Mixed models for correlated compositional data: applied to microbiome studies in Indonesia

Martin, I.

3

The mixed model for the analysis of a repeated-measurement multivariate count data

Abstract

Clustered overdispersed multivariate count data are challenging to model due to the presence of correlation within and between samples. Typically, the first source of correlation needs to be addressed but its quantification is of less interest. Here we focus on the correlation between time-points. In addition, the effects of covariates on the multivariate counts distribution need to be assessed. To fulfill these requirements, a regression model based on the Dirichlet-multinomial distribution for association between covariates and the categorical counts is extended by using random effects to deal with the additional clustering. This model is the Dirichlet - multinomial mixed regression model. Alternatively, a negative binomial regression mixed model can be deployed where the corresponding likelihood is conditioned on the total count. It appears that these two approaches

This chapter has been published as: Ivonne Martin, Hae-Won Uh, Taniawati Supali, Makedonka Mitreva, Jeanine J. Houwing-Duistermaat (2019). The mixed model for the analysis of a repeated measurement multivariate count data. *Statistics in Medicine*, 38(12): 2248 - 2268.

are equivalent when the total count is fixed and independent of the random effects. We consider both subject-specific and categorical-specific random effects. However, the latter has a larger computational burden when the number of categories increases. Our work is motivated by microbiome datasets obtained by sequencing of the amplicon of the bacterial 16S rRNA gene. These data have a compositional structure and are typically overdispersed. The microbiome dataset is from an epidemiological study carried out in a helminth-endemic area in Indonesia. The conclusions are: time has no statistically significant effect on microbiome composition, the correlation between subjects is statistically significant, and treatment has a significant effect on the microbiome composition only in infected subjects who remained infected.

3.1 Introduction

Microbiome data are overdispersed multivariate counts; for each sample, counts across multiple taxa are observed. If one is interested in the change of the microbiome composition over time, subjects are measured longitudinally [Ramanan et al. (2016)]. Such data are subject to two sources of correlation, namely the correlation between the counts of a sample and between multiple samples across time of a subject. For this type of data, the available statistical models are still limited.

The microbiome dataset considered in this paper is obtained by sequencing the amplicon of the bacterial 16S rRNA gene, where the sequencing procedure follows the HMP standardized protocol [HMP (2012)]. Chimeric sequences were filtered out and the resulting sequences are either categorized based on similarity into Operational Taxonomical Units (OTUs) followed by annotation, or directly annotated using relevant databases (e.g. Ribosomal Database Project, Greengenes or Silva). The counts for a specific category represent the abundances of the bacteria at a biological taxonomy level. Datasets generated through this sequencing process comprise features that have not been adequately accounted for by currently available statistical methods [Li (2015)]. Firstly, the dataset might be represented by a matrix of taxonomical counts with a compositional structure, which imposes a correlation between taxa [Gloor et al. (2017)]. Secondly, overdispersion might exist due to unobserved heterogeneity in the sampling procedure, the presence of taxa with rare abundance (zero-inflation), and pooling of categories. Another source might be differences in total sequence reads per sample, which might be caused by technical difficulties or by sampling or individual variability. This is commonly addressed by dividing the bacteria for each categories with the total count of the smallest reads (normalization), which results in a constant total bacterial count for all samples. Alternatively, an offset can be used in the model.

Our work is motivated by the microbiome measurements from an epidemio-

logical study carried out in a helminth endemic rural area in Indonesia [Martin et al. (2018)]. The primary research question of this study is to analyze the joint effect of helminth infections and albendazole treatment on the microbial composition comprising multiple bacterial taxa. It has been hypothesized that the presence of helminths is linked with the microbial dysbiosis. However, recent findings report inconsistencies, probably due to limitation in the study design [Ramanan et al. (2016); Cooper et al. (2013); Lee et al. (2014)]. For our study, the stool samples were collected and measured on a subset of subjects participating in a randomized placebo-controlled trial. Thus, we included the microbiome data from infected subjects who received placebo, which makes our study unique. The bacterial count and the helminth infection status were assessed in samples before and 21 months after the first treatment. Details of the study can be found elsewhere [Wiria et al. (2010)]. In a previous paper [Martin et al. (2018)], we identified an effect of treatment on the microbiome composition in subjects who were infected at baseline and at follow up. This relationship was studied in the post treatment samples, whereas the microbiome composition at baseline was not used. Here, we model all the available data simultaneously and hence need to address the correlation structure.

The objective of this paper is to develop a parametric model for the analysis of the overdispersed multivariate count data in the repeated measurement setting. To date, several statistical parametric methods for analysis of microbiome data are available, which take into account the features of the data such as overdispersion and the presence of rare taxa. One approach is to consider a univariate taxa of interest and model the association of this taxa with biological covariates. Several regression models for this simplified problem exist. Zero-inflated models or hurdle models have been proposed to deal with rare taxa [Xu et al. (2015)]. These models are also available for longitudinal studies. This approach however ignores the multivariate structure of the data. A second approach which considers the compositional feature of the microbiome data, models the multivariate count outcome across taxa by a multinomial distribution. To deal with overdispersion, the underlying parameters are assumed to follow the conjugate distribution [Chen and Li (2013)]. This formulation has an advantage that the marginal distribution has a closed form formula.

The correlation due to repeated measurements within the same person is often modelled by including a normally distributed random effect in the linear predictor, i.e., generalized linear mixed model. The overdispersion is typically accounted for by the conjugate distribution Chen and Li (2013); Zhang and Zhou (2017); Guimarães and Lindrooth (2007). Molenberghs et al. (2007, 2010) and Booth et al. (2003) introduced a combined model, where the conjugate distribution for the overdispersion is used and the correlation over time is modelled by normally distributed random effects, i.e., generalized linear mixed model. The authors only consider single categorical count data; hence these models cannot

be directly applied to our data, where we have to acknowledge the compositional feature. Therefore, in spirit of the combined model, we propose an extension of the Dirichlet - multiomial regression model with random effects to incorporate the correlation due to repeated measurements. We will use the reparameterization of the Guimarães and Lindrooth (2007), in which the overdispersion is a function of the covariates and the random effects.

This manuscript is organized as follows. In Section 3.2, we briefly describe the formulation of the loglinear model in the setting of multivariate count data and derive the likelihood of the multinomial distribution obtained by conditioning on the total count. We show the derivation of this method in the case where the count is overdispersed. The model is then extended to include the correlation due to repeated measurements over time. In Section 3.3, simulation studies are described to investigate the performance of the proposed methods and the results of the analyses of the motivating dataset are presented in Section 3.4. In Section 3.5, we conclude and discuss the proposed method.

3.2 Methods

A novel mixed model is considered for the relationship between counts of six phyla categories and the binary variables of infection status and treatment allocation before and after the first treatment round. Due to the normalization, the total count per sample is fixed at 2000 at each time point. Before introducing our new model, we will review various models for categorical count data in the cross-sectional setting: namely for independent count data (the loglinear and the multinomial logistic regression model), and for count data subject to overdispersion (the negative binomial and the Dirichlet-multinomial model) [Agresti (2013); Tutz (2012)].

We first introduce the following notations. Let $C_i^{(t)} = \left\{C_{i1}^{(t)}, \ldots, C_{iJ}^{(t)}\right\}$ be the J dimensional vector of the multivariate microbial count with $C_{ij}^{(t)}$ the abundance of bacteria taxa $j\,(j = 1, \ldots, J)$ for subject $i\,(i = 1, \ldots, N)$ at time point t. The total count for each subject i at time-point t is fixed and denoted as $C_{i+}^{(t)} = \sum_{j=1}^{J} C_{ij}^{(t)}$. Let P be the number of categorical covariates and $X_i^{(t)}$ be the P dimensional vector of covariate values for subject i at time point t. When modelling microbiome data as described above, either the sequence count itself can be considered, or the normalized count related to the total sequence read, i.e. compositional data. Multiple counts distributed over categories are usually represented by a contingency table. We briefly review models for the cross-sectional setting and therefore suppress the superscript t in the model formulation in Subsection 3.2.1.

3.2.1 Cross-sectional setting

The loglinear model for two categorical variables

The loglinear model is commonly used to model the association between multivariate categorical count data and predictors of categorical or continuous value. In the case where all variables are categorical, the data can be represented by a contingency table. Consider two categorical variables E and F, with J and K levels, respectively. The count outcome c_{jk} is associated with the jth level of predictor E and kth level of predictor F, which could be described in a $J \times K$ contingency table (Table 3.1) as follows.

		F		
	j \ k	**1**	…	***K***
E	**1**	c_{11}	…	c_{1K}
	2	c_{21}	…	c_{2K}
	⋮	⋮	⋮	⋮
	J	c_{J1}	…	c_{JK}
Marginal		c_{+1}	…	c_{+K}

Table 3.1: The $J \times K$ Contingency Table

Each cell's count outcome c_{jk} is assumed to follow a Poisson distribution with a mean μ_{jk}. Here, the saturated loglinear model for such contingency table is given by

$$\log\left(\mu_{jk}\right) = \lambda_0 + \lambda_j^E + \lambda_k^F + \lambda_{jk}^{EF}, \tag{3.1}$$

where λ_0, $\lambda_0 + \lambda_j^E$, $\lambda_0 + \lambda_k^F$, and $\lambda_0 + \lambda_k^F + \lambda_j^E + \lambda_{jk}^{EF}$ represent the overall mean, the marginal mean of categorical variable E at the jth level, the marginal mean of variable F at the kth level, and the mean when variables E and F taking the value j and k, respectively. Because there are $J \times K$ cells, the $J + K + JK + 1$ parameters of the saturated loglinear model (3.1) are not uniquely identifiable and thus constraints are needed to ensure the model identifiability. Two sets of constraints are commonly used, namely the baseline and the symmetrical constraint given by

$$\lambda_1^E = \lambda_1^F = \lambda_{j1}^{EF} = \lambda_{1k}^{EF} = 0$$

and

$$\sum_{j=1}^{J} \lambda_j^E = \sum_{k=1}^{K} \lambda_k^F = \sum_{j=1}^{J} \lambda_{j1}^{EF} = \sum_{k=1}^{K} \lambda_{1k}^{EF} = 0, \qquad \text{for all } j,k,$$

respectively. In this manuscript, we use the baseline constraint.

