

RESEARCH ARTICLE

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

Ivonne Martin^{1,2}  | Hae-Won Uh³ | Taniawati Supali⁴ | Makedonka Mitreva^{5,6} | Jeanine J. Houwing-Duistermaat^{3,7}

¹Department of Mathematics, Parahyangan Catholic University, Bandung, Indonesia

²Biomedical Data Sciences, section Medical Statistics, Leiden University Medical Centre, Leiden, The Netherlands

³Department of Biostatistics and Research Support, UMC Utrecht, Utrecht, The Netherlands

⁴Department of Parasitology, Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia

⁵McDonnell Genome Institute, Washington University in St. Louis, St. Louis, Missouri

⁶Department of Medicine, Washington University in St. Louis, St. Louis, Missouri

⁷Department of Statistics, University of Leeds, Leeds, UK

Correspondence

Ivonne Martin, Department of Mathematics, Parahyangan Catholic University, Bandung, Indonesia; or Biomedical Data Sciences, section Medical Statistics, Leiden University Medical Centre, Leiden, The Netherlands.
Email: I.Martin@lumc.nl

Present Address

Ivonne Martin, Biomedical Data Sciences, LUMC

Funding information

Royal Netherlands Academy of Arts and Sciences (KNAW), Grant/Award Number:

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 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 data sets obtained by sequencing of the amplicon of the bacterial 16S rRNA gene. These data have a compositional structure and are typically overdispersed. The microbiome data set is from an epidemiological study carried out in a helminth-endemic area in Indonesia. The conclusions are as follows: 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.

KEYWORDS

conditional model, count, Dirichlet-multinomial, generalized linear mixed model, microbiome, multivariate, overdispersion

This is an open access article under the terms of the Creative Commons Attribution NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2019 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

57-SPIN3-JRP; MIMOmics
FP7-Health-F5-2012, Grant/Award
Number: 305280; Directorate General of
Resources for Science Technology and
Higher Education (DGRSTHE) of the
Republic of Indonesia

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.¹ 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 data set 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.² Chimeric sequences were filtered out and the resulting sequences are either categorized based on similarity into operational taxonomical units followed by annotation or directly annotated using relevant databases (eg, Ribosomal Database Project, Greengenes, or Silva). The counts for a specific category represent the abundances of the bacteria at a biological taxonomy level. Data sets generated through this sequencing process comprise features that have not been adequately accounted for by currently available statistical methods.³ Firstly, the data set might be represented by a matrix of taxonomical counts with a compositional structure, which imposes a correlation between taxa.⁴ 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 epidemiological study carried out in a helminth endemic rural area in Indonesia.⁵ 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.^{1,6,7} 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.⁸ In a previous paper,⁵ 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.⁹ 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.¹⁰ 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 modeled by including a normally distributed random effect in the linear predictor, ie, generalized linear mixed model. The overdispersion is typically accounted for by the conjugate distribution.¹⁰⁻¹² Molenberghs et al^{13,14} and Booth et al¹⁵ introduced a combined model, where the conjugate distribution for the overdispersion is used and the correlation over time is modeled by normally

distributed random effects, ie, 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-multinomial regression model with random effects to incorporate the correlation due to repeated measurements. We will use the reparameterization of the work of Guimarães and Lindrooth,¹² in which the overdispersion is a function of the covariates and the random effects.

This manuscript is organized as follows. In Section 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, simulation studies are described to investigate the performance of the proposed methods and the results of the analyses of the motivating data set are presented in Section 4. In Section 5, we conclude and discuss the proposed method.

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).^{16,17}

We first introduce the following notations. Let $C_i^{(t)} = \{C_{i1}^{(t)}, \dots, C_{iJ}^{(t)}\}$ be the J -dimensional vector of the multivariate microbial count with $C_{ij}^{(t)}$ the abundance of bacteria taxa j ($j = 1, \dots, J$) for subject i ($i = 1, \dots, N$) 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 modeling microbiome data as described earlier, either the sequence count itself can be considered or the normalized count related to the total sequence read, ie, 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 Section 2.1.

2.1 | Cross-sectional setting

2.1.1 | 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 j th level of predictor E and k th level of predictor F , which could be described in a $J \times K$ contingency table (Table 1) as follows.

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

$$\log(\mu_{jk}) = \lambda_0 + \lambda_j^E + \lambda_k^F + \lambda_{jk}^{EF}, \quad (1)$$

TABLE 1 The $J \times K$ contingency table

		F		
		k	...	K
j		1	...	K
E	1	c_{11}	...	c_{1K}
	2	c_{21}	...	c_{2K}
	\vdots	\vdots	\vdots	\vdots
	J	c_{J1}	...	c_{JK}
Marginal		c_{+1}	...	c_{+K}

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 j th level, the marginal mean of variable F at the k th 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 (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

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

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

Note that, in model (1), the response (bacterial categories) E and predictor F are exchangeable. The loglinear model (1) could be written in the regression format for the bacterial outcome as follows:

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

with $[.]$ as the indicator function. To show the equivalence between two models, note that 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 $\{\xi_{01}, \xi_{02}, \dots, \xi_{0J}\}$ with ξ_{0j} for $j = 2, \dots, J$ corresponds to $\lambda_0 + \lambda_j^E$. For subjects with their predictor in other categories k , the regression model $\{\xi_{01} + \xi_{11k}, \xi_{02} + \xi_{12k}, \dots, \xi_{0J} + \xi_{1Jk}\}$, where ξ_{1jk} for $j = 2, \dots, J$, 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 λ or of ξ can be obtained by maximizing the following likelihood function. Specifically, for subject i , it is given by

$$\begin{aligned} L_i(\lambda) &= \prod_j f_{\text{Pois}}(\lambda; c_{ijk}) \\ &= \prod_j \frac{\exp(-\mu_{jk}) \mu_{jk}^{c_{ijk}}}{c_{ijk}!}, \end{aligned} \quad (2)$$

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 previously, 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:

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

Here, the baseline constraint is applied on the first equation, whereas, for the second equation, this is not needed because there are only $J \times P$ parameters. This last equation represents the loglinear model written in terms of regression coefficients ξ and covariate values, where Treatment and Infection are binary variables. To assess the statistical significance of the p th covariate ($p = 1, \dots, P$) on the multivariate count distribution, the null hypothesis $\xi_p = \mathbf{0}$ should be tested. We will use the standard likelihood ratio test, which follows a χ^2 distribution with J degrees of freedom.

