Universidad de Valparaíso CHILE

UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

NEW DEVELOPMENTS IN THE ESTIMATION OF STATISTICAL MODELS FOR COMPLEX LONGITUDINAL AND REPEATED DATA

Tesis presentada para optar al grado de Doctor en Matemática

Autor:
Jhon Erick Barrera Pérez

Profesor guía:

Cristian Meza
Universidad de Valparaíso

Examinadores:

Karine Bertin (Presidenta)
Universidad de Valparaíso

Luis Mauricio Castro (Informante)
Pontificia Universidad Católica de Chile

Marta Ávalos-Fernández
INRIA-Université de Bordeaux

Ana Arribas-Gil
Universidad Carlos III de Madrid

Héctor Araya
Universidad Adolfo Ibáñez

Valparaíso, Chile
Mayo, 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título; ☒ Tesis de Postgrado;

Título del trabajo: New Developments in the Estimation of Statistical Models for Complex Longitudinal and Repeated Data

Nombre del candidato(a): Jhon Erick Barrera Pérez

Carrera / Grado: Doctorado en Matemátca

Campus: Casa Central Valparaiso **;** **Departamento**: Matemática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Cristian Enrique Meza Becerra, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 2025/07/24 ; **Firma:** _____

Estudiante o Candidato(a):

Fecha: 2025/07/24 ; **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

ACKNOWLEDGEMENTS

First of all, I would like to thank my family, because even though distance keeps us apart, I can recognize them in everything I do, in my words, and in all the good that is or will one day be inside me.

I am also grateful to my thesis advisor, Dr. Cristian Meza, who with his patience, advice, and support has been an invaluable help throughout the journey over the past four years, as well as an example and guide of the professional I would like to become one day. To the professors and administrative staff at CIMFAV, especially Héctor Olivero and Rolando Rebolledo, for the great lessons I have learned working alongside them. A special thanks also goes to Marta Ávalos-Fernández for her hospitality during my stay in Bordeaux, as well as Ana Arribas-Gil, who contributed her experience and knowledge to much of the work presented in this thesis. I sincerely hope that we can continue to collaborate for many years to come.

Finally, to all the friends and colleagues I have had the pleasure of meeting over the past four years. Thank you for showing me that my true home is not a place, but the people with whom one feels at home.

The research presented in this thesis was partially funded by ANID Becas/Doctorado Nacional 21231659.

ABSTRACT

Advancements in measurement collection methods, in conjunction with the gradual automation of numerous processes, provide researchers with a substantial volume of data. To extract valuable information from this data, it is essential to model it according to its particular characteristics. In this context, the present thesis endeavors to develop novel methodologies and mechanisms for estimating complex statistical models based on longitudinal count and proportion data. These models are developed with a focus on situations of overdispersion, inflation in zeros, and temporal autocorrelations. Furthermore, the thesis demonstrates the application of these methods to the context of human microbiota analysis.

First, an analysis of the Zero Inflated Beta Regression (ZIBR) model, proposed for proportional compositional data (between zero and one), and the Zero Inflated Beta-Binomial Mixed Regression (ZIBBMR) model, for integer count data, were conducted. These mixed-effects models, which have been developed for the analysis of the abundance and presence of bacterial taxa in human microbiota samples, are distinguished by their hierarchical structure and are estimated by approximations of the observed likelihood. We hereby propose a novel estimation method based on a stochastic variation of the Expectation Maximization (EM) algorithm, as well as tools for statistical inference on the significance of covariates. The efficacy of this approach is evidenced by its superior performance in parameter estimation and the detection of the influence of clinical covariates, as demonstrated by experiments involving both simulated and actual data.

Secondly, we employed our methods to data from the COBRA-ENV and MODUL-CF experiments carried out in France, which record microbiota data from asthma and cystic fibrosis patients, respectively. The techniques developed enabled the identification of bacterial and fungal taxa exhibiting divergent behaviors among individuals with asthma and those without, as well as the construction of inference networks between species. Additionally, these techniques facilitated the identification of variables associated with the temporal evolution of clinical markers in patients diagnosed with cystic fibrosis. Finally, we explore the potential extensions of the methods to analogous models in the domain of microbiota, as well as the prospective avenues for advancement towards joint modeling of longitudinal and time-to-event data.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

FUNDAMENTALS: MICROBIOTA DATA, MIXED-EFFECTS MODELS AND THE SAEM ALGORITHM

We will commence this thesis with a discussion of the motivation behind it, which is the analysis of human microbiota data, as detailed in Section 1.1. The unique characteristics of these data pose significant statistical modeling challenges, requiring the development of techniques uniquely tailored to address these complexities. In Section 1.2, we will introduce mixed effects models, which are one of the main approaches for describing not only microbiota data but also many other longitudinal datasets that present grouping structures and, therefore, correlations between repeated measurements. Finally, Section 1.3 exposes the methodologies that will be employed to address the estimation and inference tasks of the models that will be proposed in this work. These methodologies have been demonstrated to be effective in other analogous contexts.

1.1 Human microbiota data

The human microbiota is the collection of microorganisms that inhabit different ecological niches of the human body, including the intestine, skin, oral cavity, respiratory tract, and urogenital tract. These microorganisms, mostly bacteria, perform essential functions for host homeostasis, such as nutrient digestion, immune system modulation, and protection against pathogens (Lloyd-Price et al., 2016). The characterization of microbiota has been made possible by advances in mass sequencing technologies, which allow direct analysis of microbial DNA present in a biological sample without the need for cultivation. There are two predominant approaches to metagenomic data collection: 16S rRNA gene sequencing and shotgun metagenomics, each with distinct methodological and statistical implications.