Note that in model (3.1) the response (bacterial categories) E, and predictor F, are exchangeable. The loglinear model (3.1) could be written in the regression format for the bacterial outcome as follows.

$$\log\left(\mu_{jk}\right) = \xi_{0j} + \xi_{1jk}[F = k], \quad j = 1,\ldots,J, \quad k = 1,\ldots,K,$$

with [.] the indicator function. To show the equivalence between two models, note the following j runs over the category and k runs over the predictor levels. For a subject with their predictor in category $k = 1$, the regression model $\{\xi_{01}, \xi_{02}, \ldots \xi_{0J}\}$ with ξ_{0j} for $j = 2,\ldots,J$ corresponds to $\lambda_0 + \lambda_j^E$. For subjects with their predictor in other categories k, the regression model $\{\xi_{01} + \xi_{11k}, \xi_{02} + \xi_{12k}, \ldots, \xi_{0J} + \xi_{1Jk}\}$ where ξ_{1jk} for $j = 2,\ldots J$ corresponds to $\lambda_k^F + \lambda_{jk}^{EF}$. Thus, in the context of regression, the λ_{jk}^{EF} represents the effect of the categorical variable F on outcome category j relative to the reference category.

To estimate the parameters, we assume that each cell's entry represents a realization from the Poisson distribution. The maximum likelihood estimate of $\boldsymbol{\lambda}$ or of $\boldsymbol{\xi}$ can be obtained by maximizing the following likelihood function. Specifically, for subject i, it is given by

$$\begin{aligned} L_i(\boldsymbol{\lambda}) &= \prod_j f_{\text{Pois}}\left(\boldsymbol{\lambda}; c_{ijk}\right) \\ &= \prod_j \frac{\exp\left(-\mu_{jk}\right)\mu_{jk}^{c_{ijk}}}{c_{ijk}!}, \end{aligned} \tag{3.2}$$

where person i belongs to category k and has counts in each bacteria category j. The model could be straightforwardly generalized to incorporate more categorical covariates which results into more than two-way contingency table. For instance, when incorporating the infection and treatment status we will have a three way contingency table. As before, the categorical variable E corresponds to the bacteria category, variable F to the treatment randomization arm and G to the infection status. The corresponding loglinear model can be written as follows

$$\begin{aligned} \log\left(\mu_{jkl}\right) &= \lambda_0 + \lambda_j^E + \lambda_k^F + \lambda_{jk}^{EF} + \lambda_l^G + \lambda_{jl}^{EG} + \lambda_{kl}^{FG} + \lambda_{jkl}^{EFG}, \quad j = 2,\ldots,J; k = l = 2 \\ &\quad \text{or in the regression format as} \\ &= \xi_{0j} + \xi_{1j}\text{Treatment} + \xi_{2j}\text{Infection} + \xi_{3j}\text{Treatment} \times \text{Infection}, \quad j = 1,\ldots,J. \end{aligned} \tag{3.3}$$

Here the baseline constraint is applied on the first equation, while for the second equation this is not needed since there are only $J \times P$ parameters. This last equation represents the loglinear model written in terms of regression coefficients $\boldsymbol{\xi}$ and covariate values, where Treatment and Infection are binary variables. To

assess the statistical significance of the pth covariate ($p = 1, \ldots, P$) on the multivariate count distribution, the null hypothesis $\boldsymbol{\xi}_p = \mathbf{0}$ should be tested. We will use the standard Likelihood Ratio Test which follows a χ^2 distribution with J degrees of freedom.

Multinomial logistic regression

In our data example, the total bacterial count is fixed to a constant for all samples. Under this constraint of a fixed total count, it is sufficient to model the counts for $J-1$ categories and $(J-1) \times P$ parameters are uniquely identified. Guimarães and Lindrooth (2007) showed that the distribution of the multivariate counts under the constraint that the total is a constant could be derived from the distribution of the unconstrained multivariate counts above by using the conditional log likelihood given the total count. When the counts in each category are independently Poisson distributed with mean μ_{jkl}, the total count c_{+kl} follows a Poisson distribution with mean $\sum_{j=1}^{J} \mu_{jkl} = \mu_{+kl}$. The distribution of the multivariate counts conditional on the total for each subject i is therefore given by

$$\begin{aligned}
\Pr\left(\boldsymbol{c}_i = \{c_{1kl}, \ldots, c_{Jkl}\} \,|c_{+kl}\right) &= \frac{\Pr\left(c_{1kl}, \ldots, c_{Jkl}, c_{+kl}\right)}{\Pr\left(c_{+kl}\right)} \\
&= \frac{\prod_{j=1}^{J} f_{\text{Pois}}\left(c_{jkl}; \mu_{jkl}\right)}{f_{\text{Pois}}\left(c_{+kl}; \mu_{+kl}\right)} = c_{+kl}! \prod_{j=1}^{J} \left(\frac{1}{c_{jkl}!}\right) \left(\frac{\mu_{jkl}}{\mu_{+kl}}\right)^{c_{jkl}} \\
&\sim \text{Multinomial}\left(c_{+kl}; \pi_{1kl}, \ldots, \pi_{Jkl}\right), \\
&\quad \text{where } \pi_{jkl} = \frac{\mu_{jkl}}{\mu_{+kl}}. \qquad (3.4)
\end{aligned}$$

Thus, under the baseline constraint and the constraint that the total count is fixed, the distribution of the multivariate count is equivalent to the multinomial distribution with parameter $\pi_j = \dfrac{\mu_j}{\mu_+}$. This model is the multinomial logistic regression model. Note that the parameters $\boldsymbol{\lambda}$ of the loglinear model (3.1) cancel out. In the multinomial logistic regression model, the parameters of the reference category are typically assumed to be equal to zero, although other constraints can be used as well.

Overdispersed count data

When the count data are overdispersed, the variance of the cell count is no longer equal to its expected value and the Poisson distribution cannot be used. A common approach to deal with overdispersion is to assume that the conditional mean

of the count outcome is a random variable following the conjugate distribution. Consider a count at category j and let $\exp(\eta_{ij})$ be the random effect for overdispersion following the Gamma distribution (conjugate for Poisson) with parameter θ. Guimarães and Lindrooth (2007) formulated the model for an overdispersed count outcome as follows:

$$\begin{aligned}
&C_{ij}|\exp(\eta_{ij}) \sim \text{Pois}(\widetilde{\mu}_{ij}), \quad j = 1,\ldots,J \\
&\widetilde{\mu}_{ij} = \exp(\eta_{ij})\,\mu_{ij}, \quad \text{where } \exp(\eta_{ij}) \sim \Gamma\left(\text{shape} = \theta^{-1}\mu_{ij}, \text{rate} = \theta^{-1}\mu_{ij}\right) \\
&\widetilde{\mu}_{ij} = \exp(\eta_{ij})\,\mu_{ij} \sim \Gamma\left(\text{shape} = \theta^{-1}\mu_{ij}, \text{rate} = \theta^{-1}\right).
\end{aligned}$$

Here, μ_{ij} corresponds to the mean of the count in the non-overdispersed model. Now the marginal distribution for the count at category j in person i, C_{ij} can be obtained by integrating out the random effect $\exp(\eta_{ij})$ as

$$\begin{aligned}
\Pr(C_{ij}) &= \int_0^\infty \Pr(C_{ij}|\exp(\eta_{ij}))\, g(\exp(\eta_{ij}))\, d\exp(\eta_{ij}) \\
&= \frac{\Gamma\left(\theta^{-1}\mu_{ij} + C_{ij}\right)}{C_{ij}!\Gamma(\theta^{-1}\mu_{ij})} \left(\frac{1}{\theta^{-1}+1}\right)^{C_{ij}} \left(\frac{\theta^{-1}}{\theta^{-1}+1}\right)^{\theta^{-1}\mu_{ij}}.
\end{aligned}$$

This corresponds to a negative binomial distribution with parameters $\left(\theta^{-1}\mu_{ij}, \frac{\theta^{-1}}{1+\theta^{-1}}\right)$. By the properties of the negative binomial random variable, the total count for subject i also follows the negative binomial distribution

$$C_{i+} \sim \text{NB}\left(\theta^{-1}\mu_{i+}, \frac{\theta^{-1}}{1+\theta^{-1}}\right).$$

The likelihood for subject i in this setting is given by

$$\begin{aligned}
L_i(\theta, \boldsymbol{\lambda}) &= \prod_j f_{\text{NB}}(\boldsymbol{\lambda}, \theta; C_{ij}) \\
&= \prod_j \frac{\Gamma\left(\theta^{-1}\mu_{ij} + C_{ij}\right)}{C_{ij}!\Gamma(\theta^{-1}\mu_{ij})} \left(\frac{1}{\theta^{-1}+1}\right)^{C_{ij}} \left(\frac{\theta^{-1}}{\theta^{-1}+1}\right)^{\theta^{-1}\mu_{ij}}. \qquad (3.5)
\end{aligned}$$

Note that in this setting, the parameter θ which models the overdispersion and the intercept λ_0 are both not identifiable. An often used solution is to absorb the overdispersion parameter into the grand mean λ_0, i.e. $\theta^{-1}\exp(\lambda_0) = \delta_0^{-1}$ Guimarães and Lindrooth (2007).

Overdispersed multinomial

We briefly review the overdispersed count data introduced by Guimarães and Lindrooth (2007) as follows. To guarantee that the parameters of the count for

each category follows a Gamma distribution with the same rate parameter, the overdispersion parameter $\exp(\eta_{ij})$ needs to be a function of the linear predictor μ_{ij}. For such a distribution, Theorem 1 of Mosimann (1962) can be applied. This theorem states that if $\boldsymbol{C}_i = \{C_{i1}, C_{i2}, \ldots, C_{iJ}\}$ are independently Gamma distributed random variables with parameters $(\widetilde{\mu}_{i1}, \widetilde{\mu}_{i2}, \ldots, \widetilde{\mu}_{iJ})$ with the same scale parameter θ^{-1}, then the random variables $\boldsymbol{\Pi}_i = \{\Pi_{i1}, \Pi_{i2}, \ldots, \Pi_{iJ}\}$ with $\Pi_{ij} = \dfrac{C_{ij}}{\sum_{j=1}^{J} C_{ij}}$ have a multivariate beta distribution (Dirichlet distribution) with parameters $\{\widetilde{\mu}_{i1}, \widetilde{\mu}_{i2}, \ldots, \widetilde{\mu}_{iJ}\}$. Note that the Dirichlet distribution is the conjugate for the multinomial distribution. Hence, the marginal distribution for the random variable $\boldsymbol{\Pi}_i$ is obtained by integrating out the Dirichlet random effects. Now, the corresponding Dirichlet - multinomial distribution is given by

$$\Pr(\boldsymbol{\Pi}_i) = \frac{\Gamma(\widetilde{\mu}_{i+})\, C_{i+}!}{\Gamma(\widetilde{\mu}_{i+} + C_{i+})} \prod_{j=1}^{J} \frac{\Gamma(\widetilde{\mu}_{ij} + C_{ij})}{\Gamma(\widetilde{\mu}_{ij})\, C_{ij}!}. \tag{3.6}$$