2.1.2 | 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 Lindrooth¹² 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 earlier 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

$$\begin{aligned} \Pr(\mathbf{c}_i = \{c_{1kl}, \dots, c_{Jkl}\} | c_{+kl}) &= \frac{\Pr(c_{1kl}, \dots, c_{Jkl}, c_{+kl})}{\Pr(c_{+kl})} \\ &= \frac{\prod_{j=1}^J f_{\text{Pois}}(c_{jkl}; \mu_{jkl})}{f_{\text{Pois}}(c_{+kl}; \mu_{+kl})} = c_{+kl}! \prod_{j=1}^J \left(\frac{1}{c_{jkl}!} \right) \left(\frac{\mu_{jkl}}{\mu_{+kl}} \right)^{c_{jkl}} \\ &\sim \text{Multinomial}(c_{+kl}; \pi_{1kl}, \dots, \pi_{Jkl}), \quad \text{where } \pi_{jkl} = \frac{\mu_{jkl}}{\mu_{+kl}}. \end{aligned} \quad (4)$$

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 $\pi_j = \frac{\mu_j}{\mu_+}$. This model is the multinomial logistic regression model. Note that the parameters λ of the loglinear model (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.

2.1.3 | 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 formulated the model for an overdispersed count outcome as follows:

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

Here, μ_{ij} corresponds to the mean of the count in the nonoverdispersed 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

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

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

$$\begin{aligned} L_i(\theta, \lambda) &= \prod_j f_{\text{NB}}(\lambda, \theta; C_{ij}) \\ &= \prod_j \frac{\Gamma(\theta^{-1} \mu_{ij} + C_{ij})}{C_{ij}! \Gamma(\theta^{-1} \mu_{ij})} \left(\frac{1}{\theta^{-1} + 1} \right)^{C_{ij}} \left(\frac{\theta^{-1}}{\theta^{-1} + 1} \right)^{\theta^{-1} \mu_{ij}}. \end{aligned} \quad (5)$$

Note that, in this setting, parameter θ , which models the overdispersion and the intercept λ_0 , is not identifiable. An often used solution is to absorb the overdispersion parameter into the grand mean λ_0 , ie, $\theta^{-1} \exp(\lambda_0) = \delta_0^{-1}$.¹²

2.1.4 | Overdispersed multinomial

We briefly review the overdispersed count data introduced by Guimarães and Lindrooth 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 in the work of Mossiman¹⁸ can be applied. This theorem states that, if $\mathbf{C}_i = \{C_{i1}, C_{i2}, \dots, C_{iJ}\}$ are independently Gamma distributed random variables with parameters $(\tilde{\mu}_{i1}, \tilde{\mu}_{i2}, \dots, \tilde{\mu}_{iJ})$ with the same scale parameter θ^{-1} , then the random variables $\mathbf{\Pi}_i = \{\Pi_{i1}, \Pi_{i2}, \dots, \Pi_{iJ}\}$ with $\Pi_{ij} = \frac{C_{ij}}{\sum_{j=1}^J C_{ij}}$ have a multivariate beta distribution (Dirichlet distribution) with parameters $\{\tilde{\mu}_{i1}, \tilde{\mu}_{i2}, \dots, \tilde{\mu}_{iJ}\}$. Note that the Dirichlet distribution is the conjugate for the multinomial distribution. Hence, the marginal distribution for the random variable $\mathbf{\Pi}_i$ is obtained by integrating out the Dirichlet random effects. Now, the corresponding Dirichlet-multinomial distribution is given by

$$\Pr(\mathbf{\Pi}_i) = \frac{\Gamma(\tilde{\mu}_{i+})C_{i+}!}{\Gamma(\tilde{\mu}_{i+} + C_{i+})} \prod_{j=1}^J \frac{\Gamma(\tilde{\mu}_{ij} + C_{ij})}{\Gamma(\tilde{\mu}_{ij})C_{ij}!}. \quad (6)$$

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

$$\begin{aligned} L_i(\lambda, \theta) = \Pr(\mathbf{C}_i | C_{i+}) &= \frac{\prod_{j=1}^J f_{\text{NB}}(C_{ij}; \tilde{\mu}_{ij})}{f_{\text{NB}}(C_{i+}; \tilde{\mu}_{i+})} \\ &= \frac{\Gamma(\theta^{-1}\mu_{i+})C_{i+}!}{\Gamma(\theta^{-1}\mu_{i+} + C_{i+})} \prod_{j=1}^J \frac{\Gamma(\theta^{-1}\mu_{ij} + C_{ij})}{\Gamma(\theta^{-1}\mu_{ij})C_{ij}!}. \end{aligned} \quad (7)$$

By $\tilde{\mu}_{ij} = \theta^{-1}\mu_{ij}$, it follows that the likelihood (7) is equivalent to the Dirichlet-multinomial distribution (6). Here, parameter θ is unidentifiable. Similar to (5), we apply the parameterization in the work of Guimarães and Lindrooth,¹² 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 nonoverdispersed multinomial model, the intercepts of the overdispersed multinomial model do not cancel out.

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, ie, 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^{13,14} and Booth et al.¹⁵ Molenberghs and Booth modeled the mean of an outcome as a multiplication of overdispersion and the linear predictor. However, to guarantee that the theorem 1 in the work of Mossimann holds, ie, 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; and 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 Section 2.1. Specifically, for the first model, we assume that, conditional on the random effects $\exp(\eta_{ij}^{(t)})$ and u_i , the count $C_{ij}^{(t)}$ follows a Poisson distribution with mean equal to

$$\begin{aligned} \mathbb{E} \left[C_{ij}^{(t)} \mid \exp(\eta_{ij}^{(t)}), u_i \right] &= \exp(\eta_{ij}^{(t)}) \exp(\tilde{\mu}_{ij}^{(t)}), \quad \text{where } \tilde{\mu}_{ij}^{(t)} = \mathbf{X}_i \boldsymbol{\xi}_j + u_i, \quad j = 1, \dots, J. \\ \exp(\eta_{ij}^{(t)}) &\sim \Gamma \left(\text{shape} = \theta^{-1} \exp(\tilde{\mu}_{ij}^{(t)}), \text{rate} = \theta^{-1} \exp(\mathbf{X}_i \boldsymbol{\xi}_j + u_i) \right). \end{aligned} \quad (8)$$

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:

$$\begin{aligned} L_{\text{UNBM}}(\xi, \theta, \sigma_u^2) &= \prod_i \Pr(C_i^{(t)}) = \prod_i \int_{u_i} \Pr(C_{i1}^{(t)}, \dots, C_{iJ}^{(t)}, u_i) du_i \\ &= \prod_i \int_{u_i} \prod_{t=1}^2 \prod_{j=1}^J \Pr(C_{ij}^{(t)} | u_i) \Pr(u_i) du_i, \end{aligned} \quad (9)$$

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

$$\begin{aligned} L_{\text{CNBM}}(\xi, \theta, \sigma_u^2) &= \prod_i \Pr(C_i^{(1)}, C_i^{(2)} | C_{i+}^{(1)}, C_{i+}^{(2)}) = \prod_i \frac{\int_{u_i} \Pr(C_i^{(1)}, C_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)} | u_i) \Pr(u_i) du_i}{\int_{u_i} \Pr(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i) du_i} \\ &= \prod_i \frac{\int_{u_i} \prod_{t=1}^2 \prod_{j=1}^J \Pr(C_{ij}^{(t)} | u_i) \Pr(u_i) du_i}{\int_{u_i} \prod_{t=1}^2 \Pr(C_{i+}^{(t)} | u_i) \Pr(u_i) du_i}. \end{aligned} \quad (10)$$