- **16S rRNA gene sequencing.** 16S ribosomal gene sequencing is a targeted technique that focuses on specific regions of the 16S gene, which is present in all bacteria and archaea. This gene contains conserved regions that allow its amplification

by PCR, as well as variable regions that enable taxonomic identification (Johnson et al., 2019). The general protocol involves the extraction of total DNA from the sample, the amplification of one or more regions of the 16S gene (e.g., V3-V4 or V4-V5), sequencing using platforms such as Illumina, and taxonomic assignment using databases such as SILVA or Greengenes. This technique has important advantages, such as lower cost and high scalability for population studies. However, it has significant limitations, including the inability to infer gene functions and biases derived from the choice of primers and amplified regions (Pollock et al., 2018).

- **Shotgun metagenomics.** In contrast to 16S sequencing, shotgun metagenomics does not target a specific gene, but rather randomly sequences all DNA present in the sample. This approach allows not only taxonomic identification at a finer level (down to species or strains), but also functional analysis of the microbiome, including metabolic pathways, antibiotic resistance genes, and mobile elements (Quince et al., 2017). The procedure consists of total DNA fragmentation, genomic library preparation, massive sequencing, and complex bioinformatic analysis that may include genome assembly, functional annotation, and quantification of taxonomic and functional abundances. This approach offers a comprehensive view of the microbial ecosystem, although it requires greater financial investment and substantial computational resources.

While we have outlined the various advantages and disadvantages of the two techniques, it should be noted that the 16S rRNA technique has limited taxonomic resolution compared to the shotgun technique. As illustrated in Figure 1.1, the former can only differentiate between bacteria at the genus level, while the latter can identify species or even strains.

1.1.1 Statistical analysis

After applying either of the two techniques mentioned above, the raw microbiota data consists of whole sequence counts assigned to different taxa per sample. These counts reflect the number of sequencing reads associated with each operational taxonomic unit (OTU), sequence variant (ASV), or species. Mathematically, this produces a matrix of discrete abundances, generally sparse and high-dimensional, in which the rows correspond to samples and the columns to taxa. One of the traditional approaches to analyzing this type of data is to treat it directly as counts, using statistical models suitable for discrete variables, such as those based on Poisson or negative binomial distributions. This maintains the integrity of the observed data, but requires observing heterogeneity in sequencing depth between samples, which must be corrected using normalization techniques such as rarefaction, library size normalization, or scaling factors (Anders and Huber, 2010).

Alternatively, and in many cases complementarily, microbiota data can be transformed into relative proportions by dividing the count of each taxon by the total number of reads observed in the sample. This results in continuous data in the range (0, 1) that also have the property of always summing to 1, which is more concisely referred to as compositional data (Aitchison, 1982). This approach enables the utilization of statistical methodologies defined for continuous variables, with the Beta distribution being the most prevalent for modeling purposes. Nevertheless, this treatment of proportions is subject

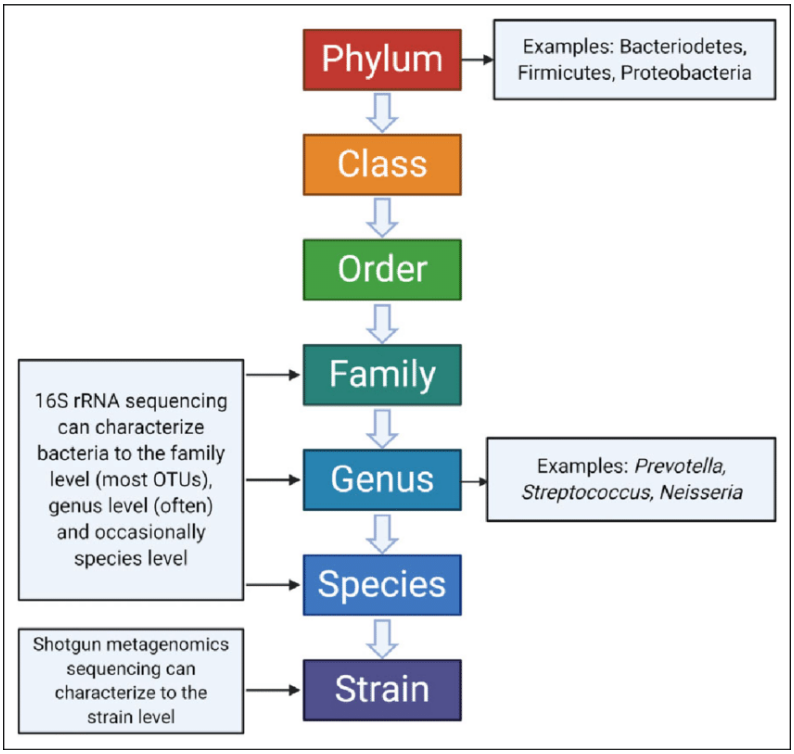


Figure 1.1: *Theoretical depth of the cited sequencing techniques for microbiota analysis. Image taken from Maki et al. (2021)*

to methodological limitations. Firstly, it is crucial to acknowledge the necessity to avoid the interpretation of these values as absolute abundances, due to the fact that the conversion of these values is contingent on the total number of sequences present in each sample. However, this quantity lacks biological significance, thereby complicating the interpretation of the absolute abundance values. Secondly, the constant sum imposed by normalization can induce spurious correlations between components, a phenomenon referred to as the *compositional trap* (Gloor et al., 2017).

In consideration of these limitations, it is possible to apply transformations to these data, provided that the modeling requires assumptions of normality or homoscedasticity. Among the most commonly employed are the *logit* transformation, which, if we consider the original proportions as p, is defined by

$$\operatorname{logit} p = \log\left(\frac{p}{1-p}\right).$$

This transformation is well known and applied in numerous contexts besides microbiota analysis. The *arcsine-square root* transformation, which is calculated by

$$\arcsin\left(\sqrt{p}\right)$$

and advocated by Sokal and Rohlf (1995) for data close to 0 or 1, is another option. However, Warton and Hui (2011) asserts that the logit transformation is more practical and interpretable in most cases. A final example of transformation, not as widely used as the previous ones, is the *complementary log-log* transformation, defined by

$$\operatorname{cloglog}(p) = \log\left(-\log\left(1-p\right)\right)$$

and which, according to Liu et al. (2017), is preferable for data with high asymmetry.