Alternatively, we consider the conditional likelihood of the multivariate negative binomial given the total count. The contribution for the ith subject is given by

$$\begin{aligned} L_i(\boldsymbol{\lambda}, \theta) = \Pr(\boldsymbol{C}_i | C_{i+}) &= \frac{\prod_{j=1}^{J} f_{\mathrm{NB}}(C_{ij}; \widetilde{\mu}_{ij})}{f_{\mathrm{NB}}(C_{i+}; \widetilde{\mu}_{i+})} \\ &= \frac{\Gamma\left(\theta^{-1}\mu_{i+}\right) C_{i+}!}{\Gamma(\theta^{-1}\mu_{i+} + C_{i+})} \prod_{j=1}^{J} \frac{\Gamma\left(\theta^{-1}\mu_{ij} + C_{ij}\right)}{\Gamma(\theta^{-1}\mu_{ij})\, C_{ij}!}. \end{aligned} \tag{3.7}$$

By $\widetilde{\mu}_{ij} = \theta^{-1}\mu_{ij}$, it follows that the likelihood (3.7) is equivalent to the the Dirichlet-multinomial distribution (3.6). Here, the parameter θ is unidentifiable. Similar to (3.5), we apply the parameterization in Guimarães and Lindrooth (2007) where the overdispersion is absorbed in the grand mean λ_0 such that $\theta^{-1}\exp(\lambda_0) = \delta_0^{-1}$ in the reference category. In contrast to the non-overdispersed multinomial model, the intercepts of the overdispersed multinomial model do not cancel out.

3.2.2 Repeated measurement of overdispersed count

In addition to the overdispersion due to the presence of multiple bacteria within one sample, there is also correlation between measurements of the same person at the two time-points, i.e. at the pre- and post-treatment. To deal with this correlation, we propose to include a random effect u_i in the linear predictor of the model and assume that conditional on this random effect the observations of the two time points are independent. We further assume that the random effect u_i follows a normal distribution with zero mean and variance σ_u^2. The idea of using different distributions for the random effects representing overdispersion and

correlation was introduced by Molenberghs et al. (2007, 2010) and Booth et al. (2003). Molenberghs and Booth modelled the mean of an outcome as a multiplication of overdispersion and the linear predictor. However, to guarantee that the Theorem 1 of Mossimann holds, i.e. that the proportion of each bacterial category has Dirichlet-multinomial distribution, we need to model the overdispersion as a function of the linear predictor.

In the rest of this section, we describe three different mixed models for multivariate count data with overdispersion in the repeated measurement setting using random effects: conditional on the random effect u_i, the counts follow the multivariate negative binomial distribution; the counts follow the conditional multivariate negative binomial distribution given the total count; the proportions (cell's count divided by total count) follow the Dirichlet-multinomial distribution. In all models, we will add the random effect u_i to the linear predictor. These models are therefore extensions of the models for overdispersed multivariate count given in Subsection 3.2.1. Specifically for the first model, we assume that conditional on the random effects $\exp\left(\eta_{ij}^{(t)}\right)$ and u_i, the count $C_{ij}^{(t)}$ follows a Poisson distribution with mean equal to

$$\begin{aligned}
\mathrm{E}\left[C_{ij}^{(t)} \,|\, \exp\left(\eta_{ij}^{(t)}\right), u_i\right] &= \exp\left(\eta_{ij}^{(t)}\right) \exp\left(\widetilde{\mu}_{ij}^{(t)}\right), \\
\text{where } \widetilde{\mu}_{ij}^{(t)} &= \boldsymbol{X}_i \boldsymbol{\xi}_j + u_i, \quad j = 1, \ldots, J. \\
\exp\left(\eta_{ij}^{(t)}\right) &\sim \Gamma\left(\text{shape} = \theta^{-1} \exp\left(\widetilde{\mu}_{ij}^{(t)}\right), \text{rate} = \theta^{-1} \exp\left(\boldsymbol{X}_i \boldsymbol{\xi}_j + u_i\right)\right) \qquad (3.8)
\end{aligned}$$

Thus, given the random effect u_i, the two vectors of counts $C_i^{(t)}$ for $t = 1$ and $t = 2$ are independently distributed and follow the negative binomial distribution. The corresponding likelihood can be written as follows

$$\begin{aligned}
L_{\text{UNBM}}\left(\boldsymbol{\xi}, \theta, \sigma_u^2\right) = \prod_i \Pr\left(\boldsymbol{C}_i^{(t)}\right) &= \prod_i \int_{u_i} \Pr\left(C_{i1}^{(t)}, \ldots, C_{iJ}^{(t)}, u_i\right) d\, u_i \\
&= \prod_i \int_{u_i} \prod_{t=1}^{2} \prod_{j=1}^{J} \Pr\left(C_{ij}^{(t)} | u_i\right) \Pr\left(u_i\right) d\, u_i \qquad (3.9)
\end{aligned}$$

and we denote the regression model under this likelihood to be the unconstrained negative-binomial mixed model (UNBM).

For the second approach, we consider the counts follow the conditional multivariate distribution given the total count. When each categorical count conditional on the total count follows the negative binomial with the same rate parameter, the total count $C_{i+}^{(t)} | u_i$ follows the negative binomial distribution with

parameters $\left(\sum_{j=1}^{J}\exp\left(\eta_{ij}^{(t)}\right)\exp\left(\widetilde{\mu}_{ij}^{(t)}\right), \frac{\theta^{-1}}{1+\theta^{-1}}\right)$. Thus, the corresponding conditional likelihood is given by

$$\begin{aligned} L_{\text{CNBM}}\left(\boldsymbol{\xi}, \theta, \sigma_u^2\right) &= \prod_i \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) \\ &= \prod_i \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)} | u_i\right) \Pr\left(u_i\right) d\, u_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\, u_i} \\ &= \prod_i \frac{\int_{u_i} \prod_{t=1}^{2} \prod_{j=1}^{2} \Pr\left(C_{ij}^{(t)} | u_i\right) \Pr\left(u_i\right) d\, u_i}{\int_{u_i} \prod_{t=1}^{2} \Pr\left(C_{i+}^{(t)} | u_i\right) \Pr\left(u_i\right) d\, u_i}. \end{aligned} \tag{3.10}$$

The model corresponding to this likelihood is denoted as the conditional negative-binomial mixed model (CNBM). However, when the total counts depends on u_i the total count should be a random variable. This is not the case in our dataset. Therefore, we propose the third method with the assumption that the total count is independent of u_i.

In the third approach, we model the multivariate counts in terms of the relative abundance. We assume that the vector of proportions $\Pi_i^{(t)}$ conditional on the random effect u_i follows the Dirichlet multinomial distribution, i.e.

$$\begin{aligned} \left\{\frac{C_{i1}^{(t)}}{C_{i+}^{(t)}}, \ldots, \frac{C_{i1}^{(t)}}{C_{i+}^{(t)}}\right\} | \left\{\alpha_{i1}^{(t)}, \ldots, \alpha_{ij}^{(t)}\right\}, u_i &\sim \text{Mult}\left(\tilde{\pi}_{i1}^{(t)}, \ldots, \tilde{\pi}_{iJ}^{(t)}\right) \\ \left\{\tilde{\pi}_{i1}^{(t)}, \ldots, \tilde{\pi}_{iJ}^{(t)}\right\} &\sim \text{Dir}\left(\alpha_{i1}^{(t)}, \ldots, \alpha_{iJ}^{(t)}\right) \\ \alpha_{ij}^{(t)} &= \theta^{-1}\mu_{ij}^{(t)} \end{aligned} \tag{3.11}$$

where the $\mu_{ij}^{(t)}$ is the linear predictor as in the loglinear model for the Poisson count. With this parameterization, the expected multinomial parameter becomes

$$\widetilde{\pi}_{ij}^{(t)} = \frac{\exp\left(\eta_{ij}^{(t)}\right)\exp\left(\widetilde{\mu}_{ij}^{(t)}\right)}{\sum_{j=1}^{J}\exp\left(\eta_{ij}^{(t)}\right)\exp\left(\widetilde{\mu}_{ij}^{(t)}\right)}.$$

The likelihood for each subject i is then formulated as follows

$$
\begin{aligned}
L_{\mathrm{DMM}}\left(\boldsymbol{\xi}, \theta, \sigma_u^2\right) &= \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}}\right) \\
&= \int_{u_i} \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}} | u_i\right) \Pr\left(u_i\right) d\, u_i \\
&= \int_{u_i} \prod_{t=1}^{2} \frac{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)}\right) C_{i+}^{(t)}!}{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)} + C_{i+}^{(t)}\right)} \prod_{j=1}^{J} \frac{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)} C_{ij}^{(t)}\right)}{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)}\right) C_{ij}^{(t)}!} \Pr\left(u_i\right) d\, u_i. \quad (3.12)
\end{aligned}
$$

The corresponding regression model under this likelihood is denoted as the Dirichlet - multinomial mixed model (DMM). It is shown in the Appendix A, that in the case where the total count does not depend on the random effect u_i, the likelihoods (3.10) and (3.12) are equivalent.

The variance of the random effect u (σ_u^2) represents the correlation between the samples of the same subject across time. However, this value is hard to interpret and the marginal correlation between categorical count outcomes might be more interesting. This correlation is given by

$$
\mathrm{Corr}\left(C_{ij}{}^{(t)}, C_{ij*}^{(t^*)}\right) = \frac{\sigma_{C_{ij}{}^{(t)}, C_{ij*}^{(t^*)}}}{\sqrt{\sigma^2_{C_{ij}^{(t)}} \cdot \sigma^2_{C_{ij*}^{(t(*))}}}}.
$$

The marginal correlation can be computed from Monte Carlo estimates of the first and second moments.

The program language R is used for all the computations except for data application with categorical-specific random effects. When maximizing the likelihoods the integrals are approximated by the adaptive Gauss-Hermite quadrature method [Liu and Pierce (1994)], and we used the functions available in the ecoreg package [Jackson et al. (2008)] to compute the integral. R implementations are available in github (https://github.com/IvonneMartin/CombinedMultinomial)

3.2.3 The categorical-specific random effect

In the above parameterization, we assume that the subject-specific effect u_i is univariate and is the same for all bacteria categories and time-points. Alternatively, a J dimensional vector of random effects can be used. Equation (3.8) now becomes

$$
\begin{aligned}
\widetilde{\mu}_{ij}^{(t)} &= \boldsymbol{X}_i \boldsymbol{\xi}_j + u_{ij}, \quad j = 1, \ldots, J. \\
u_{ij} &\sim \mathrm{MVN}\left(\mathbf{0}, \Delta_{J\times J}\right) \quad (3.13)
\end{aligned}
$$

Here, each bacterial category has its own realization of the random effect and the random effects solely model correlation between the categories over time. The vector $\boldsymbol{u}_i$ of length J follows a multivariate normal distribution with a J by J diagonal variance matrix Δ with σ_j^2 as diagonal elements. In addition to the general model (3.13), we consider a model with common variance $\sigma_j^2 = \sigma_u^2$, for all j to reduce the parameter space. Since the overdispersion already takes care of the correlation among the categories, this model might be better interpretable. However, a drawback of this model is that computation of the likelihood function involves an intractable J dimensional integral.