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 data set. 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, ie,

$$\begin{aligned} \left\{ \frac{C_{i1}^{(t)}}{C_{i+}^{(t)}}, \dots, \frac{C_{iJ}^{(t)}}{C_{i+}^{(t)}} \right\} \left| \left\{ \alpha_{i1}^{(t)}, \dots, \alpha_{iJ}^{(t)} \right\}, u_i \right. &\sim \text{Mult}(\tilde{\pi}_{i1}^{(t)}, \dots, \tilde{\pi}_{iJ}^{(t)}) \\ \left\{ \tilde{\pi}_{i1}^{(t)}, \dots, \tilde{\pi}_{iJ}^{(t)} \right\} &\sim \text{Dir}(\alpha_{i1}^{(t)}, \dots, \alpha_{iJ}^{(t)}) \\ \alpha_{ij}^{(t)} &= \theta^{-1} \mu_{ij}^{(t)}, \end{aligned} \quad (11)$$

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

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

The likelihood for each subject i is then formulated as follows:

$$\begin{aligned} L_{\text{DMM}}(\xi, \theta, \sigma_u^2) &= \Pr\left(\frac{C_i^{(1)}}{C_{i+}^{(1)}}, \frac{C_i^{(2)}}{C_{i+}^{(2)}}\right) = \int_{u_i} \Pr\left(\frac{C_i^{(1)}}{C_{i+}^{(1)}}, \frac{C_i^{(2)}}{C_{i+}^{(2)}} \middle| u_i\right) \Pr(u_i) du_i \\ &= \int_{u_i} \prod_{t=1}^2 \frac{\Gamma(\theta^{-1} \mu_{i+}^{(t)}) C_{i+}^{(t)}!}{\Gamma(\theta^{-1} \mu_{i+}^{(t)} + C_{i+}^{(t)})} \prod_{j=1}^J \frac{\Gamma(\theta^{-1} \mu_{ij}^{(t)} C_{ij}^{(t)})}{\Gamma(\theta^{-1} \mu_{ij}^{(t)}) C_{ij}^{(t)}!} \Pr(u_i) du_i. \end{aligned} \quad (12)$$

The corresponding regression model under this likelihood is denoted as the Dirichlet-multinomial mixed model (DMM). It is shown in Appendix A that, in the case where the total count does not depend on the random effect u_i , the likelihoods (10) and (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

$$\text{Corr}(C_{ij}^{(t)}, C_{ij}^{(t^*)}) = \frac{\sigma_{C_{ij}^{(t)}, C_{ij}^{(t^*)}}}{\sqrt{\sigma_{C_{ij}^{(t)}}^2 \cdot \sigma_{C_{ij}^{(t^*)}}^2}}.$$

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,¹⁹ and we used the functions available in the ecoreg package²⁰ to compute the integral. R implementations are available in github (<https://github.com/IvonneMartin/CombinedMultinomial>).

2.3 | The categorical-specific random effect

In the aforementioned 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 (8) now becomes

$$\begin{aligned} \tilde{\mu}_{ij}^{(t)} &= \mathbf{X}_i \boldsymbol{\xi}_j + u_{ij}, \quad j = 1, \dots, J. \\ u_{ij} &\sim \text{MVN}(\mathbf{0}, \Delta_{J \times J}). \end{aligned} \quad (13)$$

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 \mathbf{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 (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 | SIMULATION STUDY

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):

$$\begin{aligned} H_0 : \boldsymbol{\xi}_p &= \mathbf{0} \quad \text{vs} \quad H_1 : \text{at least one of } \boldsymbol{\xi}_p \neq 0, \\ H_0 : \sigma_u^2 &= 0 \quad \text{vs} \quad H_1 : \sigma_u^2 > 0. \end{aligned}$$

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 data set 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 data sets with three categories from the UNBM model.

3.1.1 | Data set generation

To reduce the computational burden, data sets 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, λ 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 data set (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 $i, i = 1, \dots, N$, we randomly generate binary covariates X_i^t for each time-point t and a random effect $u_i \sim N(0, \sigma_u^2)$.
2. The mean for each category j is computed as $\tilde{\mu}_{ij}^{(t)} = \theta^{-1} \exp(\lambda + u_i)$, where λ 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, $\sigma_u = (\sigma_{u_1}, \sigma_{u_2}, \sigma_{u_3})$ is (0.5,0.6,0.5), (0.8,0.9,0.8), or (1,0.9,1).

For the second set of simulation, six bacterial categories are used and parameters for the simulation are obtained from the data set. 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 $\tilde{\mu}_{ij} = \theta^{-1} \exp(\log(S) + \lambda + u_i)$. Here, the offset $\log(S)$ is incorporated to take into account the total bacteria count S . For each scenario mentioned earlier, 1000 replicates were generated. The models were fitted to each of the replicates.

3.2 | Simulation results

3.2.1 | Evaluation of DMM model

The performance of the method in estimating the parameters is described in Figure 1. Overall, the bias and MSE appear 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 a 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 data set 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 specificity when the sample size and variance component are large (Table S2B).

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, whereas the data sets were generated by using a vector of random effects following the multivariate normal distribution. The results are given in Table 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, whereas, 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 is larger. Moreover, these correlations tend to increase with a larger variance of the random effects.

3.2.2 | 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. Data sets 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

