gaussian) mixed models:

$$y_{ijh} \sim \begin{cases} 0 & \text{with probability } p_{ijh} \\ \text{N}(y_{ijh} \mid \mu_{ijh}, \sigma_h^2) & \text{with probability } 1\text{- } p_{ijh} \end{cases} \quad (3)$$

The function `lme.zig` in `NBZIMM` implements ZIGMMs for certain transformed data and uses an EM algorithm for fitting ZIGMMs.

3 FEATURES

The data inputs and specifications of fixed and random effects in the above functions are the same as in the standard function `lme`, allowing us to incorporate all the forms of fixed and random effects implemented in `lme` into our models. The function `lme` can specify various forms of within-subject residual correlation structures(Pinheiro and Bates, 2000), for instance, autoregressive of order 1, AR(1), which also have be incorporated into our framework. The fitted models from the above functions can be summarized using functions in the package `nlme`. For example, the function `summary` returns the estimates, standard deviations and p-values of fixed effects, and the estimates of the variances of random effects, *etc*. These features are remarkable, making `NBZIMM` easy to use, flexible and comprehensive in modeling, and stable and fast in computation.

Microbiome data usually consist of numerous taxa, however, the functions `glmm.nb`, `glmm.zinb`, and `lme.zig` only model one taxa at a time. The `NBZIMM` package includes a wrapper function, `mms`, to screen all included taxa, using NBMMs, ZINBMMs, LMMs, or ZIGMMs, by repeated calls to `glmm.nb`, `glmm.zinb`, `lme`, or `lme.zig`, respectively. The function `fixed` can be used to extract the estimates, standard deviations and p-values for fixed effects of all the taxa and covariates from the output of `mms`, while `get.fixed` extracts the estimates, standard deviations and p-values of fixed effects for a given taxon or covariate.

For datasets with numerous taxa and multiple covariates, it is needed to display the analytic results graphically. The `NBZIMM` package provides two functions to visualize the estimates, standard deviations and p-values of fixed effects from the output of `mms`. The function `plot.fixed` plots the estimates, intervals and p-values for numerous fixed effects. It uses different colors to distinguish between significant and insignificant effects. The function `heat.p` displays ggplot2-based heatmap to visualize p-values and sign of significant effects for numerous taxa and multiple covariates. Examples of `plot.fixed` and `heat.p` output are shown in Figures 1 and 2, using the longitudinal microbiome data of Romero *et al.*(2014) that is available in `NBZIMM`. The model included four fixed effects: Age, GA_Days, Race, and pregnant status, and a random intercept.

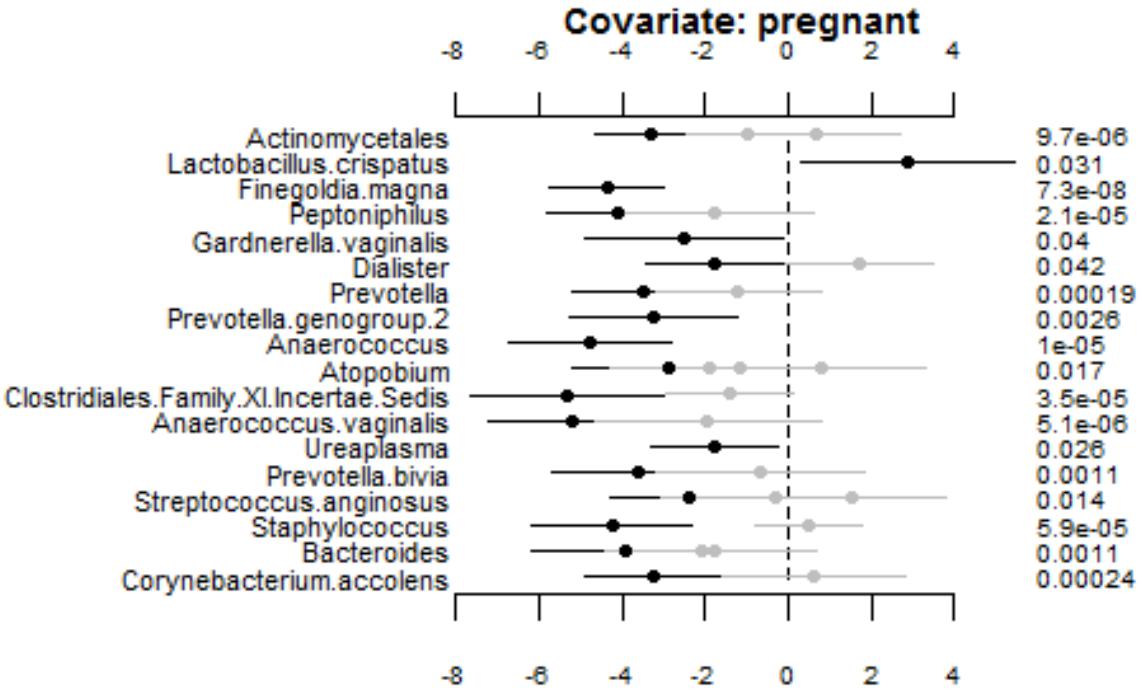

Figure 1. Pregnant effects on taxa. Only significant taxa are labeled.