Regardless of the methodology employed in the analysis of the original data, the modeling of microbial abundance presents significant challenges. In particular, one of the more important features of the microbiota is that temporal variation occurs frequently because of the interaction between microorganisms and hosts (Gerber, 2015). In this context, longitudinal models are a more appropriate tool for analyzing the variables that affect such variations. As mentioned by Kodikara et al. (2022), further challenges are inherent to microbiota data, such as:

1. **Within-subject correlation**: A correlation structure at the level of each subject arises because data are collected from the same subject over time. Consequently, because samples from the same subject are not independent, statistical models capable of capturing such correlation structures must be considered.

2. **Sparsity**: Frequently, at some time points, the observed microbiota value is zero. This is because there is no physical presence of a particular microorganism, or the method for collecting data undersamples a particular microbiota population. Another reason for observing zero is the measurement error. Consequently, the observed data are called zero-inflated, and a model capable of handling excess zeros must be considered.

3. **Over-dispersion**: Because of external factors and random colonization of specific populations of microorganisms, laboratory procedures for determining the composition of the bacterial ecosystem, and the sparsity observed in the data, among other factors, microbiota counts are often heterogeneous, and therefore, the data are overdispersed.

4. **High-dimensional structure**: Each sample presents information for thousands of species, genera, and operational taxonomic units. Therefore, in most cases, the number of features (or variables) is higher than the number of observations. In this context, using models that deal with the *curse of dimensionality* is a must.

In the aforementioned article, the authors also list some existing models for the modeling tasks. Most of these will involve some kind of mixed effects structure, qualifying as mixed effects models, which, given their importance and use, we will describe in the next section.

1.2 Mixed effects models

Mixed effects models are a robust analytical instrument for the examination of longitudinal or clustered data, particularly when observations are collected repeatedly over time or across units that share common group structures. These models are designed to accommodate both within-subject and between-subject sources of variability, which are often intrinsic to biological and clinical data.

The underlying assumption of this framework is that the response variable is associated with a function of covariates, incorporating both fixed and random effects. The fixed effects represent population-level parameters and are assumed to be common across

all individuals in the study. These capture the average or systematic influence of covariates across the entire sample. In contrast, the random effects account for subject-specific deviations from this average trend, enabling the model to capture individual-level heterogeneity. In the case of repeated measures, random intercepts and/or slopes are introduced to represent variation in baseline levels or response trajectories across individuals.

This dual structure allows mixed models to disentangle two essential sources of variability: (1) within-subject variation, which reflects fluctuations in repeated observations over time for a given individual, and (2) between-subject variation, which reflects inherent differences among individuals. By modeling these components explicitly, mixed models provide more accurate estimates and inference than methods that ignore data dependence or heterogeneity.

The models considered in this thesis build upon the foundation of linear models, extending through increasingly general cases to arrive at generalized linear mixed models (GLMMs). These allow for flexible response distributions beyond the normal case (e.g., binomial or Poisson), while still preserving the hierarchical structure of random effects. It is precisely models of this kind—those that incorporate both the richness of individual variation and the structure of repeated observations—that will be studied and applied in this work.

1.2.1 Linear models

These models play an essential role in statistics, so much so that they are considered a fundamental pillar of this branch of mathematics (Vonesh and Chinchilli, 1996). Let us consider some observations $y_1, y_2, \ldots, y_n$ of a continuous variable coming from n different individuals (or time points) and a set of other variables $x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(d)}, \quad 1 \leq j \leq n$, that we assume can explain the first one. Linear models consider that the relation between these variables can be expressed by the next model:

$$y_j = \alpha_1 x_j^{(1)} + \alpha_2 x_j^{(2)} + \ldots + \alpha_d x_j^{(d)} + \epsilon_j, \tag{1.1}$$

where $\alpha_1, \ldots, \alpha_d$ are constants known as *regression coefficients* and ϵ_j is a sequence of error residuals, which are normally distributed with mean 0 and variance σ^2. In the case of longitudinal data, measurements are collected at observation times $t_1, t_2, \ldots, t_n$. In this context, y_j is the $j-$th observation measured at time t_j and $x_j^{(1)}, x_j^{(2)}, \ldots, x_j^{(d)}$ are the values of the d explanatory variables, also called *regression covariables* at time t_j.

Given the simplicity of its definition and the generality of its applications, the linear model is widely used to examine countless phenomena through the statistical lens. In particular, its estimation is extremely simple. To calculate estimators of the regression coefficients $\alpha_1, \ldots, \alpha_d$, it suffices to express the Model 1.1 as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = \left(x_j^{(k)}, 1 \leq j \leq n, 1 \leq k \leq d\right)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$. This expression is known as the *matrix form* of the model. Considering this form, the estimation can be performed using the *maximum likelihood (ML)* method or the *least squares (LS)* method. However, at least for the estimation of $\boldsymbol{\alpha}$, these two methods

coincide and, under very simple conditions known as *classical linear model assumptions* (Wooldridge, 2003), it can be shown that this estimator is given by:

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

1.2.2 Nonlinear models

Even with all their advantages, linear models may be insufficient for dealing with complex situations. In particular, in many statistical analysis contexts, the relationship between independent variables and the response cannot be adequately captured by a linear combination of parameters. In these cases, *nonlinear models* offer a more flexible framework, allowing for the modeling of complex functional relationships between predictors and the response variable (Bates and Watts, 1988). We can then extend Model 1.1, proposing nonlinear models as

$$y_j = f(\mathbf{x}_j, \boldsymbol{\theta}) + \varepsilon_j, \tag{1.3}$$

where y_j is the observed response for the $j-$th observation, x_j is the vector of predictor covariates, θ is a set of unknown parameters, $f(\cdot)$ is a nonlinear function with respect to the parameters θ, and ϵ_j represents the random error, normally distributed with mean 0 and variance σ^2.

This type of model is especially useful when theoretical or empirical knowledge about the functional form of the relationship is available (e.g., logarithmic, exponential, sigmoid), as is often the case in microbial population dynamics, enzyme kinetics, or treatment response. In some cases, these functional forms come from solving systems of ordinary differential equations.