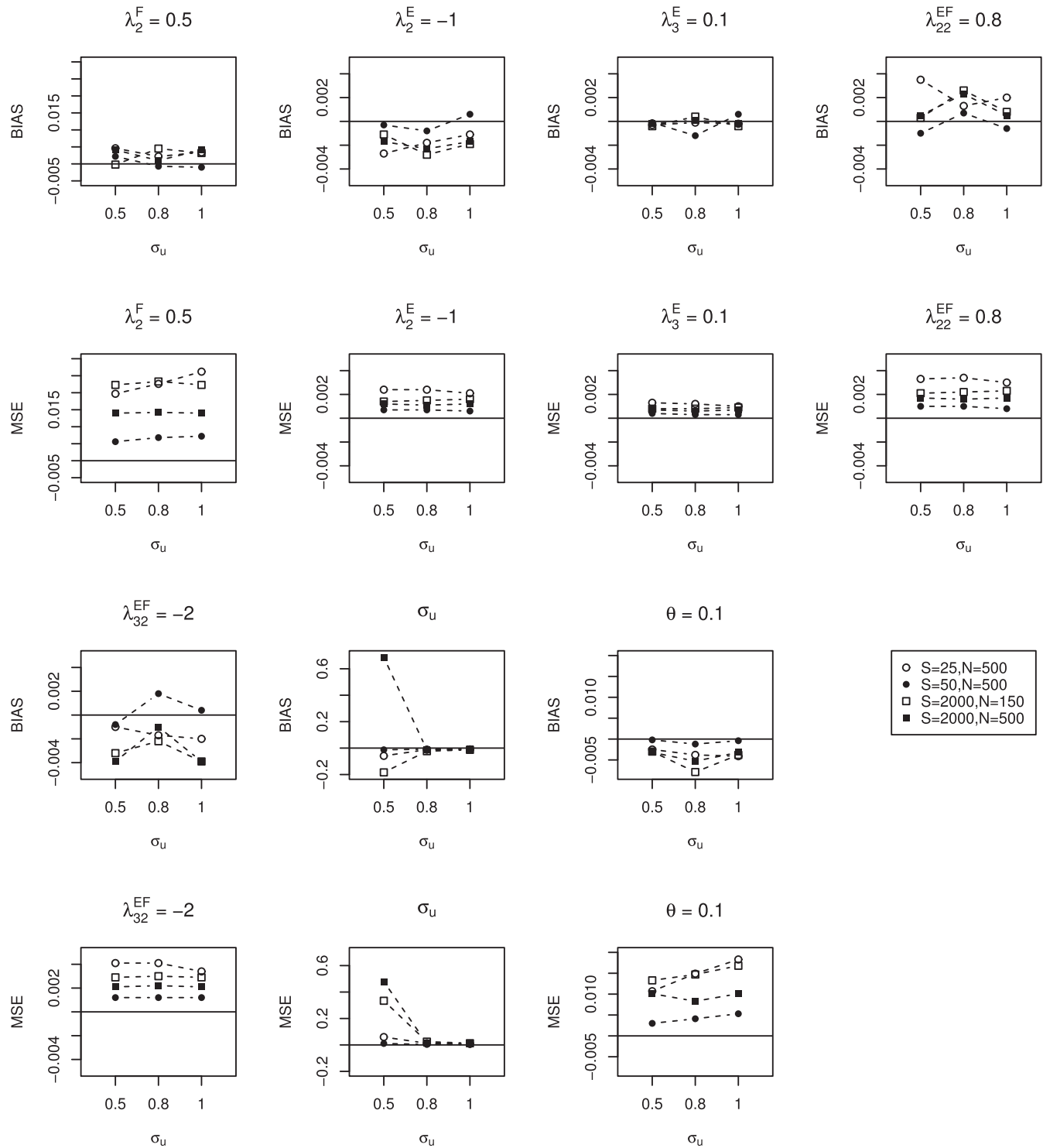


FIGURE 1 Bias and mean squared error (MSE) of data sets generated from the Dirichlet-multinomial mixed model with subject-specific random effect. λ : a vector of parameters in loglinear model. σ_u : the standard deviation of the between individual variation. θ the overdispersion

(CNBM), and the Dirichlet-multinomial mixed model (DMM) are given in Figure 2 for the fixed effect parameters and Figure 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 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 $N = 150$

TABLE 2 The mean estimates (standard deviation) over 1000 replicates when data sets were generated from the Dirichlet-multinomial mixed model with categorical-specific random effect with common variance

	$\lambda_2^F = 0.5$	$\lambda_2^E = -1$	$\lambda_3^E = 0.1$	$\lambda_{22}^{EF} = 0.8$	$\lambda_{32}^{EF} = -2$	$\log(\sigma_u) = -1.309$	$\log(\theta) = -2.302$	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)
	$\lambda_2^F = 0.5$	$\lambda_2^E = -1$	$\lambda_3^E = 0.1$	$\lambda_{22}^{EF} = 0.8$	$\lambda_{32}^{EF} = -2$	$\log(\sigma_u) = -0.693$	$\log(\theta) = -2.302$	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)
	$\lambda_2^F = 0.5$	$\lambda_2^E = -1$	$\lambda_3^E = 0.1$	$\lambda_{22}^{EF} = 0.8$	$\lambda_{32}^{EF} = -2$	$\log(\sigma_u) = -0.223$	$\log(\theta) = -2.302$	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)
	$\lambda_2^F = 0.5$	$\lambda_2^E = -1$	$\lambda_3^E = 0.1$	$\lambda_{22}^{EF} = 0.8$	$\lambda_{32}^{EF} = -2$	$\log(\sigma_u) = 0$	$\log(\theta) = -2.302$	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 data sets were fitted with the DMM model with subject-specific random effect and rows started with Cat-sp represents the estimates (standard deviation) when data sets were fitted with DMM model with categorical-specific random effect having common variance.

$\lambda, \sigma_u, \theta$ are as explained in Figure 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.

to $N = 500$). When using the conditional distribution, given the total, the estimates of the fixed effect parameters in Figure 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 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, improve 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 to be less biased, whereas the overdispersion parameters are underestimated. When fitting the DMM model to the data, the estimates of the random effect parameters are biased in all scenarios.

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 data set assuming (common) 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 data set considered here was measured in a subset of a randomized clinical trial performed in a helminth-endemic area in Nangapanda subdistrict, Indonesia, described elsewhere⁸ 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 infected 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 the work of Wiria et al.⁸ The 16s rRNA gene from the stool samples was 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 five most prevalent phyla and pooled the remaining into one category, resulting in six phyla categories: *Actinobacteria*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, unclassified, and pooled category.