3.3 Simulation study

3.3.1 Simulation setting

Three sets of simulation studies were conducted to evaluate the performance of the proposed methods. With regard to estimation of the fixed effect parameters and variance components, we first investigated the performance of the DMM models for a subject- and categorical-specific random effects. We reported the bias and MSE as well as the sensitivity and specificity for these parameters. The sensitivity and the specificity of the likelihood ratio test statistics were computed for the following pairs of hypotheses (for fixed and variance of random effect, respectively).

$$\begin{aligned} &\mathrm{H}_0 : \boldsymbol{\xi}_p = \mathbf{0} \quad \text{vs} \quad \mathrm{H}_1 : \text{at least one of } \boldsymbol{\xi}_p \neq 0, \\ &\mathrm{H}_0 : \sigma_u^2 = 0 \quad \text{vs} \quad \mathrm{H}_1 : \sigma_u^2 > 0. \end{aligned}$$

In the second set, we want to estimate the marginal correlation given the distribution of the random effect. The purpose of this study is to verify whether the marginal intraclass correlation observed in our motivating dataset can be represented by our models (UNBM and DMM). For this purpose, we vary the standard deviation of the random effect and we used 10,000 Monte-Carlo simulation for estimating the marginal intraclass correlation.

In the third set, we aimed to study the robustness of the parameter estimates by fitting the DMM models when the true model is UNBM. For this purpose, we generated datasets with three categories from the UNBM model.

Dataset generation

To reduce the computational burden, datasets with only three categories at two different time-points t were considered. The total count per sample S was 25, 50 or 2000, and the number of samples N was 150 or 500. Two sets of parameters were

used, namely $\boldsymbol{\lambda}$ was fixed at $\{\lambda_2^F, \lambda_2^E, \lambda_3^E, \lambda_{22}^{EF}, \lambda_{32}^{EF}\} = \{0.5, -1, 0.1, 0.8, -2\}$ as well as the parameters from the dataset (results are given in Supporting Information Table S1). To increase the power, the parameter values of the first set are relatively larger. Note that the parameter λ_0 is fixed at zero to guarantee identifiability of the overdispersion parameter. The overdispersion parameter was fixed at $\theta = 0.1$. For the standard deviation of the random effects, we considered values σ_u of 0.5, 0.8 and 1.

Specifically, for the Dirichlet-multinomial mixed (DMM) model with a univariate random effect, multivariate counts were generated as follows.

1. For each subject $i, i = 1, \ldots, N$, we randomly generate binary covariates X_i^t for each time point t and a random effect $u_i \sim N\left(0, \sigma_u^2\right)$.

2. The mean for each category j is computed as $\tilde{\mu}_{ij}^{(t)} = \theta^{-1} \exp\left(\boldsymbol{\lambda} + u_i\right)$ where the $\boldsymbol{\lambda}$ correspond to ξ.

3. A multivariate count with mean $\tilde{\mu}_{ij}^{(t)}$ is generated.

For the DMM model with multivariate random effect, a similar procedure was used except that the random effects in step (1) are now generated from the multivariate normal distribution with a diagonal covariance matrix Σ. We considered three sets of values for the standard deviations of random effects, namely $\boldsymbol{\sigma}_u = \left(\sigma_{u_1}, \sigma_{u_2}, \sigma_{u_3}\right)$ is (0.5,0.6,0.5), (0.8,0.9,0.8) or (1,0.9,1).

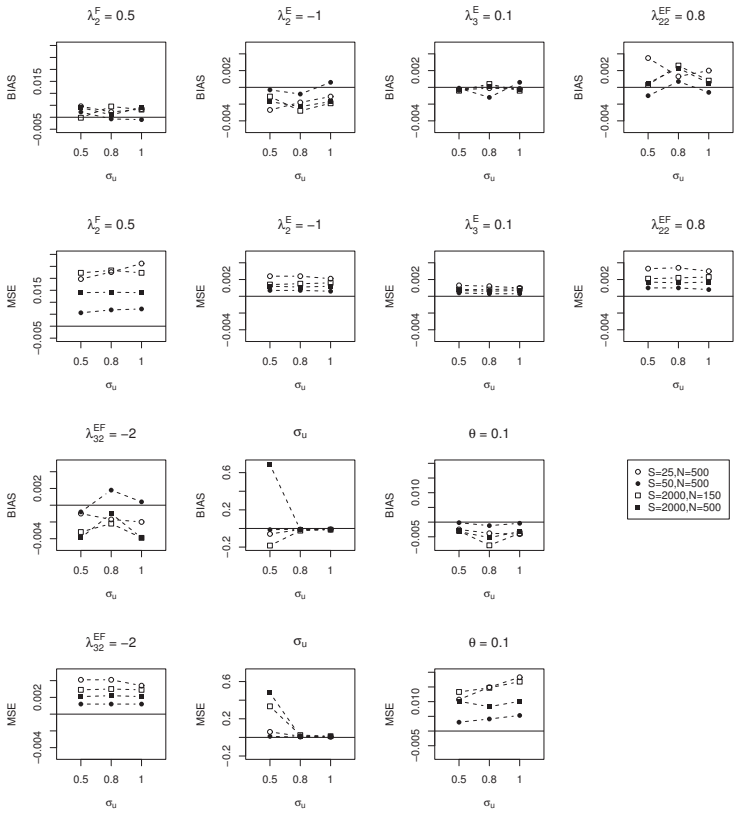
For the second set of simulation, 6 bacterial categories are used and parameters for the simulation are obtained from the dataset. Finally for the unconstrained negative binomial mixed (UNBM) model, the second step was replaced by computation of the expected count outcome for each category j of $\tilde{\mu}_{ij} = \theta^{-1} \exp\left(\log(S) + \boldsymbol{\lambda} + u_i\right)$. Here the offset $\log(S)$ is incorporated to take into account the total bacteria count S. For each scenario mentioned above, 1000 replicates were generated. The models were fitted to each of the replicates.

3.3.2 Simulation results

Evaluation of DMM model

The performance of the method in estimating the parameters is described in Figure 3.1. Overall, the bias and MSE appears to be improved when either the total bacterial count (from $S = 25$ to $S = 50$ and the sample size was $N = 500$), or the sample size was increased (from $N = 150$ to $N = 500$ and the total count was $S = 2000$). For small value of σ_u, both the bias and the MSE of this estimate are relatively large. Similar results are obtained for the model with categorical-specific random effects (Figure S1). The sensitivity of the likelihood ratio test for the fixed effects parameters that are obtained from the dataset are very low for all scenarios except when the total sample size is large (Table S2A). For testing the zero

variance component, the likelihood ratio test has a high sensitivity and specifity when the sample size and variance component are large (Table S2B).



$\boldsymbol{\lambda}$: a vector of parameters in loglinear model.

σ_u: the standard deviation of the between individual variation.

θ : the overdispersion.

Figure 3.1: Bias and MSE of datasets generated from the DMM model with subject-specific random effect.

	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u)$ =-1.309	$\log(\theta)$ = -2.302	Loglik
Subj-sp	0.418(0.130)	-0.959(0.059)	0.096(0.050)	0.765(0.071)	-1.900(0.075)	-1.680(0.754)	-1.886(0.094)	-3918.581(18.333)
Cat-sp	0.438(0.172)	-1.007(0.116)	0.108(0.109)	0.809(0.084)	-2.017(0.086)	-0.704(0.005)	-2.366(0.125)	-3961.971(14.865)
	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u)$ =-0.693	$\log(\theta)$ = -2.302	Loglik
Sub-sp	0.359(0.130)	-0.882(0.080)	0.096(0.077)	0.695(0.088)	-1.739(0.095)	-0.973(0.334)	-1.320(0.099)	-4064.461(18.503)
Cat-sp	0.462(0.170)	-1.000(0.122)	0.110(0.117)	0.794(0.085)	-1.997(0.083)	-0.607(0.028)	-2.253(0.129)	-3988.959(16.392)
	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u)$ =-0.223	$\log(\theta)$ = -2.302	Loglik
Sub-sp	0.300(0.132)	-0.766(0.099)	0.092(0.105)	0.602(0.11)	-1.509(0.121)	-0.766(0.196)	-0.698(0.112)	-4171.966(17.482)
Cat-sp	0.455(0.194)	-1.004(0.173)	0.099(0.17)	0.795(0.089)	-1.985(0.096)	-0.237(0.049)	-2.235(0.165)	-4011.262(19.136)
	λ_2^F =0.5	λ_2^E =-1	λ_3^E = 0.1	λ_{22}^{EF} = 0.8	λ_{32}^{EF} =-2	$\log(\sigma_u) = 0$	$\log(\theta)$ = -2.302	Loglik
Sub-sp	0.270(0.129)	-0.691(0.112)	0.092(0.117)	0.541(0.122)	-1.376(0.133)	-0.699(0.187)	-0.367(0.117)	-4193.591(19.342)
Cat-sp	0.449(0.200)	-1.003(0.213)	0.100(0.197)	0.795(0.088)	-1.985(0.101)	-0.020(0.046)	-2.225(0.177)	-3993.517(23.146)

Each rows started with Sub-sp represents the estimates (standard deviation) when datasets were fitted with the DMM model with subject-specific random effect and rows started with Cat-sp represents the estimates (standard deviation) when datasets were fitted with DMM model with categorical-specific random effect having common variance.
$\boldsymbol{\lambda}, \sigma_u, \theta$ are as explained in Figure 3.1.
Loglik represents the loglikelihood value obtained using the corresponding model.
Rows in gray represent the estimation when the standard deviation of the normally distributed random effect is small.

Table 3.2: The mean estimates (standard deviation) over 1000 replicates when datasets were generated from the DMM model with categorical-specific random effect with common variance.