Obviously, estimation for nonlinear models is not as straightforward as for linear models. In particular, if we consider ML estimation, the usual approach is to optimize the likelihood function using numerical techniques, such as those typically performed by algorithms such as Newton-Raphson, quasi-Newton, or the Levenberg–Marquardt algorithm (Seber and Wild, 2005).

1.2.3 Linear mixed effects models

A further limitation of Model 1.1 is its assumption that all observed values are generated from a single source, whether it be the same individual, the same process, or the same experiment. However, it is important to note that certain phenomena may originate from different observation subunits. Consequently, measurements from the same subunit are expected to be more similar to each other than to those from other subunits. This process ultimately results in the formation of a grouped data structure.

Suppose now that we perform a study on N individuals and seek to construct a general model for all observations collected for the N individuals. We will denote y_{it} the observation taken of individual i at time t and $x_{it}^{(1)}, x_{it}^{(2)}, \ldots, x_{it}^{(d)}$ the values of the d explanatory variables for individual i at time t. If we further assume that the parameters of the model vary from one individual to another, then for any individual i, Model 1.1 becomes

$$y_{it} = a_{i1}x_{it}^{(1)} + a_{i2}x_{it}^{(2)} + \ldots + a_{id}x_{it}^{(d)} + \epsilon_{it}, \quad 1 \leq i \leq N, \quad 1 \leq t \leq T_i. \tag{1.4}$$

Let us assume that each individual parameter a_{ik} can be additivity broken down into a fixed effect β_k and an individual effect β_{ik}, i.e;

$$a_{ik} = \beta_k + \beta_{ik}, \quad 1 \leq k \leq d$$

where, β_k is the value of $k-$th model parameter in the population, and β_{ik} the deviation of a_{ik} from this value. We can write (1.4) in matrix form:

$$\boldsymbol{y_i} = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{X}_i\boldsymbol{\beta_i} + \boldsymbol{\epsilon_i}, \tag{1.5}$$

where $\boldsymbol{X}_i$ is the $n_i \times d$ design matrix made up of d explanatory variables $x^{(1)}, x^{(2)}, \ldots, x^{(d)}$, $\boldsymbol{y_i}$ and $\boldsymbol{\epsilon_i}$ the $n_i \times 1$ vectors of observations and residual errors respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{\beta_i}$ the $d \times 1$ vectors of fixed and individual effects.

In the context of a probabilistic framework, the individual parameters a_{ik} are regarded as random variables and the part dependent on the individual β_{ik} are designated as *random effects*. In the context of a linear mixed effects (LME) model, the parameters $\boldsymbol{\beta}_i$ and $\boldsymbol{\epsilon}_i$ are defined as independent vectors normally distributed with each component possessing a mean of zero and respective variances $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_i$. Adding a further general component, we can consider different design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ for the two parts of a_{ik} (Laird and Ware, 1982):

$$\boldsymbol{y_i} = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i. \tag{1.6}$$

This development enables the utilization of distinct covariates for the two components of the coefficients, thereby distinguishing between those that influence interindividual variability and those that affect intraindividual variability. It is possible to refer to an equivalent way of writing Model 1.6, which employs a hierarchical structure that will be very useful throughout this thesis:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{\beta}_i &\sim N\left(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i\right) \\ \boldsymbol{\beta}_i &\sim N\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right) \end{aligned} \tag{1.7}$$

From this, it can be concluded that $\boldsymbol{y_i}$ is also a Gaussian vector:

$$\boldsymbol{y}_i \sim N\left(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{X}_i\boldsymbol{\Sigma}\boldsymbol{X}_i' + \boldsymbol{\Sigma}_i\right).$$

1.2.4 Nonlinear mixed effects models

As in Sections 1.2.1 and 1.2.2, the model (1.7) can be generalized to accept nonlinear structures, thus defining *nonlinear mixed effects* models (NLME) (Lindstrom and Bates, 1990). These structures can be presented for both the mean and the variability of the observations, using the following scheme:

$$\begin{aligned} y_{it} &= f\left(\boldsymbol{x}_{it}; \boldsymbol{\phi}_{it}\right) + g\left(\boldsymbol{z}_{it}; \boldsymbol{\phi}_{it}, \xi\right)\epsilon_{it}, \quad 1 \leq i \leq N, \quad 1 \leq t \leq T_i \\ \boldsymbol{\phi}_{it} &= \boldsymbol{X}_{it}\boldsymbol{\beta} + \boldsymbol{W}_{it}\boldsymbol{\beta}_i, \quad \boldsymbol{\beta}_i \sim N\left(0, \Sigma\right), \end{aligned} \tag{1.8}$$

where the structural model $f(\cdot)$ and variability model $g(\cdot)$ are known nonlinear functions, $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ are covariates for the observations, $\boldsymbol{X}_{it}$ and $\boldsymbol{W}_{it}$ are design matrices for the random effects, ξ are variability parameters, and $\boldsymbol{\phi}$ is the linear combination of fixed ($\boldsymbol{\beta}$)

and random ($\boldsymbol{\beta}_i$) effects. Note that equation (1.8) reduces to a LME model when a linear form of $f(\cdot)$ and $g(\cdot)$ is chosen.

