La loi négative binomiale pour le traitement des données des microbiotes

SAHBANE Abdesstar

08 November 2020

Contents

1 Introduction

Recent advances in next-generation sequencing (NGS) technology enable researchers to collect a large volume of metagenomic sequencing data. These data provide valuable resources for investigating interactions between the microbiome and host environmental/clinical factors. In addition to the well-known properties of microbiome count measurements, for example, varied total sequence reads across samples, over-dispersion and zero-inflation, microbiome studies usually collect samples with hierarchical structures, which introduce correlation among the samples and thus further complicate the analysis and interpretation of microbiome count data.

In this article, we propose negative binomial mixed models (NBMMs) for detecting the association between the microbiome and host environmental/clinical factors for correlated microbiome count data. Although having not dealt with zero-inflation, the proposed mixed-effects models account for correlation among the samples by incorporating random effects into the commonly used fixed-effects negative binomial model, and can efficiently handle over-dispersion and varying total reads. We have developed a flexible and efficient IWLS (Iterative Weighted Least Squares) algorithm to fit the proposed NBMMs by taking advantage of the standard procedure for fitting the linear mixed models.

2 Methods

Typical microbiome data generated by the 16S rRNA gene sequencing or the shotgun metagenomic sequencing consist of the following components : 1) Counts, Cij , for n samples and m features. The features may refer to bacterial taxa at different hierarchical levels (species, genus, classes, etc.), groups of correlated taxa, gene functions, or pathways, etc.; 2) Total sequence read, T_i , for each sample; 3) Host factors, X_i , representing host clinical/environmental or genetic variables; 4) Sample variables, Z_i , representing sample collection identifier in the hierarchical study design, such as family structure, repeated measures from multiple body sites or time points. The goal is to detect associations between microbiome features C_ij and host factors X_i . The total sequence reads vary from sample to sample by orders of magnitude and can largely bias comparison of counts across samples, and thus should be accounted for in the analysis. Sample variables Z_i introduce hierarchical, spatial, and temporal dependence of microbiome counts, and should be included in the analysis as random factors.

Similar to most existing methods, we separately analyze each feature (count response) in a univariate fashion. For notational simplification, we denote $y_i = C_{ij}$ for any given feature j. We assume that the count response y i follows the negative binomial distribution:

$$y_i \sim NB\left(y_i \mid \mu_i, \theta\right) = \frac{\Gamma\left(y_i+\theta\right)}{\Gamma(\theta) y_i!} \cdot \left(\frac{\theta}{\mu_i+\theta}\right)^{\theta} \cdot \left(\frac{\mu_i}{\mu_i+\theta}\right)^{y_i}$$

where μ_i and are the mean and the shape parameter, respectively, and γ is the gamma function. The negative binomial distribution can be expressed as a gamma mixture of Poisson distribution : $y_i \sim$ Poisson $\left(y_i \mid \mu_i \varepsilon_i\right)$ and $\varepsilon_i \sim \text{Gamma}(\theta, \theta)$. It can be derived that $\mathrm{E}\left(y_i\right) = \mu_i, \mathrm{Var}\left(y_i\right) = \mu_i + \frac{\mu_i^2}{\theta}$, and $\mathrm{Var}\left(y_i\right) \geq \mathrm{E}\left(y_i\right)$. Thus, the shape parameter controls the amount of over-dispersion. when $\theta = +\infty, \mathrm{Var}\left(y_i\right) = \mu_i$ and the negative binomial model converges to a Poisson model that cannot deal with over-dispersion.

Our negative binomial mixed models (NBMMs) relate the mean parameters μ_i to the host factors X_i (including the intercept), the sample variables Z_i and the total sequence reads T_i via the link function logarithm:

$$\log\left(\mu_i\right) = \log\left(T_i\right) + X_i\beta + Z_i b$$

where $\log(T_i)$ is the offset, which corrects for the variation of the total sequence reads across the samples, β is the vector of fixed effects for the host factors X_i and b is the vector of K random effects for the sample variables Z_i . The random effects are used to model the correlation among the samples and the multiple sources of variation, and thus

to avoid biased inference on the effects of the host factors X_i . The vector of the random effects is usually assumed to follow the multivariate normal distribution : $b \sim N_K(0, \Psi)$ where Ψ is a positive-definite variance-covariance matrix that determines the form and complexity of random effects. Although in principle our NBMMs can deal with various patterns of Ψ, we here describe the method with a simple case where the random effects are independent, i.e.,$b \sim N_K\left(\mathrm{o}, \tau^2 I\right)$.