Since the model with the categorical-specific random effect is time consuming to fit, we also investigate the robustness of assuming a subject-specific random effect while the datasets were generated by using a vector of random effects following the multivariate normal distribution. The results are given in Table 3.2. It appears that for a random effect with smaller standard deviation ($\log(\sigma_u)$ of -1.309), the biases of the estimates of fixed effect parameters and of $\log(\sigma_u)$ are relatively small, while for a random effect with larger standard deviation $\log(\sigma_u) = 0$ (σ_u of 1) the biases are relatively large.

In Table S3, the marginal correlations are given for the subject-specific random effects. It appears that the correlation between categories are all negative and the correlation between samples across time are very small. These results are not affected by the standard deviation of the random effect for our considered values. Table S4 lists the marginal correlations using categorical-specific random effects where each category-specific random effect has the same standard deviations σ_u. We notice that a part of the correlations between categories is now positive and the correlation between the same categories across time are larger. Moreover, these correlations tend to increase with a larger variance of the random effects.

Simulations under the UNBM model

The marginal correlations for the UNBM with a subject-specific random effect are listed in Table S5. It appears that the correlations between categories are positive as well as negative. The correlations of the same category between time points are all positive and increase with σ_u. A similar result is observed for the UNBM model with categorical-specific random effects (Table S6) although here the correlation varies more across categories.

Next, we investigated the robustness of the models. Datasets were generated using the multivariate negative binomial mixed model without conditioning on the total count (UNBM model). The results of fitting the unconditional multivariate negative binomial mixed model (UNBM), the multivariate negative binomial mixed model conditional on the total (CNBM) and the Dirichlet-multinomial mixed model (DMM) are given in Figure 3.2 for the fixed effect parameters and Figure 3.3 for the variance component.

In general, the fixed effect parameters obtained from these three different models are unbiased except the estimates of the intercepts (λ_2^F) for the CNBM model and the DMM model. Since the model used for analysis and generating the data are the same, the estimates of the fixed effect parameters in Figure 3.2 are unbiased and the variance of the estimator decreases when the total count was increased (from $S = 25$ to $S = 50$) or the sample size is increased (from $N = 150$ to $N = 500$). When using the conditional distribution given the total, the estimates of the fixed effect parameters in Figure 3.2 are biased when the total bacterial count is small ($S = 25$ and $S = 50$). When the total count is relatively large ($S = 2000$), the estimates of the fixed effects (including the intercept λ_2^F) are less biased. When

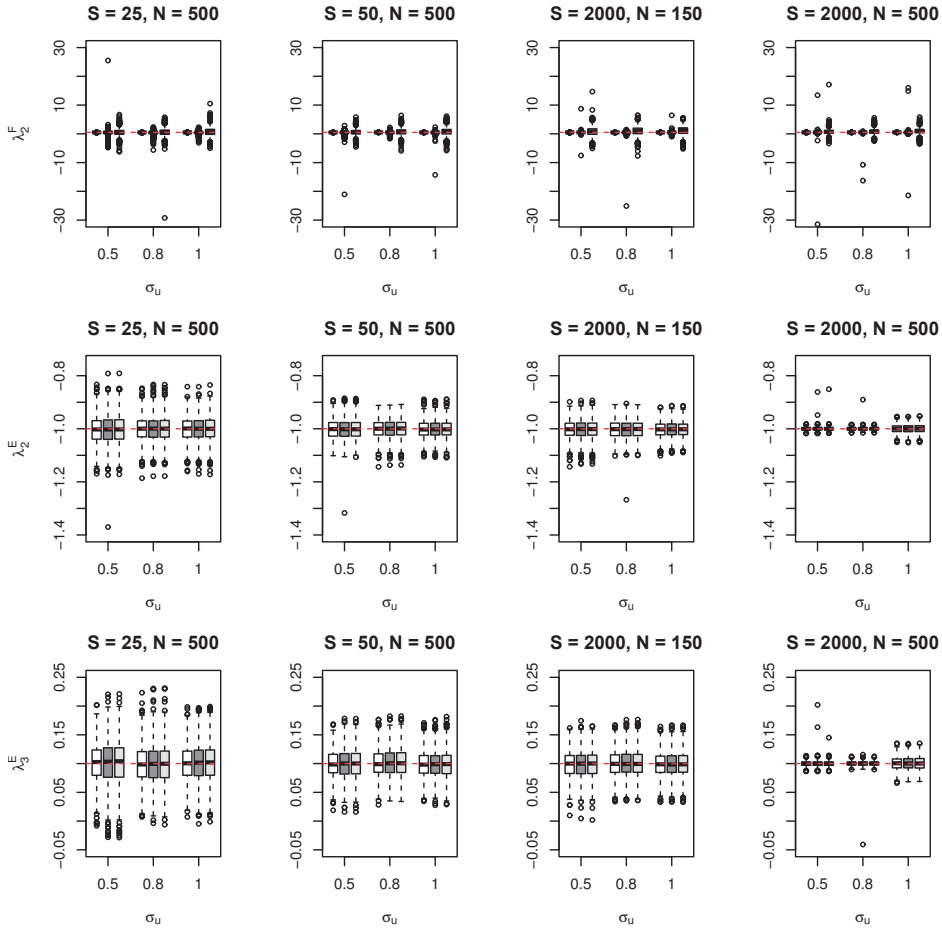


Figure 3.2: Estimates of the fixed effect and overdispersion parameters obtained from three different models (UNBM, CNBM and DMM) when datasets were generated using UNBM model. (first part)

estimating the fixed effect parameters using the DMM model, the estimate of the fixed effects are unbiased except for the intercept term λ_2^F and increasing the sample size does not improve the estimation.

The estimates of the random effect parameters in the UNBM model are unbiased and by increasing the total bacterial count or the sample size improves the precision. In the CNBM model, when the total bacterial count is small ($S = 25$ and $S = 50$), we observe that the standard deviation of u_i is overestimated and that the bias in the estimate of the overdispersion parameter is small. When the total count is large $S = 2000$, the estimate of the standard deviation of u_i appears

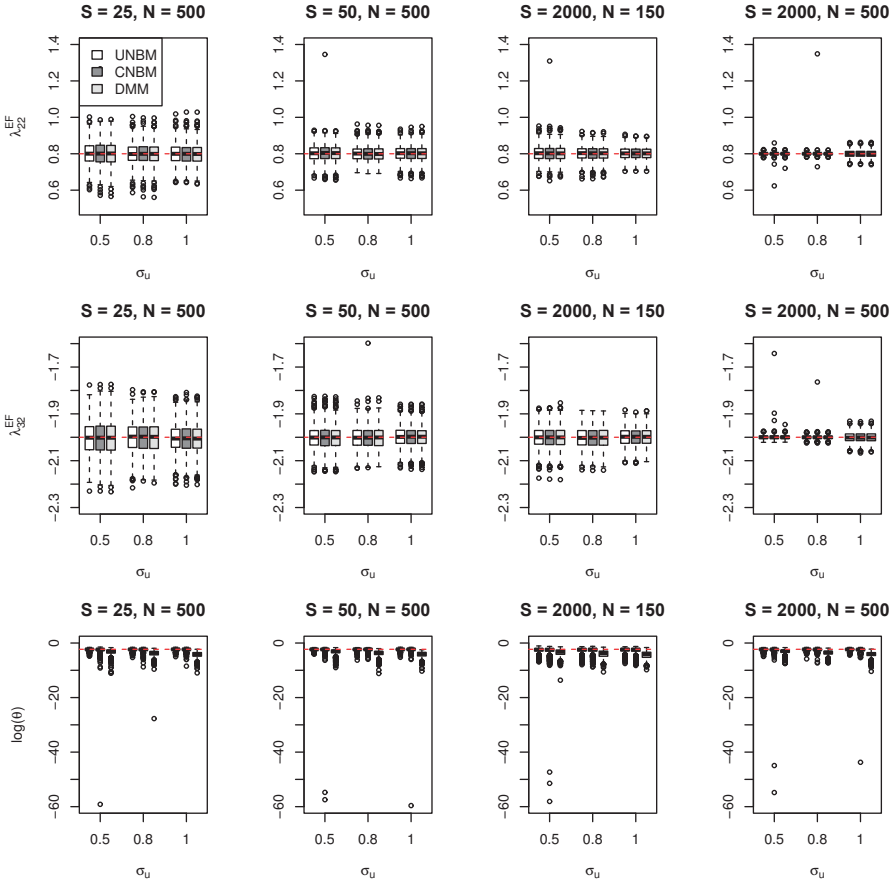


Figure 3.2: Estimates of the fixed effect and overdispersion parameters obtained from three different models (UNBM, CNBM and DMM) when datasets were generated using UNBM model. (cont.)

to be less biased while the overdispersion parameters is underestimated. When fitting the DMM model to the data, the estimates of the random effect parameters are biased in all scenarios.

3.4 Data Application

We used the DMM models to analyze the effect of helminth infections and treatment on microbiome composition. For this purpose, we first consider the fixed effect structure and fitted several DMM models to our dataset assuming (com-

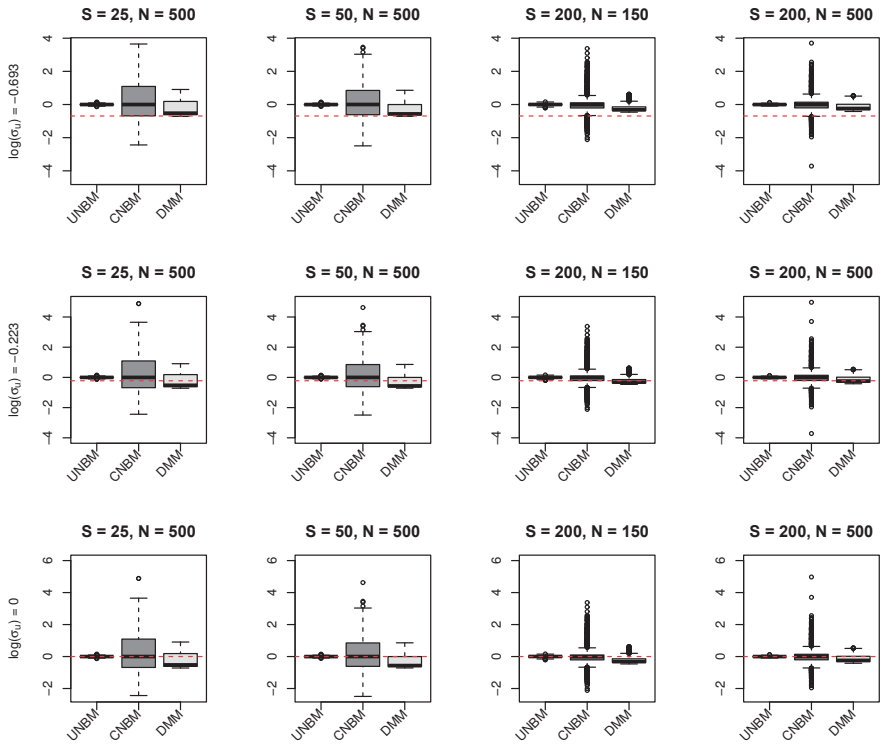


Figure 3.3: Estimates for the variance components obtained from three different models (UNBM, CNBM and DMM) when datasets were generated using UNBM model.

mon) random effect for each category. Next, we will investigate the best random effect structure and we will verify whether the parameter estimates of the fixed effects are affected by the random effect structure.