Assuming furthermore that ϵ_{ij} is the measurement error independent of $\boldsymbol{\beta}_i$ and normally distributed with mean 0 and variance σ^2, the model can be written hierarchically as follows:

$$\begin{aligned} y_{it}|\boldsymbol{\phi}_{it}, \xi &\sim N\left(f\left(\boldsymbol{x}_{it}; \boldsymbol{\phi}_{it}\right), \sigma^2\left(g\left(\boldsymbol{z}_{it}; \boldsymbol{\phi}_{it}, \xi\right)^2\right)\right) \\ \boldsymbol{\phi}_{it}|\boldsymbol{\beta}_i, \boldsymbol{\beta} &= \boldsymbol{X}_{it}\boldsymbol{\beta} + \boldsymbol{W}_{it}\boldsymbol{\beta}_i \\ \boldsymbol{\beta}_i &\sim N\left(\mathbf{0}, \Sigma\right) \end{aligned} \tag{1.9}$$

1.2.5 Generalized linear mixed effects models

Up to this point, the emphasis has been placed on models for which the response variable, whether it be linear or nonlinear, assumes continuous values. However, this is not always the case. In instances where observations are discrete, binary, or continuous but with values in a closed interval, a more general modeling framework is required. In light of the contributions from Nelder and Wedderburn (1972) and McCullagh and Nelder (1982), the utilization of *Generalized Linear Models* (GLM) emerged as a proposed solution to address these circumstances.

To define these models, we must first discuss the *exponential family* of distributions. The exponential family includes many distribution such as Normal, Poisson, Binomial or Gamma. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ a vector of observations from a distribution in the exponential family. Its density can be expressed in the form:

$$p(\boldsymbol{y}|\boldsymbol{\eta}, \psi) = \exp\left(\frac{\boldsymbol{y}\boldsymbol{\eta} - b(\boldsymbol{\eta})}{a(\psi)} + c(\boldsymbol{y}, \psi)\right), \tag{1.10}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are specific functions, $\boldsymbol{\eta}$ is called the canonical parameter representing location, and ψ is called the dispersion parameter representing scale. It can be shown that

$$\mathbb{E}\left[\boldsymbol{y}\right] = \frac{\partial b(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}, \quad \mathbb{V}ar\left[\boldsymbol{y}\right] = a(\psi) \cdot \frac{\partial^2 b(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^2}.$$

The basic structure of a GLM is:

$$\boldsymbol{\eta} = g(\boldsymbol{m}), \quad \text{and} \quad \boldsymbol{m} = \mathbb{E}\left[\boldsymbol{y}\right] = g^{-1}(\boldsymbol{\eta}), \tag{1.11}$$

where $\boldsymbol{\eta}$ is named as the linear predictor, and g is a monotonic differentiable function which relates the mean with the linear predictor, and therefore called *link function*. In addition, a GLM requires the choice of a distribution (within the exponential family). In the Gaussian case, we have that $\boldsymbol{\eta} = \boldsymbol{m}$ and $\psi = \sigma^2$, and thus, the link function is the identity. There are many choices of link functions and usually the canonical link is selected (for more details see Verbeke et al. (2010) and McCulloch and Searle (2004)).

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT_i})'$ be the T_i repeated observations of the response within individual i. Along the same line of introducing random effects, we assume that conditioning on these effects $\boldsymbol{\beta}_i$, the repeated measures $\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{iN}$ are independent and

each follows a distribution in the exponential family. Then, a *Generalized Linear Mixed Effects Model (GLMM)* can be written as:

$$\begin{aligned} g\left(\boldsymbol{m_i}\right) &= h\left(\boldsymbol{x}_{ij}, \boldsymbol{\phi}_i\right) \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_i \\ \boldsymbol{\phi}_i &= \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{W}_{ij}\boldsymbol{\beta}_i, \quad \boldsymbol{\beta}_i \sim \mathcal{N}\left(0, \Sigma\right), \end{aligned} \tag{1.12}$$

where $\boldsymbol{m}_{ij} = \mathbb{E}\left[m_{ij}|\boldsymbol{\beta}, \boldsymbol{\beta_i}\right]$ is the conditional mean, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_i$ are fixed effects parameters and random effects respectively, and $\boldsymbol{\Sigma}$ is a covariance matrix.

If a shift to a more general framework is desired, it is important to acknowledge the development of the Generalized Additive Models for Location, Scale and Shape (GAMLSS) by Rigby and Stasinopoulos (2005). This development enables the modeling of not only the mean components, but also of dispersion, asymmetry, and kurtosis, through link functions and random effects. In the course of this thesis, the proposed framework will be utilized as a basis for comparison with the techniques developed in subsequent chapters.

1.2.6 Zero-inflated and zero-adjusted models

Zero-inflated models (ZIMs) are statistical models designed for count or continuous data with an excess of zeros, meaning that the observed frequency of zeros exceeds what is expected under standard distributions like Poisson, negative binomial, or Beta. This phenomenon is common when two distinct data-generating mechanisms are present (Min and Agresti, 2005). Depending on whether or not the original distribution allowed for a value of zero, we are talking about a zero-inflated or zero-adjusted model, respectively.

Let us assume that y_{it} follows a distribution given by $f(y_{it}; x_{it}, \varphi_i, \theta)$ the density (or mass) function, with φ_i being the random effects, x_{it} covariates and θ being fixed parameters. As mentioned earlier, if we assume that y_{it} can take zero values (as in the case of a Poisson model), the zero-inflated model is defined by a new distribution given by $\hat{f}(y_{it}; x_{it}, \varphi_i, \theta, p_{it})$:

$$\hat{f}(y_{it}; x_{it}, \varphi_i, \theta, p_{it}) = \begin{cases} p_{it} + (1 - p_{it}) \cdot f(0; x_{it}, \varphi_i, \theta) & \text{if } y_{it} = 0 \\ (1 - p_{it}) \cdot f(y_{it}; x_{it}, \varphi_i, \theta) & \text{if } y_{it} > 0, \end{cases}$$

or in the case where zero values cannot be taken (the Beta distribution, for example):

$$\hat{f}(y_{it}; x_{it}, \varphi_i, \theta, p_{it}) = \begin{cases} p_{it} & \text{if } y_{it} = 0 \\ (1 - p_{it}) \cdot f(y_{it}; x_{it}, \varphi_i, \theta) & \text{if } y_{it} > 0. \end{cases}$$

The $p_{it} \in (0, 1)$ value quantifies the probability of structural zeros for the model, and thus it may be beneficial to introduce a dependence of this value on covariates or random effects. Most often, this is done using a link function, which can be the logit or probit function. However, this adds another layer of complexity to model estimation tasks, so it is advisable to check whether this dependency is well justified. More details on the estimation of these models can be found in Ospina and Ferrari (2012).