3 The IWLS algorithm for fitting the NBMMs

We propose an IWLS (Iterative Weighted Least Squares) algorithm to fit the NBMMs by extending the commonly used algorithms for fitting generalized linear models (GLMs) and generalized linear mixed models (GLMMs). For any fixed shape parameter , the negative binomial density is of the exponential form :

$$NB\left(y_i \mid \mu_i, \theta\right) = \exp\left\{\frac{y_i \vartheta_i - b\left(\vartheta_i\right)}{\phi} + c\left(y_i, \phi\right)\right\}$$

où $\vartheta_i = \log \frac{\mu_i}{\mu_i+\theta}, \phi = 1, b\left(\vartheta_i\right) = -\theta \log\left(1 - e^{\log \frac{\mu_i}{\mu_i+\theta}}\right) = -\theta \log\left(1 - e^{\vartheta_i}\right)$,
$et\ \mathrm{c}(y_i, \phi) = \log\left(\frac{\Gamma(y_i+\theta)\theta^\theta}{\Gamma(\theta) y_i!}\right)$
Therefore, the negative binomial model is a special case of generalized linear models (GLMs) for any fixed θ fixe . If θ is an unknown parameter, the negative binomial model is not a GLM. However, the NBMMs can be fit by iteratively updating the parameters (β , b , τ^2) and θ . Conditional on θ , the NBMM is a special GLMM and thus the parameters (β , b , τ^2) can be updated by using the GLMMs procedure. Conditional on (β , b), the shape parameter θ can be updated by maximizing the NB likelihood using the standard Newton–Raphson algorithm.
Conditional on , we update the parameters (β , b , τ^2) by extending the IWLS algorithm or equivalently the Penalized Quasi-Likelihood procedure for fitting GLMMs. The IWLS algorithm proceeds to approximate the generalized linear model likelihood by a weighted normal likelihood and then update the parameters from the weighted normal model. Conditional on the shape parameter θ, the fixed effects β and the random effects b, the negative binomial likelihood $NB\left(y_i \mid \mu_i, \theta\right)$ can be approximated by the weighted normal likelihood: :

$$NB\left(y_i \mid \mu_i, \theta\right) \approx N\left(t_i \mid \eta_i, w_i^{-1}\right)$$

where $\eta_i = \log\left(T_i\right) + X_i\beta + Z_i b$, the 'normal response data' t_i and the 'weights' w_i are called the pseudo-response and the pseudo-weights, respectively. The pseudo-response t_i

and pseudo-weights w_i are calculated by:

$$t_i = \hat{\eta}_i - \frac{L'\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right)}{L'\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right)}, \text{ et } w_i = -L''\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right)$$

where $\hat{\eta}_i = \log(T_i) + X_i\widehat{\beta} + Z_i\hat{b}, L\left(y_i \mid \hat{\eta}_i, \hat{\theta}\right) = \log NB\left(y_i \mid \widehat{\mu}_i, \hat{\theta}\right)$,
$L'(y_i \mid \eta_i, \theta) = dL(y_i \mid \eta_i, \theta)/d\eta_i, L''(y_i \mid \eta_i, \theta) = d^2L(y_i \mid \eta_i, \theta)/d\eta_i^2$,
and $(\widehat{\beta}, \hat{b})$ and $\hat{\theta}$ are the current estimates of (β, b) et θ, respectively. Therefore, the NBMMs can be approximated by the linear mixed model with w_i as weights:

$$t_i = \log(T_i) + X_i\beta + Z_ib + w_i^{-1/2}e_i, b \sim N_K\left(0, \tau^2\right), e \sim N_n\left(0, \sigma^2 I\right)$$

The parameters $(\beta, b, \tau^2, \sigma^2)$ are then updated from this linear mixed model by using the standard algorithm for fitting LMMs.
In summary, the IWLS for fitting the NBMMs is an iterative algorithm and proceeds as follows:

1. Initialize β, b, and θ some plausible values;
2. for $j = 1, 2, \cdots$
   (a) Based on the current values $\left(\beta^{(j-1)}, b^{(j-1)}, \theta^{(j-1)}\right)$, , calculate pseudo-response $t_i^{(j)}$ and pseudo-weights $w^{(j)}$;
   (b) update $(\beta, b, \tau^2, \sigma^2)$ by fitting the LMM (6);
   (c) Update θ by the standard Newton–Raphson algorithm.
3. Repeat Step 2) until convergence.

We use the criterion

$$\left(\eta^{(j)} - \eta^{(j-1)}\right)^2 < \varepsilon\left(\eta^{(j)}\right)^2,$$

to assess convergence, where

$$\eta^{(j)} = \sum_{i=1}^{n}\left(\log(T_i) + X_i\beta^{(j)} + Z_ib^{(j)}\right),$$

and ε is a small value (say 10^-5).

References

[1] Zaixiang Tang Lei Zhang Xiangqin Cui Andrew K. Benson Nengjun Yi Xinyan Zhang, Himel Mallick. Negative binomial mixed models for analyzing microbiome count data. 2017.