The microbiome dataset considered here was measured in a subset of a randomized clinical trial performed in a helminth-endemic area in Nangapanda subdistrict, Indonesia, described elsewhere [Wiria et al. (2010)] and is publicly available at Nematode.net (http://nematode.net/Data/Indonesia_16S/S1_Table.xlsx). In brief, households were randomized to receive either a single dose of 400 mg albendazole or placebo, once every three months for a period of one and a half years. To assess the effect of treatment on the prevalence of soil transmitted helminth infections, yearly stool samples were collected on a voluntary basis. *T. trichiura* infection was detected by microscopy and a multiplex real time PCR was used to detect the DNA of hookworm (*Ancylostoma duodenale* or *Necator americanus*) and *Ascaris lumbricoides*. A subject was regarded as infected if it was in-

fected with at least one helminth species.

For the current study, paired DNA samples before and at 21 months after the first treatment round from 150 inhabitants in Nangapanda were selected based on the treatment allocation and infection status, as well as the availability of complete stool data at pre- and post-treatment. The procedure for sample collection and processing was already described in Wiria et al. (2010). The 16s rRNA gene from the stool samples were processed through the 454 pyrosequencing technique, and the classification of the sequence resulted in counts of 18 bacterial phyla. For the current analyses, we retained the 5 most prevalent phyla and pooled the remaining into one category, resulting in six phyla categories: *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, unclassified, and pooled category.

The description of relative abundance of each bacterial phyla at each time points are given in Table S7. *Firmicutes* has the highest relative abundance at each time points (around 68%), followed by *Actinobacteria* (around 12%), *Proteobacteria* (around 10%), *Bacteroidetes* (around 6 %) and Unclassified and pooled category (each around 1%). The dispersions are estimated by the ratio between the variance and mean. All bacteria counts show dispersion larger than 1 indicating the presence of overdispersion. Since zero-inflation might lead to overdispersion, we investigated the number of the samples with zero counts for the six categories at the two time points. Only for the following three categories, a small number of samples with zero counts was observed: *Bacteroidetes* (5 samples at post-treatment), Unclassified bacteria (1 at pre-treatment and post-treatment), and the pooled category (15 at pre-treatment and 6 at post-treatment). The corresponding histograms can be found in Figure S2. From this, we conclude that zero-inflation is not present, hence the overdispersion is probably caused by other sources. We will therefore account for overdispersion by additional random effects.

Table 3.3 gives the observed correlations between categories and of categories between time points. The order j for $C_j^{(t)}$ are *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, *Proteobacteria*, Unclassified and pooled category. The observed correlations between *Firmicutes* and the three most abundant bacteria (*Actinobacteria*, *Proteobacteria* and *Bacteroidetes*) are relatively high and negative (around -0.50), indicating an increase of *Firmicutes* corresponds to the decrease of these bacterial categories. These correlations are relatively similar for both time points, except for the correlation between *Firmicutes* and *Actinobacteria* which becomes smaller at the second time point (-0.27). The correlations between *Firmicutes* and Unclassified, and the pooled category, are relatively small. The intraclass correlations of bacterial categories between the two time points are always positive. *Firmicutes* and *Actinobacteria* show the highest correlation between two time points (0.14 and 0.17).

The baseline characteristics of the study participants were given in Table 3.4. In each of the randomization arms, there are four possible combinations of infection status at pre- and post-treatment. Namely, uninfected subjects who either

	$C_1^{(1)}$	$C_2^{(1)}$	$C_3^{(1)}$	$C_4^{(1)}$	$C_5^{(1)}$	$C_6^{(1)}$	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$	$C_4^{(2)}$	$C_5^{(2)}$	$C_6^{(2)}$
$C_1^{(1)}$	1	-0.46	-0.43	-0.48	-0.12	-0.23						
$C_2^{(1)}$	·	1	-0.29	0.13	0.02	0						
$C_3^{(1)}$	·	·	1	-0.27	-0.19	0						
$C_4^{(1)}$	·	·	·	1	0.1	0.06						
$C_5^{(1)}$	·	·	·	·	1	0.01						
$C_6^{(1)}$	·	·	·	·	·	1						
$C_1^{(2)}$	**0.14**	-0.11	-0.05	-0.01	0	-0.13	1	-0.27	-0.53	-0.57	0.04	-0.14
$C_2^{(2)}$	-0.14	**0.17**	0.04	0.03	-0.01	-0.05	·	1	-0.27	-0.15	-0.05	0.01
$C_3^{(2)}$	0.04	0.05	**0.01**	-0.07	-0.08	-0.1	·	·	1	-0.07	-0.22	-0.11
$C_4^{(2)}$	-0.11	-0.02	0.01	**0.07**	0.05	0.3	·	·	·	1	0.02	0.09
$C_5^{(2)}$	0.06	-0.25	0.09	0.01	**0.05**	0.01	·	·	·	·	1	-0.05
$C_6^{(2)}$	-0.07	0.08	-0.06	-0.01	0.23	**0.17**	·	·	·	·	·	1

$C_j^{(t)}$ represents the bacterial phyla $j, j = 1, \ldots, 6$ at time point t. The order of j are *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, *Proteobacteria*, Unclassified and pooled category.

Table 3.3: The observed marginal correlation of the motivating dataset.

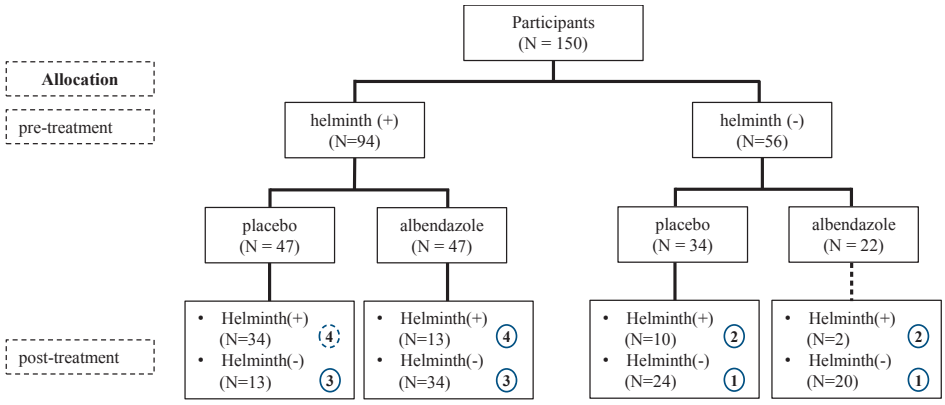


Figure 3.4: The profile of the microbiome study. The chart shows the number of subjecs infected with at least one of the prevalent soil transmitted helminths (Helminth (+)) or free of helminth infections (Helminth (-)) that belonged to either the placebo or albendazole treatment group, at pre-treatment and 21 months after the first treatment round. The circled number represents the condition explained in Section 3.4.

remained uninfected (condition 1) at post-treatment or became infected at post-treatment (condition 2) and infected subjects who either became uninfected at post-treatment (condition 3) or remained infected at post-treatment (condition 4). The number of samples in each conditions at pre- and post-treatment are given in Figure 3.4. It has been shown previously that treatment had an effect on the

Characteristics	albendazole arm (N = 69)	placebo arm (N = 81)
Age (in years) ,mean(SD)	27.38 (16.5)	27.85 (16.91)
Sex, female, n(%)	39 (56.5)	45 (55.6)
Helminth Infections, n(%)		
A. lumbricoides	17 (24.6)	18 (22.2)
Hookworm	26 (37.7)	23 (28.4)
N. americanus	25 (36.2)	23 (28.4)
A. duodenale	2 (2.9)	2 (2.5)
T. trichiura	20 (28.9)	22 (27.2)
Any helminth	47 (68.12)	47 (58.0)
Proportion (in %) of the 6 most abundant bacteria phyla, mean (SD)		
Actinobacteria	12.5 (8.9)	11.0 (7.9)
Bacteroidetes	7.4 (11.3)	6.4 (11.0)
Firmicutes	66.8 (13.5)	70.0 (13.7)
Proteobacteria	9.8 (7.9)	9.2 (8.4)
Unclassified*)	2 (2.22)	2.7 (3.2)
Pooled#)	1.5 (3. 7)	0.7 (1.2)

Table 3.4: **Characteristics at baseline for study participants**.
*)Unclassified represents sequences that cannot be assigned to a phyla.
#)Pooled category consists of the remaining 13 phyla having average relative abundance among samples less than 1%.

composition at post-treatment in infected subjects who remained infected (condition 4)[Martin et al. (2018)]. Here, we want to reanalyze this dataset by using a joint model for the microbiome data at pre- and post-treatment to assess the treatment effect in the infected subjects who remained infected. Additionally, we want to estimate the time effect, while adjusting for other variables such as infection status and treatment allocation. The following loglinear model is considered. Let D, E, F, G, H represent the categorical variables: bacterial taxa, infection (INF), treatment (TRT), baseline infection status (BHelm), and time (t) with J, K, L, M, N levels for each variable. For bacterial phyla, the *Firmicutes* was considered as a reference category. Now the following model was fitted to the data

$$\begin{aligned}\log\left(\mu_{ijklm}^{DEFGH}\right) = &\left(\log\left(\delta_0^{-1}\right) + \lambda_j^D\right) + \left(\lambda_k^E + \lambda_{jk}^{DE}\right) + \left(\lambda_n^H + \lambda_{jn}^{DH}\right) + \\ &\left(\lambda_{ln}^{FH} + \lambda_{jln}^{DFH}\right) + \left(\lambda_{mn}^{GH} + \lambda_{jmn}^{DGH}\right) + \left(\lambda_{lmn}^{FGH} + \lambda_{jlmn}^{DFGH}\right) + \\ &\left(\lambda_{klmn}^{EFGH} + \lambda_{jklmn}^{DEFGH}\right) + u_i, \\ &\text{with the baseline constraint at } J = K = L = M = N = 1, \\ &u_i \sim N\left(0, \sigma_u^2\right) \end{aligned} \tag{3.14}$$

Alternatively the model could be written in terms of regression coefficients as follows.