1.2.7 Inference methods for mixed effects models

Subsequent to establishing the framework within which mixed models are proposed, the discussion will proceed to the statistical inference tasks that can be performed on them.

Assuming the vector of observations for each individual is represented by $\boldsymbol{y}_i$, the random effects are denoted by $\boldsymbol{\beta}_i$, and the fixed parameters of the model are indicated by $\boldsymbol{\theta}$, which include, among others, the fixed effects $\boldsymbol{\beta}$ and the covariance matrix of the random effects $\boldsymbol{\Sigma}$, the log-likelihood function can be established as follows:

$$\mathcal{LL}\left(\boldsymbol{\theta};\boldsymbol{y}\right) = \prod_{i=1}^{N} \int p_1(\boldsymbol{y}_i|\boldsymbol{\beta}_i;\boldsymbol{\theta}) p_2(\boldsymbol{\beta}_i|\boldsymbol{\theta}) d\boldsymbol{\beta}_i, \tag{1.13}$$

where p_1 and p_2 are the conditional densities for observations and random effects respectively, $\boldsymbol{y} = (y_1, \ldots, y_N)'$, $\boldsymbol{\theta}$ is the collection of all parameters, and $\boldsymbol{\beta}_i$ are random effects.

In the context of maximum likelihood estimation, the subsequent logical step entails the identification of values that optimize the function expressed in Equation 1.13. However, with the exception of Model 1.4, the expressions of these optimal estimators are not immediately apparent. This is primarily due to the fact that the integrals involved in calculating the log-likelihood are analytically impossible to handle. Consequently, numerous alternative approaches have been proposed to streamline these operations. Some of these methods involve linearizing the model, a procedure implemented in the `lme` package (Pinheiro and Bates, 2006) for the software `R`, or approximating integrals using numerical methods, a technique that has been used, among other places, in the `nlme4` package (Bates et al., 2015) for the same previous software. Iterative techniques have also been proposed, which in most cases are modifications of the *algorithm Expectation-Maximization (EM)* proposed by Dempster et al. (1977) to perform maximum likelihood estimation in censored or incomplete data problems.

However, linearization-based approaches have both statistical and practical drawbacks. First, they do not converge to the maximum likelihood estimates. While bias is typically negligible for fixed effects, bias in variance components can be significant, mainly when there is high interindividual variability. This has been shown to increase the type I error of likelihood tests (Comets and Mentré, 2001; Bertrand et al., 2008) with the potential of building wrong models. Second, stochastic algorithms have been shown to yield unbiased estimates (Savic et al., 2011), while linearization-based methods exhibit substantial bias when applied to non-continuous data, as demonstrated by Molenberghs and Verbeke (2006).

In the next section, we will explain in detail the algorithm we will use for estimation and inference tasks throughout this work, defined as a stochastic version of the EM algorithm, the so-called SAEM algorithm (Delyon et al., 1999).

1.3 The SAEM algorithm for estimation and inference

The Expectation-Maximization (EM) algorithm (Dempster et al., 1977) is an iterative method widely used to estimate parameters in statistical models with incomplete data, censored data, or latent structures. When the observed data $\boldsymbol{y}$ are incomplete or involve unobserved latent variables $\boldsymbol{z}$, the EM algorithm maximizes the marginal likelihood $L(\boldsymbol{\theta};\boldsymbol{y})$ by iteratively estimating the complete-data log-likelihood and optimizing it. In this case, the EM algorithm iterates the following two steps until convergence:

1. **E-step (Expectation).** Compute the expected value of the complete-data log-likelihood, given the observed data and current parameter estimates:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\theta}^{(k)}} \left[\log p(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\theta})\right]$$

2. **M-step (Maximization).** Maximize this expected log-likelihood with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$$

This scheme ensures some interesting properties, such as the fact that the log-likelihood does not decrease at each iteration. However, the performance depends on the ability to compute or approximate the conditional expectation in the E-step. In these cases, there are more robust variants of the algorithm that are proposed to address these limitations.

The Stochastic Approximation Expectation-Maximization (SAEM) algorithm (Kuhn and Lavielle, 2004, 2005) is a powerful tool for estimating population parameters in complex mixed effect models. This algorithm is applicable for the iterative computation of ML estimates in a wide variety of incomplete data statistical problems in which the Expectation step of the EM algorithm is not explicit; in particular in mixed effects models, where the individual random effects are treated as non-observed data.

Let $\boldsymbol{y} = (y_{it}, 1 \leq i \leq N, 1 \leq t \leq T_i)$ and $\boldsymbol{\varphi} = (\varphi_i, 1 \leq i \leq N)$ denote the observed and non-observed data, respectively, so the complete data of the model are $(\boldsymbol{y}, \boldsymbol{\varphi})$. In this case, the SAEM algorithm consists of replacing the usual E-step of EM with a stochastic approximation procedure with the aim of finding an estimator of $\boldsymbol{\theta}$. Given an initial point $\boldsymbol{\theta}^{(0)}$, iteration q of the algorithm is as follows:

- **Simulation (S) step:** Draw a realization $\boldsymbol{\varphi}^{(q)}$ from the conditional distribution $p\left(\cdot| \boldsymbol{y}; \boldsymbol{\theta}^{(q-1)}\right)$.
- **Stochastic Approximation (SA) step:** Update $s_q(\boldsymbol{\theta})$, the approximation of the conditional expectation $\mathbb{E}\left[\log p\left(\boldsymbol{y}, \boldsymbol{\varphi}^{(q)}; \boldsymbol{\theta}\right) |\boldsymbol{y}, \boldsymbol{\theta}^{(q-1)}\right]$:

$$s_q(\boldsymbol{\theta}) = s_{q-1}(\boldsymbol{\theta}) + \gamma_q \left(\log p\left(\boldsymbol{y}, \boldsymbol{\varphi}^{(q)}; \boldsymbol{\theta}\right) - s_{q-1}(\boldsymbol{\theta})\right)$$

  where $\{\gamma_q\}_{q\in\mathbb{N}}$ is a decreasing sequence of stepsizes with $\gamma_1 = 1$.
- **Maximization (M) step:** Update $\boldsymbol{\theta}^{(q)}$ according to $\boldsymbol{\theta}^{(q)} = \arg\max_{\boldsymbol{\theta}} s_q(\boldsymbol{\theta})$.