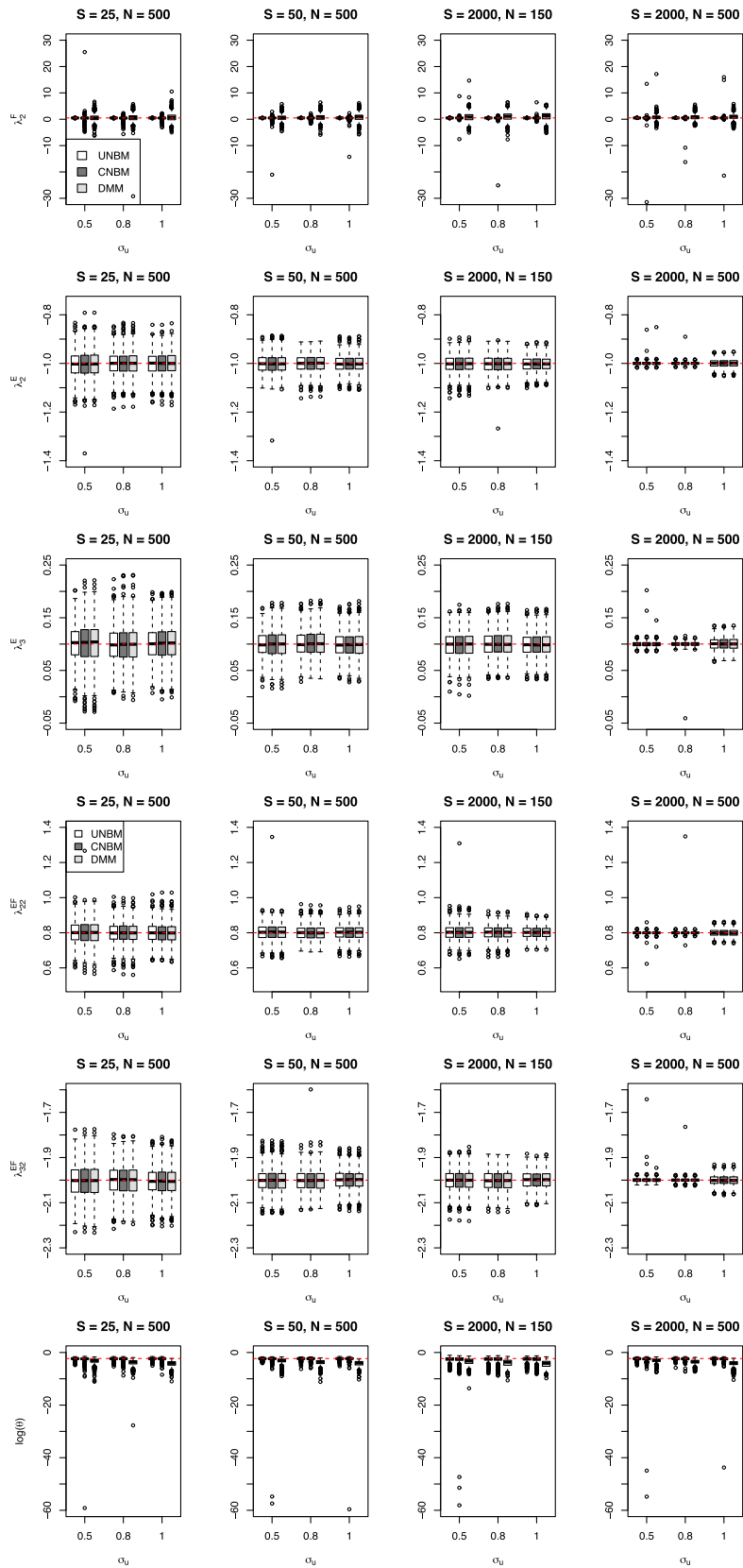


FIGURE 2 Estimates of the fixed effect and overdispersion parameters obtained from three different models (UNBM, CNBM, and DMM) when data sets were generated using UNBM model. CNBM, conditional negative-binomial mixed model; DMM, Dirichlet-multinomial mixed; UNBM, unconstrained negative binomial mixed [Colour figure can be viewed at wileyonlinelibrary.com]

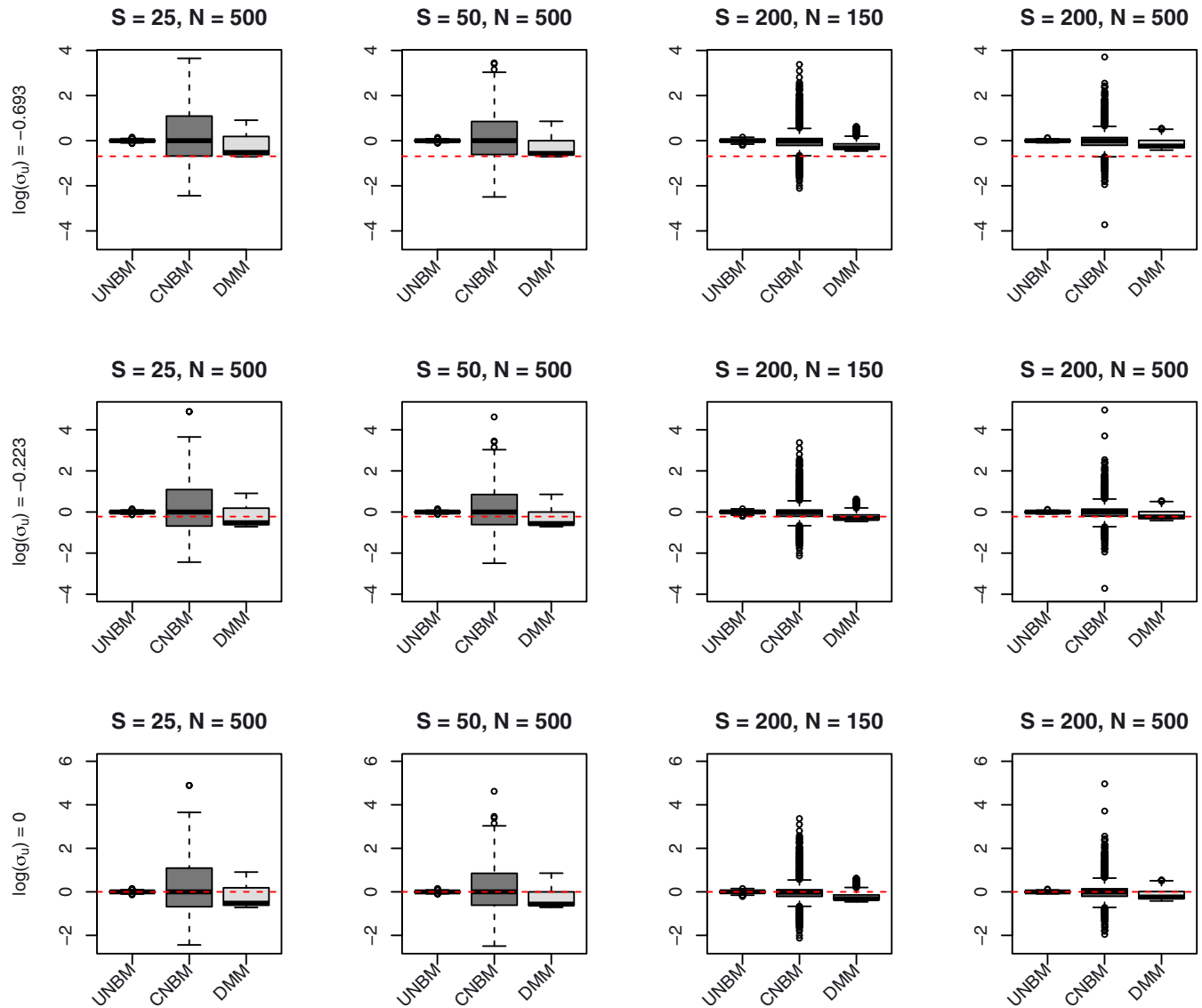


FIGURE 3 Estimates for the variance components obtained from three different models (UNBM, CNBM, and DMM) when data sets were generated using the UNBM model. CNBM, conditional negative-binomial mixed model; DMM, Dirichlet-multinomial mixed; UNBM, unconstrained negative binomial mixed [Colour figure can be viewed at wileyonlinelibrary.com]

The description of relative abundance of each bacterial phyla at each time points is 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 pretreatment 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 overdispersion by additional random effects.