$$\begin{aligned}\log\left(\mu_{ij}^{(t)}\right) = &\xi_{0j} + \xi_{1j}\text{INF} + \xi_{2j}t + \xi_{3j}\text{TRT} \times t + \xi_{4j}\text{BHelm} \times t + \\ &\xi_{5j}\text{BHelm} \times \text{TRT} \times t + \xi_{6j}\text{INF} \times \text{BHelm} \times \text{TRT} \times t + u_i\end{aligned}$$

where $\xi_{0j} = \log\left(\delta_0^{-1}\right) + \lambda_j^D$, $\xi_{1j} = \lambda_l^F + \lambda_{jl}^{DF}$, and so forth. In this model, there are 6×7 estimable covariate effects on each bacterial phyla. In condition 4, the difference in the microbiome composition between the albendazole and placebo arm is represented by $\xi_{3j} + \xi_{5j} + \xi_{6j}$, while in condition 3, the difference in the microbiome composition between two arms by $\xi_{3j} + \xi_{5j}$. In the subjects who are uninfected at baseline the treatment effect is represented by ξ_{3j}, irrespective of their infection status at post-treatment. The change of microbiome composition, when subjects were uninfected at baseline, remained uninfected at post-treatment, and received placebo, is modelled by ξ_{2j}. Two interaction terms with BHelm were included in this model (3.14) (i.e. the coefficient ξ_{4j} and ξ_{5j}) to model the effect of having infection at pre-treatment and still being infected at follow up, irrespective of treatment by albendazole. The coefficient ξ_{4j} represents the effect of having infection at pre-treatment in the placebo group. We first included a subject-specific random effect u_i in the model. Statistical significance for each covariate was assessed by the likelihood ratio test with 6 degrees of freedom and the significance of the random effect was assessed using the likelihood ratio test with mixture of $\chi^2_{[0,1]}$ distribution.

The parameter estimates from the loglinear model with subject-specific random effects (3.14) are given in Table S8. The between subject variation over time is estimated by the standard deviation σ_u of 0.269 (s.e. of 0.053). The variance of this random effect is significantly different from zero (p-value < 0.001, LRT with mixture of $\chi^2_{[0,1]}$ distribution), indicating that the microbiome counts of a person over time are correlated. The regression coefficients for the covariates BHelm$\times t$ (ξ_{4j}) and BHelm$\times$TRT$\times t$ (ξ_{5j}) appear not to be significantly associated

with the microbiome ($p-$values > 0.05), indicating that having infection at pre-treatment does not influence the microbiome composition. These two covariates were present at the second time point for subjects in condition 3 and 4. Being the terms $\xi_{4j} + \xi_{5j}$ almost zero for all categories, the change of microbiome in these conditions appears to be not affected by these two covariates.

To obtain a model with less parameters, we first eliminated the covariate BHelm$\times$TRT$\times t$. The covariate BHelm$\times t$ was also not significant in this reduced model ($p-$value of 0.795). Hence, we reduced the model (3.14) further by eliminating this covariate. In this updated model, BHelm$\times$TRT$\times t$ was still not significant ($p-$value of 0.843). Finally, we fitted the following model

$$\log\left(\mu_{ij}^{(t)}\right) = \xi_{0j} + \xi_{1j}\text{INF} + \xi_{2j}t + \xi_{3j}\text{TRT} \times t + \xi_{4j}\text{INF} \times \text{BHelm} \times \text{TRT} \times t + u_i. \tag{3.15}$$

In this final model for fixed effects assuming a subject-specific random effect (3.15), 6×4 parameters represent the covariate effects on the microbiome composition. The treatment effect is modelled by ξ_{3j} for all conditions except for condition 4. The difference in the microbiome composition in condition 4 between the albendazole and placebo arm is represented by $\xi_{3j} + \xi_{4j}$. The estimated log odds ratio for each bacterial category compared to *Firmicutes* is given in Table S9. Also for this model the standard deviation of random subject-specific effect u_i is significantly greater than zero ($p-$value < 0.001). Albendazole has no direct effect in subjects who remained uninfected as the odds ratios for each bacterial category are approximately 1. On the other hand, when subjects remained infected, the odds of *Actinobacteria* to *Firmicutes* at the second time point compared to the first time point increases about 55% while the odds ratio for *Bacteroidetes* to *Firmicutes* decreases about 62%.

Next we considered a 6 dimensional random effects structure for this data. We fitted DMM model (3.15). The results are listed in Tables 3.5 and S10. Overall, the estimates of the fixed effects and overdispersion are very similar for these random effect structures. This is in line with the result of the simulation study. However, when we fitted the DMM model with categorical-specific random effects, we observed the following; while the estimated variance component over time for the first three categories are relatively large ($\sigma^2_{u_1} = 0.369$ to $\sigma^2_{u_3} = 0.536$), for the last three categories (*Proteobacteria*, Unclassified and Pooled) are small and hence the random effects for these categories can be omitted.

Finally, we investigated whether the correlations induced by the model correspond to the observed correlations; the marginal correlation induced by the DMM model with a subject-specific random effect (Table S11A),a categorical specific random effect with common variance (Table 3.6) and with categorical-dependent variance for the random effects (Table S11B). For all DMM models, the pairwise correlations at each time points between *Firmicutes* and the other three preva-

Categories	INF	t	TRT$\times t$	Bhelm$\times$INF$\times$TRT$\times t$
Actinobacteria	-0.006 (-0.218, 0.207)	0.050 (-0.155, 0.256)	0.046 (-0.235, 0.326)	0.326 (-0.042, 0.694)
Bacteroidetes	0.220 (-0.056, 0.496)	-0.119 (-0.395, 0.157)	-0.012 (-0.381, 0.356)	-0.916 (-1.573, -0.259)
Protobacteria	0.171 (-0.054, 0.396)	0.056 (-0.161, 0.273)	0.035 (-0.256, 0.326)	0.026 (-0.376, 0.427)
Unclassified	-0.024 (-0.304, 0.257)	0.129 (-0.149, 0.407)	-0.099 (-0.476, 0.277)	-0.159 (-0.727, 0.410)
Pooled	0.166 (-0.158, 0.490)	0.195 (-0.124, 0.515)	-0.030 (-0.449, 0.388)	-0.180 (-0814, 0.454)
Loglik	-8285.5	$\widehat{\theta}$ (s.e)	0.08 (0.01)	
$\hat{\sigma}_u$ (s.e)	0.22 (0.03)			

*Fitted with SAS procedure NLMIXED with 3 quadrature points of Adaptive Gauss-Hermite approximation.

Table 3.5: The log odds ratio (95% CI) when dataset were fitted with DMM with categorical-specific random effect having common variance.*

lent bacterial phyla are relatively high and similar to the observed marginal correlations (Table 3.6, Table S11A-B). With regard to the correlation of categories between the two time points, the DMM model with categorical-specific random effects with common variance showed a similar correlation structure to the observed one (Table 3.6). For the DMM model with categorical-specific random effect, the correlation between the same category over time seems to be too high compared to the dataset (Table S11B). Therefore, we concluded that the DMM model with a categorical specific random effect having common variance across categories is the model which describes our data best.

	$C_1^{(1)}$	$C_2^{(1)}$	$C_3^{(1)}$	$C_4^{(1)}$	$C_5^{(1)}$	$C_6^{(1)}$	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$	$C_4^{(2)}$	$C_5^{(2)}$	$C_6^{(2)}$
$C_1^{(1)}$	1	-0.55	-0.35	-0.51	-0.3	-0.22						
$C_2^{(1)}$	·	1	-0.06	-0.09	-0.05	-0.03						
$C_3^{(1)}$	·	·	1	-0.04	-0.03	-0.02						
$C_4^{(1)}$	·	·	·	1	-0.05	-0.03						
$C_5^{(1)}$	·	·	·	·	1	-0.02						
$C_6^{(1)}$	·	·	·	·	·	1						
$C_1^{(2)}$	**0.19**	-0.12	-0.06	-0.1	-0.05	-0.04	1	-0.57	-0.3	-0.51	-0.3	-0.23
$C_2^{(2)}$	-0.12	**0.13**	0.03	0.03	0.02	0.02	·	1	-0.07	-0.08	-0.06	-0.04
$C_3^{(2)}$	-0.05	0.02	**0.05**	0.02	0.01	0.01	·	·	1	-0.05	-0.02	-0.01
$C_4^{(2)}$	-0.1	0.02	0.02	**0.12**	0.02	0.01	·	·	·	1	-0.05	-0.03
$C_5^{(2)}$	-0.05	0.02	0.01	0.02	**0.05**	0.01	·	·	·	·	1	-0.02
$C_6^{(2)}$	-0.04	0.02	0.01	0.01	0.01	**0.03**	·	·	·	·	·	1

Table 3.6: The estimated marginal correlation of the dataset obtained by DMM model with categorical-specific random effect having common variance across categories.

3.5 Discussion

We proposed a novel parametric multivariate method to model microbiome data from an epidemiological study using a repeated measurements design. Current parametric models that account simultaneously for overdispersion and repeated measurements use a combination of a conjugate and a normal distribution. This method was introduced by Booth et al. (2003) for count data. Molenberghs et al. (2010) reviewed the combined model for the binary [Molenberghs et al. (2012)] and time-to-event data [Efendi et al. (2014)]. The multinomially-distributed data were however not considered in these papers. The rationale of this combined model is the simplification to the parent distribution when overdispersion is absent and furthermore, the conditional distribution given the normally distributed random effect has a closed-form formula which reduces computational time. Thus, this model has an advantage over the generalized linear mixed models where multivariate normal distributions were used to model correlation due to overdispersion and repeated measurements. Our proposed model is also an extension of the econometrics model for the analysis of choice probabilities in the cross-sectional setting [Guimarães and Lindrooth (2007)]. We considered three models for the analysis of repeatedly measured microbiome data, namely models corresponding to the unconditional distribution and to conditional distribution given the total count of a sample. For the latter distribution, we considered the situations where the total counts either vary or are fixed. We showed that for the last situation, i.e. total count is fixed, the likelihood is equivalent to the likelihood of the multinomial logistic model. Since in our dataset the total number of counts per sample is constant we prefer to use the DMM model.

In a simulation study, we showed that the DMM model provides unbiased estimates for the fixed and random effects independent of the used random effect structure to model the correlation between subjects across time. The sensitivity of the likelihood ratio test for the fixed and random effect components are relatively high when the sample size is large as in the case of our data application. We also showed that the models provided similar estimates for the fixed and random effects when datasets were generated from DMM model with different random effect structure. Two structures of the random effects were considered in the DMM model; one is the simplest subject-specific random effect where the variation of each categorical count outcomes is the same, and the second is to assume a diagonal covariance structure with the same variance for each category. With regard to the marginal correlation for each category between time points, we observed that different correlations can be obtained by changing the random effect structure. The simple random effect structure provides small correlations while for the model with categorical-specific random effect, the correlations are larger and increase with the size of variance component. Hence, if the interest is solely on the fixed effects and random effect estimates, the simple model with subject-specific

random effect can be used. On the other hand, when the correlation structure between the same category across time is of interest, a more complex DMM model with categorical-specific random effects should be used.