There are some important remarks on the the working details of the SAEM algorithm. In the case of complex mixed effects models the conditional distribution of the non-observed data $p\left(\cdot| \boldsymbol{y}; \boldsymbol{\theta}^{(q-1)}\right)$ cannot be computed in closed form and simulation from it cannot be carried out directly. However, a Markov Chain Monte Carlo (MCMC) approach can be used in the Simulation step of the SAEM algorithm described above, consisting in applying, for instance, the Metropolis-Hastings algorithm (Metropolis et al., 1953) with different proposal kernels, in order to approximate $p\left(\cdot| \boldsymbol{y}; \boldsymbol{\theta}^{(q-1)}\right)$ with a Markov chain with defined transition probabilities.

Also, convergence can be improved by generating more than one Markov chain or realization at simulation and by applying a Monte Carlo scheme. That is, at the Simulation

step m realizations $\boldsymbol{\varphi}^{(q,l)} \sim p\left(\cdot \mid \boldsymbol{y}; \boldsymbol{\theta}^{(q-1)}\right)$, $1 \leq l \leq m$, are drawn, and in the SA step the approximation of the conditional expectation is updated as

$$s_q(\theta) = s_{q-1}(\boldsymbol{\theta}) + \gamma_q \left(\frac{1}{m} \sum_{l=1}^{m} \log p\left(\boldsymbol{y}, \boldsymbol{\varphi}^{(q,l)}; \boldsymbol{\theta}\right) - s_{q-1}(\boldsymbol{\theta}) \right).$$

If the complete-data model belongs to the exponential family, that is, if

$$\log p(\boldsymbol{y}, \boldsymbol{\varphi}; \boldsymbol{\theta}) = -\Psi(\boldsymbol{\theta}) + \langle S(\boldsymbol{y}, \boldsymbol{\varphi}), \xi(\boldsymbol{\theta}) \rangle$$

where $S(\boldsymbol{y}, \boldsymbol{\varphi})$ represents a sufficient statistic of the data, then, the SA step reduces to:

$$F_q = F_{q-1} + \gamma_q \left(\frac{1}{m} \sum_{l=1}^{m} S(\boldsymbol{y}, \boldsymbol{\varphi}^{(q,l)}) - F_{q-1} \right) \tag{1.14}$$

and $s_q(\boldsymbol{\theta}) = -\Psi(\boldsymbol{\theta}) + \langle F_q, \xi(\boldsymbol{\theta}) \rangle$; that is, the actualization is made only on the sufficient statistic. This scheme can be applied even to models outside the exponential family, provided that a part of the model belongs to this family. However, we cannot speak of updating a sufficient statistic, but rather of a data summary function. Under general circumstances (Delyon et al., 1999), the convergence of the parameter sequence $\boldsymbol{\theta}^{(q)}$ toward a (local) maximum of the likelihood $\hat{\boldsymbol{\theta}}$ is guaranteed, regardless of the starting point $\boldsymbol{\theta}^{(0)}$ (Celeux et al., 1995).

The sequence of stepsizes $\{\gamma_q\}_{q \in \mathbb{N}}$ is usually set to 1 during the first iterations to avoid getting stuck in local maxima. In this way the first iterations are identical to those of the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990), which is known for its slow convergence rate. To avoid this scenario, in later iterations of SAEM γ_q decreases to zero to force convergence with fewer iterations. Details of application of the SAEM algorithm to complex mixed-effects models can be found in Meza et al. (2012); Márquez et al. (2023); Arribas-Gil et al. (2014) and de la Cruz et al. (2024).

1.4 SAEM-based methods for statistical inference

Following the proposition of the SAEM algorithm as a estimation method for the complex mixed models to be addressed, the necessity arises to introduce additional methods grounded in the aforementioned algorithm for the execution of covariate statistical significance tests. As previously discussed, a notable strength of mixed models is their capacity to accommodate covariates, thereby facilitating the understanding of the underlying dynamics behind observed data. Consequently, it becomes essential to establish statistical methodologies that can accurately assess the validity of this inclusion.

To this end, the most commonly used tests are usually the following:

- **Wald test:** this test assesses whether a parameter is statistically different from a specific value (usually zero). Let $\hat{\beta}$ be the estimator of the fixed parameter, and $\text{Var}(\beta)$ its estimated variance. The Wald statistic for β is defined as:

$$W = \frac{\hat{\beta}^2}{\text{Var}(\hat{\beta})}$$

and, under the null hypothesis $H_0 : \beta = 0$, it approximately follows a χ^2_1 distribution.

- **Likelihood ratio test (LRT):** LRT compares two nested models: a complete model M_1 and a reduced model M_0, where the second is a special case of the first under a null hypothesis. The statistic is defined as:

$$\Lambda = -2\left[\log L(\hat{\theta}_0) - \log L(\hat{\theta}_1)\right] \tag{1.15}$$

where $\log L(\hat{\theta}_k)$ is the likelihood of model $M_k, k = 0, 1$. Under the null hypothesis and regular conditions, $\Lambda \sim \chi^2_d$, where d is the difference in the number of parameters.

To perform these tests, it is clear that the variances of the estimators obtained and the log-likelihood of the models are needed. Fortunately, SAEM has techniques with which approximations of these values can be calculated. These techniques are presented in the following sections.