Table 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*, and 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* corresponding to the decrease of these bacterial

TABLE 3 The observed marginal correlation of the motivating data set

	$C_1^{(1)}$	$C_2^{(1)}$	$C_3^{(1)}$	$C_4^{(1)}$	$C_5^{(1)}$	$C_6^{(1)}$	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$	$C_4^{(2)}$	$C_5^{(2)}$	$C_6^{(2)}$
$C_1^{(1)}$	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 $j, j = 1, \dots, 6$ at time-point t . The order of j are *Firmicutes*, *Actinobacteria*, *Bacteroidetes*, *Proteobacteria*, *Unclassified* and *pooled* category.

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 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 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 4. It has been shown previously that treatment had an effect on the composition at post-treatment in infected subjects who remained infected (condition 4).⁵ Here, we want to reanalyze this data set 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,

TABLE 4 Characteristics at baseline for study participants

Characteristics	albendazole (N = 69)	placebo (N = 81)
Age (in years), mean (sd)	28.38 (16.52)	27.85 (16.91)
Sex, female, n(%)	39 (56.52)	45 (55.56)
Helminth Infections, n(%)		
<i>A. lumbricoides</i>	17 (24.64)	18 (22.22)
Hookworm	26 (37.681)	23 (28.40)
<i>N. americanus</i>	25 (36.23)	23 (28.40)
<i>A. duodenale</i>	2 (2.90)	2 (2.47)
<i>T. trichiura</i>	20 (28.99)	22 (27.16)
Any helminths	47 (68.12)	47 (58.03)
Proportion (in %) of the 6 most abundant bacteria phyla, mean(sd)		
<i>Actinobacteria</i>	12.49 (8.95)	10.96 (7.98)
<i>Bacteroidetes</i>	7.41 (11.35)	6.42 (11.04)
<i>Firmicutes</i>	66.83 (13.46)	70.05 (13.66)
<i>Proteobacteria</i>	9.78 (7.86)	9.16 (8.37)
Unclassified [#]	1.95 (2.22)	2.68 (3.22)
Pooled [*]	1.54 (3.67)	0.73 (1.22)

[#]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%.

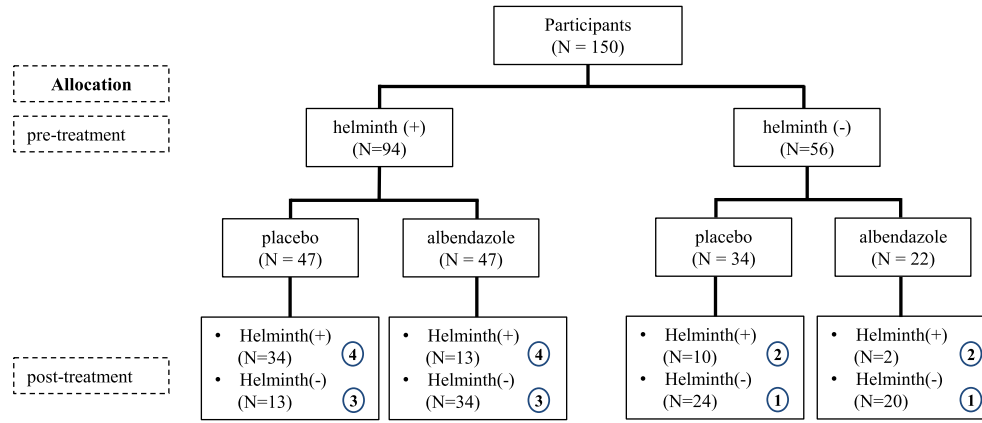


FIGURE 4 The profile of the microbiome study. The chart shows the number of subjects infected with at least one of the prevalent soil transmitted helminths (Helminth (+)) or free of helminth infections (Helminth (-)) that belonged to either the placebo or albendazole treatment group, at pre-treatment and 21 months after the first treatment round. The circled number represents the condition explained in Section 4 [Colour figure can be viewed at wileyonlinelibrary.com]

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

$$\log(\mu_{ijklm}^{DEFGH}) = (\log(\delta_0^{-1}) + \lambda_j^D) + (\lambda_k^E + \lambda_{jk}^{DE}) + (\lambda_n^H + \lambda_{jn}^{DH}) + (\lambda_{ln}^{FH} + \lambda_{ln}^{DFH}) + (\lambda_{mn}^{GH} + \lambda_{jmn}^{DGH}) + (\lambda_{lmn}^{FGH} + \lambda_{jlmn}^{DFGH}) + (\lambda_{klmn}^{EFGH} + \lambda_{jklmn}^{DEFGH}) + u_i$$

with the baseline constraint at $J = K = L = M = N = 1$, $u_i \sim N(0, \sigma_u^2)$. (14)

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

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

where $\xi_{0j} = \log(\delta_0^{-1}) + \lambda_j^D$, $\xi_{1j} = \lambda_l^F + \lambda_{jl}^{DF}$, 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 $\xi_{3j} + \xi_{5j} + \xi_{6j}$, whereas, 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 modeled by ξ_{2j} . Two interaction terms with BHelm were included in this model (14) (ie, 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 $\chi_{[0,1]}^2$ distribution.

The parameter estimates from the loglinear model with subject-specific random effects (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 $\chi_{[0,1]}^2$ distribution), indicating that the microbiome counts of a person over time are correlated. The regression coefficients for the covariates $\text{BHelm} \times t$ (ξ_{4j}) and $\text{BHelm} \times \text{TRT} \times t$ (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 conditions 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.

TABLE 5 The log odds ratio (95% CI) when data set were fitted with Dirichlet-multinomial mixed with categorical-specific random effect having common variance*

Categories	INF	t	TRT×t	Bhelm×INF×TRT×t
<i>Actinobacteria</i>	−0.006 (−0.218, 0.207)	0.050 (−0.155, 0.256)	0.046 (−0.235, 0.326)	0.326 (−0.042, 0.694)
<i>Bacteroidetes</i>	0.220 (−0.056, 0.496)	−0.119 (−0.395, 0.157)	−0.012 (−0.381, 0.356)	−0.916 (−1.573, −0.259)
<i>Proteobacteria</i>	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 (−0.814, 0.454)
Loglik	−8285.5	$\hat{\theta}$ (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.

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 model (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:

$$\log(\mu_{ij}^{(t)}) = \xi_{0j} + \xi_{1j}INF + \xi_{2j}t + \xi_{3j}TRT \times t + \xi_{4j}INF \times BHelm \times TRT \times t + u_i. \quad (15)$$

In this final model for fixed effects, assuming a subject-specific random effect (15), 6×4 parameters represent the covariate effects on the microbiome composition. The treatment effect is modeled 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. Moreover, 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%, whereas the odds ratio for *Bacteroidetes* to *Firmicutes* decreases about 62%.