For our data application, we were interested in the parameters modelling the variability between subjects and the effect of covariates on microbiome composition therefore we used a subject - specific random effect. Following the generalized linear mixed model framework, the random effect $\boldsymbol{u}_i$ is linked to the expected outcome and measures the variation of the count outcome for certain category between subjects. The variability of the categorical count between subjects is then captured by a single estimate of the standard deviation of the random effect and its significance reveals that the variability between subjects should be taken into account in the model. The estimate of the standard deviation in our data analysis is 0.269 (s.e. of 0.053, p-value < 0.001) which is relatively small hence our assumption of a subject-specific random effect is justified. The standard deviation although small is significant hence our extensive model is necessary for this data. With regard to the fixed effects, their estimates describe the contribution of the covariate to the odds ratio of two bacterial categories. One advantage from our model is to model the change of microbiome in different strata over time. For instance, we showed in the motivating dataset that the change of microbiome over time in subject who remained uninfected in the placebo arm could be inferred from the estimate of the time coefficients. Using the same model, we could also infer the change of microbiome when subjects remained unifected in the albendazole arm as well as the change of microbiome when subjects remained infected. In the previous analyses, we selected subjects who were infected at pre-treatment and fit the Dirichlet-multinomial regression at post-treatment to observe the effect of having long term infection and treatment on microbiome composition. The statistical test using that model showed that subjects who remained infected and received albendazole harbored significantly different microbiome composition compared to subjects who remained infected and received placebo. This result is confirmed by the analysis in this manuscript.

On the other hand, for the data application, when the interest is on the marginal correlation, the random effect structure has to be correctly modelled. For our dataset, we considered three structures, namely subject-specific random effects, categorical-specific random effects with common variance and with categorical-dependent variances. The second correlation structure represents our data best, suggesting that the first structure is too restricted and in the third structure, there were too many parameters for which there is not sufficient information in our data to estimate all of them.

Several challenges in modelling the microbiome data using this method exist. Firstly, in our data application, we were able to fit a categorical-specific random effects structure however the computational burden was large. More research is needed to obtain computational efficient methods. The second challenge is

related to the number of categorical count outcome involved in the study. Typically, categories with rare count (bacteria only presence in the small number of samples) are pooled. One might argue that this rare count might be due to systematic error rather than sequencing error and thus pooling could be viewed as losing the information. Future research should address the issue of the number of categories included in the analyses and consequently a development of computationally efficient method is needed to take into account the category-dependent random effect.

Several alternatives for our approach can be considered. Although modelling overdispersion with the conjugate distribution has computational advantages, it might be too simple since all correlation is modeled by one additional parameter. Extensions to more complex correlation structures would be of interest. Secondly, the choice of six categories is arbitrary. More categories can be analyzed if the dimension of the parameter space is reduced, for example using penalization [Chen and Li (2013); Xia et al. (2013)]. Thirdly, the interpretation of the fixed effect parameters are all conditional on the random effects. In practice, one might be interested in marginal parameters [Heagerty (1999); Tsonaka et al. (2015)]. To this end, marginalized models for multivariate counts need to be developed. Finally, it is of interest to analyze the microbiome data jointly with other outcomes such as diseases or immunological markers. For example, we would like to model the effect of helminths and treatment on microbiome composition and cytokines. This is a topic of an ongoing research.

3.6 Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 Bias and MSE of datasets generated from DMM model with subject-specific random effect when parameters are obtained from the true dataset.

Figure S1 Bias and MSE of datasets generated from the DMM model with categorical-specific random effect having categorydependent variances.

Table S2 The sensitivity and specificity of the hypothesis testing for (A) covariate effect and (B) variability of random effect.

Table S3 The estimated marginal correlations based on the DMM model with subject-specific random effect across different standard deviation of random effects using Monte-Carlo.

Table S4 The estimated marginal correlations based on the DMM model with categorical-specific random effects with common variance across categories using Monte-Carlo.

Table S5 The estimated marginal correlations based on the UNBM model with subject-specific random effect across different standard deviation of random effects using Monte-Carlo.

Table S6 The estimated marginal correlation based on the UNBM model with categorical-specific random effect having common variance across categories using Monte-Carlo.

Table S7 The description of bacterial count data at each time-points.

Figure S2 The distribution of bacterial phyla when zero count presents.

Table S8 **The starting model**. The estimate (95% CI) of the log odds ratio for each covariates in the microbiome dataset.

Table S9 **Final Model**. The estimate of the log odds ratio (95% CI) for each covariates in the microbiome dataset.

Table S10 The log odds ratio (95% CI) when dataset were fitted with DMM with categorical-specific random effect having category-dependent variance across categories.

Table S11 The estimated marginal correlation of the dataset obtained by DMM models.

A Derivation of the joint multivariate distribution

A.1 Joint multivariate distribution for proportions

Conditioned on the random effect $\boldsymbol{u}_i$, the relative abundances are independent. Thus, the joint distribution for the multivariate relative abundance for subject i could be formulated as follows.

$$\begin{aligned}
\Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}}\right) &= \int_{u_i} \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}}, \frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}}, u_i\right) d\,u_i \\
&= \int_{u_i} \Pr\left(\frac{\boldsymbol{C}_i^{(1)}}{C_{i+}^{(1)}} | u_i\right) \Pr\left(\frac{\boldsymbol{C}_i^{(2)}}{C_{i+}^{(2)}} | u_i\right) \Pr(u_i)\, d\,u_i \\
&= \int_{u_i} \prod_{t=1}^{2} \Pr\left(\frac{\boldsymbol{C}_i^{(t)}}{C_{i+}^{(t)}}\right) \Pr(u_i)\, d\,u_i \\
&= \int_{u_i} \prod_{t=1}^{2} \frac{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)}\right) C_{i+}^{(t)}!}{\Gamma\left(\theta^{-1}\mu_{i+}^{(t)} + C_{i+}^{(t)}\right)} \prod_{j=1}^{J} \frac{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)} + C_{ij}^{(t)}\right)}{\Gamma\left(\theta^{-1}\mu_{ij}^{(t)}\right) C_{ij}^{(t)}!} \Pr(u_i)\, d\,u_i
\end{aligned} \tag{3.16}$$

with $\mu_{i\cdot}^{(t)}$ is the loglinear mean.

A.2 Joint multivariate distribution under condition on total count.

We will show that the distribution given in equation (3.9) and (3.10) are in general not equivalent. The distribution in the equation (3.10) and (3.12) are not equivalent except for the situation where the total count is fixed.

We denote the $\boldsymbol{C}_i^{(t)}$ as the multivariate count outcome at time t for subject i and the total count to be $C_{i+}^{(t)}$. Thus the multivariate count outcome for subject i conditional on their total is as follows.

$$\begin{aligned}
\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) &= \frac{\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)}\right)}{\Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}\right)} \\
&= \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i} \\
&= \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) d\,u_i}
\end{aligned} \tag{3.17}$$

The probability $\Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right)$ could be rewritten as follows.

$$\begin{aligned}
\Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) &= \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)} | u_i\right) \Pr(u_i) \\
\text{(by conditional independence)} &= \Pr\left(C_{i+}^{(1)} | u_i\right) \Pr\left(C_{i+}^{(2)} | u_i\right) \Pr(u_i) \\
\text{(conditional distribution)} &= \frac{\Pr\left(C_{i+}^{(1)}, u_i\right)}{\Pr(u_i)} \frac{\Pr\left(C_{i+}^{(2)}, u_i\right)}{\Pr(u_i)} \Pr(u_i) \\
&= \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right)}{\Pr(u_i)} \frac{\Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)} \Pr(u_i) \\
&= \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right) \cdot \Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)}
\end{aligned}$$

Thus, the joint probability of multivariate count outcome given in equation (3.17) can be written as follows.

$$\begin{aligned}
&\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) \\
&= \frac{\int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right) \cdot \Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)} d u_i}{\int_{u_i} \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(1)}\right) \cdot \Pr\left(u_i | C_{i+}^{(2)}\right) \Pr\left(C_{i+}^{(2)}\right)}{\Pr(u_i)}} \\
&= \int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) \left[\frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(u_i | C_{i+}^{(2)}\right)}{\Pr(u_i) \int_{u_i} \frac{\Pr\left(u_i | C_{i+}^{(1)}\right) \Pr\left(u_i | C_{i+}^{(2)}\right)}{\Pr(u_i)} d u_i} \right] d u_i
\end{aligned} \tag{3.18}$$

Since $\Pr(u_i)$ is not equal to the term in bracket in equation (3.18) then equation (3.9) and (3.10) are not equivalent. However, when the total count is fixed, the following equation holds: $\Pr\left(u_i | C_{i+}^{(t)}\right) = \Pr(u_i)$. Now, using the last equation in (3.18), the joint distribution becomes

$$\begin{aligned}
\Pr\left(\boldsymbol{C}_i^{(1)}, \boldsymbol{C}_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}\right) &= \int_{u_i} \Pr\left(\boldsymbol{C}_i^{(1)} | C_{i+}^{(1)}, u_i\right) \Pr\left(\boldsymbol{C}_i^{(2)} | C_{i+}^{(2)}, u_i\right) \Pr(u_i) \, d u_i \\
&= \int_{u_i} \frac{\Pr\left(\boldsymbol{C}_i^{(1)}, C_{i+}^{(1)} | u_i\right) \Pr\left(\boldsymbol{C}_i^{(2)}, C_{i+}^{(2)} | u_i\right)}{\Pr\left(C_{i+}^{(1)}\right) \Pr\left(C_{i+}^{(2)}\right)} \Pr(u_i) \, d u_i.
\end{aligned} \tag{3.19}$$

Since the count at each category $C_{ij}^{(t)}|u_i \sim \mathrm{NB}\left(\theta^{-1}\mu_{ij}^{(t)}, \frac{\theta^{-1}}{1+\theta^{-1}}\right)$ where $\log\left(\mu_{ij}^{(t)}\right)$ $= \boldsymbol{X}_i\boldsymbol{\xi}_j + u_i$, we obtain similar formulation as the conditional likelihood at the cross-sectional setting. Thus, in the case where the total count is fixed, the formulation is equivalent to the distribution of the multivariate relative abundance (3.16).