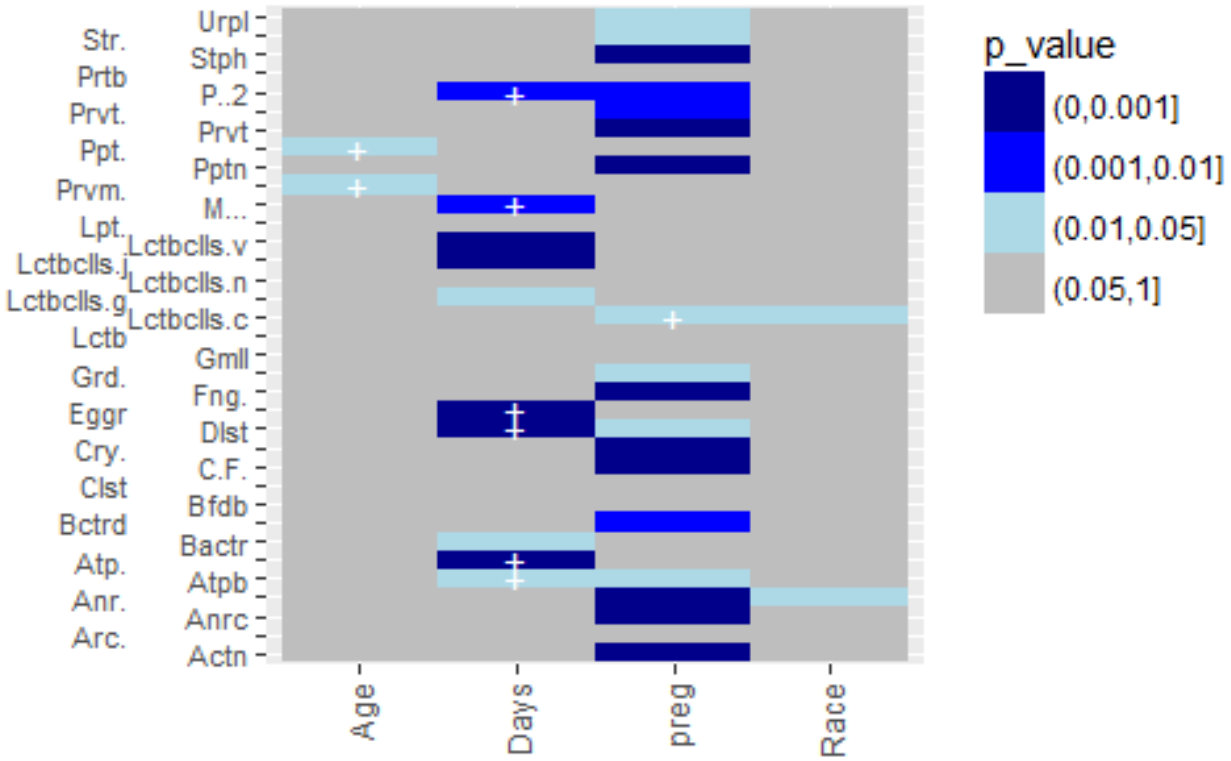

Figure 2. A heat map for p-values. The taxa names on y-axis are abbreviated. The sign "+" indicates the positive effect.

4 DISCUSSION

We have developed a freely available R package `NBZIMM` that addresses some of the analytic challenges in complex microbiome studies. Although we emphasize the application to microbiome data analysis, the package and the methods are general, and can be used to analyze other over-dispersed and zero-inflated count data with multilevel designs. The `NBZIMM` package is still under continual development. Our mixed models adopt a classical framework that is not appropriate to jointly analyze multiple correlated covariates. We will extend the mixed models by incorporating weakly informative prior distributions for the fixed effects that allow us to obtain more reliable and stable inferences(Gelman, et al., 2014). We also plan to develop mixed models for jointly analyzing multiple taxa.

REFERENCES

Gelman, A.*, et al.* (2014) *Bayesian Data Analysis*. Chapman & Hall/CRC Press, New York.

Gilbert, J.A.*, et al.* (2016) Microbiome-wide association studies link dynamic microbial consortia to disease, *Nature*, **535**, 94-103.

La Rosa, P.S.*, et al.* (2014) Patterned progression of bacterial populations in the premature infant gut, *Proc Natl Acad Sci U S A*, **111**, 12522-12527.

Paulson, J.N.*, et al.* (2013) Differential abundance analysis for microbial marker-gene surveys, *Nat Methods*, **10**, 1200-1202.

Pinheiro, J.C. and Bates, D.C. (2000) *Mixed-Effects Models in S and S-PLUS*. Springer Verlag, New York.

Romero, R.*, et al.* (2014) The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women, *Microbiome*, **2**, 4.

Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*. Springer-Verlag New York.

Zhang, X.*, et al.* (2017) Negative Binomial Mixed Models for Analyzing Microbiome Count Data, *BMC Bioinformatics*, **18**, 4.

Zhang, X., Mallick, H. and Yi, N. (2016) Zero-inflated Negative Binomial Regression for Differential Abundance Testing in Microbiome Studies, *Journal of Bioinformatics and Genomics*, **2**, 2.

Zhang, X.*, et al.* (2018) Negative Binomial Mixed Models for Analyzing Longitudinal Microbiome Data, *Frontiers in Microbiology*