Next, we considered a six-dimensional random effects structure for this data. We fitted DMM model (15). The results are listed in Tables 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. As the estimated variance component over time for the first three categories are relatively large ($\sigma_{u_1}^2 = 0.369$ to $\sigma_{u_3}^2 = 0.536$), 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 6) and with categorical-dependent variance for the random effects

TABLE 6 The estimated marginal correlation of the data set obtained by Dirichlet-multinomial mixed model with categorical-specific random effect having common variance across categories

	$C_1^{(1)}$	$C_2^{(1)}$	$C_3^{(1)}$	$C_4^{(1)}$	$C_5^{(1)}$	$C_6^{(1)}$	$C_1^{(2)}$	$C_2^{(2)}$	$C_3^{(2)}$	$C_4^{(2)}$	$C_5^{(2)}$	$C_6^{(2)}$
$C_1^{(1)}$	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 S11B). For all DMM models, the pairwise correlations at each time points between *Firmicutes* and the other three prevalent bacterial phyla are relatively high and similar to the observed marginal correlations (Table 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 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 data set (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.

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 for count data. Molenberghs et al reviewed the combined model for the binary²¹ and time-to-event data.²² 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.¹² We considered two 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, ie, total count is fixed, the likelihood is equivalent to the likelihood of the multinomial logistic model. Since, in our data set, 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 data sets 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, whereas, for the model with categorical-specific random effect, the correlations are larger and increased 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 modeling 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 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 data set 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 uninfected 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 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 modeled. For our data set, 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 no sufficient information in our data to estimate all of them.

Several challenges in modeling 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 modeling overdispersion with the conjugate distribution has computational advantages, it might be too simple because 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.^{10,23} Thirdly, the interpretation of the fixed effect parameters are all conditional on the random effects. In practice, one might be interested in marginal parameters.^{24,25} 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.

ACKNOWLEDGEMENTS

Authors would like to acknowledge people from the field team of Indonesia for sample collection and parasitological examination, and people from Mitreva lab for generating the microbiome data set. The study is funded by The Royal Netherlands Academy of Arts and Sciences (KNAW) 57-SPIN3-JRP and MIMomics FP7-Health-F5-2012 grant agreement 305280. The authors would like to thank the Directorate General of Resources for Science Technology and Higher Education (DGRSTHE) of the Republic of Indonesia for providing the scholarship for a PhD candidate in this study.

ORCID

Ivonne Martin  <https://orcid.org/0000-0003-0362-4479>

REFERENCES

1. Ramanan D, Bowcutt R, Lee SC, et al. Helminth infection promotes colonization resistance via type 2 immunity. *Science*. 2016;352(6285):608-612. <http://doi.org/10.1126/science.aaf3229>
2. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*. 2012;486(7402):215-221. <http://doi.org/10.1038/nature11209>
3. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu Rev Stat Its Appl*. 2015;2(1):73-94. <http://doi.org/10.1146/annurev-statistics-010814-020351>
4. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Front Microbiol*. 2017;2224:8. <http://doi.org/10.3389/fmicb.2017.02224>

5. Martin I, Djuardi Y, Sartono E, et al. Dynamic changes in human-gut microbiome in relation to a placebo-controlled anthelmintic trial in Indonesia. *PLoS Negl Trop Dis*. 2018;12(8):e0006620. <http://doi.org/10.1371/journal.pntd.0006620>
6. Cooper P, Walker AW, Reyes J, et al. Patent human infections with the whipworm, *Trichuris trichiura*, are not associated with alterations in the faecal microbiota. *PLoS One*. 2013;8(10):e76573. <http://doi.org/10.1371/journal.pone.0076573>
7. Lee SC, Tang MS, Lim YAL, et al. Helminth colonization is associated with increased diversity of the gut microbiota. *PLoS Negl Trop Dis*. 2014;8(5):e2880. <http://doi.org/10.1371/journal.pntd.0002880>
8. Wiria AE, Prasetyani MA, Hamid F, et al. Does treatment of intestinal helminth infections influence malaria? Background and methodology of a longitudinal study of clinical, parasitological and immunological parameters in Nangapanda, Flores, Indonesia (immunoSPIN Study). *BMC Infect Dis*. 2010;10:77. <http://doi.org/10.1186/1471-2334-10-77>
9. Xu L, Paterson AD, Turpin W, Xu W. Assessment and selection of competing models for zero-inflated microbiome data. *PLoS One*. 2015;10(7):e0129606. <http://doi.org/10.1371/journal.pone.0129606>
10. Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann Appl Stat*. 2013;7(1):418-442. <http://doi.org/10.1214/12-AOAS592>
11. Zhang X, Mallick H, Tang Z, et al. Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*. 2017;18(1):4. <http://doi.org/10.1186/s12859-016-1441-7>
12. Guimarães P, Lindrooth RC. Controlling for overdispersion in grouped conditional logit models: a computationally simple application of Dirichlet-multinomial regression. *Econ J*. 2007;10(2):439-452. <http://doi.org/10.1111/j.1368-423X.2007.00215.x>
13. Molenberghs G, Verbeke G, Demétrio CGB. An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Anal*. 2007;13(4):513-531. <http://doi.org/10.1007/s10985-007-9064-y>
14. Molenberghs G, Verbeke G, Demétrio CGB, Vieira AMC. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Stat Sci*. 2010;25(3):325-347. <http://doi.org/10.1214/10-STS328>
15. Booth JG, Casella G, Friedl H, Hobert JP. Negative binomial loglinear mixed models. *Stat Model*. 2003;3(3):179-191. <http://doi.org/10.1191/1471082X03st058oa>
16. Agresti A. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2013. *Wiley Series in Probability and Statistics*; vol. 792.
17. Tutz G. *Regression for Categorical Data*. New York, NY: Cambridge University Press; 2012. *Cambridge Series in Statistical and Probabilistic Mathematics*; vol. 34.
18. Mosimann JE. On the compound multinomial distribution, the multivariate β - distribution, and correlations among proportions. *Biometrika*. 1962;49(1/2):65-82. <http://doi.org/10.2307/2333468>
19. Liu Q, Pierce DA. A note on Gauss-Hermite quadrature. *Biometrika*. 1994;81(3):624-629. <http://doi.org/10.2307/2337136>
20. Jackson CH, Best NG, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *J R Stat Soc Ser A Stat Soc*. 2008;171(1):159-178.
21. Molenberghs G, Verbeke G, Iddi S, Demétrio CG. A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *J Multivar Anal*. 2012;111:94-109. <http://doi.org/10.1016/j.jmva.2012.05.005>
22. Efendi A, Molenberghs G, Iddi S. A marginalized combined gamma frailty and normal random-effects model for repeated, overdispersed, time-to-event outcomes. *Commun Stat - Theory Methods*. 2014;43(22):4806-4828. <http://doi.org/10.1080/03610926.2012.740125>
23. Xia F, Chen J, Fung WK, Li H. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*. 2013;69(4):1053-1063. <http://doi.org/10.1111/biom.12079>
24. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*. 1999;55(3):688-698. <http://doi.org/10.1111/j.0006-341X.1999.00688.x>
25. Tsonaka R, van der Woude D, Houwing-Duistermaat J. Marginal genetic effects estimation in family and twin studies using random-effects models. *Biometrics*. 2015;71(4):1130-1138. <http://doi.org/10.1111/biom.12350>

SUPPORTING INFORMATION

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

How to cite this article: Martin I, Uh H-W, Supali T, Mitreva M, Houwing-Duistermaat JJ. The mixed model for the analysis of a repeated-measurement multivariate count data. *Statistics in Medicine*. 2019;1-21. <https://doi.org/10.1002/sim.8101>

APPENDIX A

DERIVATION OF THE JOINT MULTIVARIATE DISTRIBUTION

A.1 | Joint multivariate distribution for proportions

Conditioned on the random effect u_i , the relative abundances are independent. Thus, the joint distribution for the multivariate relative abundance for subject i could be formulated as follows:

$$\begin{aligned}
 \Pr\left(\frac{C_i^{(1)}}{C_{i+}^{(1)}}, \frac{C_i^{(2)}}{C_{i+}^{(2)}}\right) &= \int_{u_i} \Pr\left(\frac{C_i^{(1)}}{C_{i+}^{(1)}}, \frac{C_i^{(2)}}{C_{i+}^{(2)}}, u_i\right) du_i \\
 &= \int_{u_i} \Pr\left(\frac{C_i^{(1)}}{C_{i+}^{(1)}} \middle| u_i\right) \Pr\left(\frac{C_i^{(2)}}{C_{i+}^{(2)}} \middle| u_i\right) \Pr(u_i) du_i \\
 &= \int_{u_i} \prod_{t=1}^2 \Pr\left(\frac{C_i^{(t)}}{C_{i+}^{(t)}}\right) \Pr(u_i) du_i \\
 &= \int_{u_i} \prod_{t=1}^2 \frac{\Gamma(\theta^{-1}\mu_{i+}^{(t)}) C_{i+}^{(t)!}}{\Gamma(\theta^{-1}\mu_{i+}^{(t)} + C_{i+}^{(t)})} \prod_{j=1}^J \frac{\Gamma(\theta^{-1}\mu_{ij}^{(t)} + C_{ij}^{(t)})}{\Gamma(\theta^{-1}\mu_{ij}^{(t)}) C_{ij}^{(t)!}} \Pr(u_i) du_i,
 \end{aligned} \tag{A1}$$

with $\mu_i^{(t)}$ is the loglinear mean.

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

We will show that the distribution given in Equations (9) and (10) is in general not equivalent. The distribution in Equations (10) and (12) is not equivalent except for the situation where the total count is fixed.

We denote $C_i^{(t)}$ 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:

$$\begin{aligned}
 \Pr\left(C_i^{(1)}, C_i^{(2)} \middle| C_{i+}^{(1)}, C_{i+}^{(2)}\right) &= \frac{\Pr\left(C_i^{(1)}, C_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)}\right)}{\Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}\right)} \\
 &= \frac{\int_{u_i} \Pr\left(C_i^{(1)}, C_i^{(2)}, C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) du_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) du_i} \\
 &= \frac{\int_{u_i} \Pr\left(C_i^{(1)}, C_i^{(2)} \middle| C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) du_i}{\int_{u_i} \Pr\left(C_{i+}^{(1)}, C_{i+}^{(2)}, u_i\right) du_i}.
 \end{aligned} \tag{A2}$$

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

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

Thus, the joint probability of multivariate count outcome given in Equation (A2) can be written as follows:

$$\begin{aligned} \Pr\left(\mathbf{C}_i^{(1)}, \mathbf{C}_i^{(2)} \mid \mathbf{C}_{i+}^{(1)}, \mathbf{C}_{i+}^{(2)}\right) &= \frac{\int_{u_i} \Pr\left(\mathbf{C}_i^{(1)}, \mathbf{C}_i^{(2)} \mid \mathbf{C}_{i+}^{(1)}, \mathbf{C}_{i+}^{(2)}, u_i\right) \frac{\Pr(u_i \mid \mathbf{C}_{i+}^{(1)}) \Pr(\mathbf{C}_{i+}^{(1)}) \cdot \Pr(u_i \mid \mathbf{C}_{i+}^{(2)}) \Pr(\mathbf{C}_{i+}^{(2)})}{\Pr(u_i)} du_i}{\int_{u_i} \frac{\Pr(u_i \mid \mathbf{C}_{i+}^{(1)}) \Pr(\mathbf{C}_{i+}^{(1)}) \cdot \Pr(u_i \mid \mathbf{C}_{i+}^{(2)}) \Pr(\mathbf{C}_{i+}^{(2)})}{\Pr(u_i)} du_i} \\ &= \int_{u_i} \Pr\left(\mathbf{C}_i^{(1)}, \mathbf{C}_i^{(2)} \mid \mathbf{C}_{i+}^{(1)}, \mathbf{C}_{i+}^{(2)}, u_i\right) \left[\frac{\Pr(u_i \mid \mathbf{C}_{i+}^{(1)}) \Pr(u_i \mid \mathbf{C}_{i+}^{(2)})}{\Pr(u_i) \int_{u_i} \frac{\Pr(u_i \mid \mathbf{C}_{i+}^{(1)}) \Pr(u_i \mid \mathbf{C}_{i+}^{(2)})}{\Pr(u_i)} du_i} \right] du_i. \end{aligned} \quad (\text{A3})$$

Since $\Pr(u_i)$ is not equal to the term in bracket in Equation (A3), then Equations (9) and (10) are not equivalent. However, when the total count is fixed, the following equation holds: $\Pr(u_i \mid \mathbf{C}_{i+}^{(t)}) = \Pr(u_i)$. Now, using the last equation in (A3), the joint distribution becomes

$$\begin{aligned} \Pr\left(\mathbf{C}_i^{(1)}, \mathbf{C}_i^{(2)} \mid \mathbf{C}_{i+}^{(1)}, \mathbf{C}_{i+}^{(2)}\right) &= \int_{u_i} \Pr\left(\mathbf{C}_i^{(1)} \mid \mathbf{C}_{i+}^{(1)}, u_i\right) \Pr\left(\mathbf{C}_i^{(2)} \mid \mathbf{C}_{i+}^{(2)}, u_i\right) \Pr(u_i) du_i \\ &= \int_{u_i} \frac{\Pr\left(\mathbf{C}_i^{(1)}, \mathbf{C}_{i+}^{(1)} \mid u_i\right) \Pr\left(\mathbf{C}_i^{(2)}, \mathbf{C}_{i+}^{(2)} \mid u_i\right)}{\Pr\left(\mathbf{C}_{i+}^{(1)}\right) \Pr\left(\mathbf{C}_{i+}^{(2)}\right)} \Pr(u_i) du_i. \end{aligned} \quad (\text{A4})$$

Since the count at each category $C_{ij}^{(t)} \mid u_i \sim \text{NB}\left(\theta^{-1} \mu_{ij}^{(t)}, \frac{\theta^{-1}}{1+\theta^{-1}}\right)$, where $\log\left(\mu_{ij}^{(t)}\right) = \mathbf{X}_i \boldsymbol{\xi}_j + u_i$, 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 (A1).