1.4.1 SAEM-based approximation of the variance-covariance matrix

In the case of maximum likelihood estimation, the variance-covariance matrix of $\boldsymbol{\theta}$ can asymptotically be calculated based on the Fisher information matrix of the model, which for complex models cannot be computed in a closed form. Based on the Louis's missing information principle (Louis, 1982) it is possible to compute an estimation of the Fisher information matrix. According to this principle, we have the identity:

$$\partial^2_\theta \log p(\boldsymbol{y};\theta) = \mathbb{E}\left(\partial^2_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi};\theta)|\boldsymbol{y};\theta\right) + \mathrm{Cov}\left(\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi};\theta)|\boldsymbol{y};\theta\right)$$

where

$$\begin{aligned}\mathrm{Cov}\left(\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi};\theta)|\boldsymbol{y};\theta\right) =& \mathbb{E}\left(\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi};\theta)\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi};\theta)^\top|\boldsymbol{y};\theta\right)\\ &- \mathbb{E}\left(\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi};\theta)|\boldsymbol{y};\theta\right)\mathbb{E}\left(\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi};\theta)|\boldsymbol{y};\theta\right)^\top\end{aligned}$$

Given this, the second order derivative of the observed likelihood function with respect to parameter $\hat{\theta}$, $\partial^2_\theta L(\hat{\theta};\boldsymbol{y})$, can be approximated by the sequence $\{H_q\}_{q\in\mathbb{N}}$ which is calculated at iteration q of the SAEM algorithm as:

$$\begin{aligned}D_q &= D_{q-1} + \gamma_q\left[\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)}) - D_{q-1}\right]\\ G_q &= G_{q-1} + \gamma_q\left[\partial^2_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)})\right.\\ &\quad\left.+\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)})\partial_\theta \log p(\boldsymbol{y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)})^\top - G_{q-1}\right]\\ H_q &= G_q - D_q D_q^\top.\end{aligned}$$

At convergence, $-H_q^{-1}$ can be used to approximate the covariance matrix of the parameter estimates (Zhu and Lee, 2002; Cai, 2010), which are useful for hypothesis testing for the different parameters of the model.

1.4.2 Approximation of the log-likelihood using Importance Sampling

The log-likelihood of the observed data cannot be computed in closed form for complex mixed effects models, but its estimation is required to perform the LRT and to compute information criteria for a given model. One approximation method is given by the application of the Importance Sampling algorithm (Kloek and van Dijk, 1978).

Let $\mathcal{LL}_y(\hat{\theta})$ be the log-likelihood of the model at the vector of population parameter estimates, that is $\mathcal{LL}_y(\hat{\theta}) = \log p(\boldsymbol{y}; \hat{\theta})$ where $p(\boldsymbol{y}; \hat{\theta}) = L(\hat{\theta}; \boldsymbol{y})$ is the joint probability distribution function of the observed data given $\hat{\theta}$. Notice that $\mathcal{LL}_y(\hat{\theta}) = \log p(\boldsymbol{y}; \hat{\theta}) = \sum_{i=1}^{N} \log p(y_i; \hat{\theta})$ and, for some *proposal distribution* $\tilde{p}_{\varphi_i}$ absolutely continuous with respect to p_{φ_i}, we have

$$p(y_i; \hat{\theta}) = \int p(y_i, \varphi_i; \hat{\theta}) d\varphi_i = \int p(y_i|\varphi_i; \hat{\theta}) \frac{p(\varphi_i; \hat{\theta})}{\tilde{p}(\varphi_i; \hat{\theta})} \tilde{p}(\varphi_i; \hat{\theta}) d\varphi_i = \mathbb{E}_{\tilde{p}} \left[p(y_i|\varphi_i; \hat{\theta}) \frac{p(\varphi_i; \hat{\theta})}{\tilde{p}(\varphi_i; \hat{\theta})} \right].$$

That is, $p(y_i; \hat{\theta})$ can be expressed as an expectation which can be approximated by:

1. Obtain a random sample of size K $\varphi_i^{(1)}, \ldots, \varphi_i^{(K)}$ from the proposal distribution $\tilde{p}_{\varphi_i}$;
2. Compute the empirical mean $\hat{p}_{(i,K)} = \frac{1}{K} \sum_{k=1}^{K} p(y_i|\varphi_i^{(k)}; \hat{\theta}) \frac{p(\varphi_i^{(k)}; \hat{\theta})}{\tilde{p}(\varphi_i^{(k)}; \hat{\theta})}$

An optimal proposal distribution would be the conditional distribution $p_{\varphi_i|y_i}$ since in that case the estimator of the expectation has zero variance. But since the closed form expression of the distribution is not available, we choose a proposal *close* to this optimal distribution, based on the empirically estimated conditional mean and variance, $\mu_i = \hat{\mathbb{E}} \left[\varphi_i | y_i; \hat{\theta} \right]$ and $\sigma_i^2 = \hat{\text{Var}} \left[\varphi_i | y_i; \hat{\theta} \right]$, of the simulated random effects during the Simulation step of the SAEM algorithm. Then, the proposed candidate $\varphi_i^{(k)}$, with $k = 1, \ldots, K$, is drawn with a noncentral student t-distribution $\varphi_i^{(k)} = \mu_i + \sigma_i \times T_{i,k}$, with $T_{i,k} \sim t_\nu$ i.i.d., where t_ν denotes a Student's t-distribution with ν degrees of freedom.

CHAPTER 2

SAEM-ZIBR: STOCHASTIC ESTIMATION METHOD FOR A TWO-PART MIXED MODEL FOR LONGITUDINAL COMPOSITIONAL DATA*

2.1 Introduction

The Zero-Inflated Beta Regression (ZIBR) (Chen and Li, 2016) is defined as a two-stage mixed effects model based on the work of Ospina and Ferrari (2012), that allows the inclusion of clinical covariates, both to explain the presence or absence of a certain bacterial taxon and, in case of presence, the influence of these covariates on the relative abundance of the taxon. It should be clarified that this model works with microbial abundance data as proportions rather than counts, which is why it uses the Beta distribution. This model provides a comprehensive approach to analyse longitudinal compositional microbiome data taking into account its bounded nature, skewness and the over-abundance of zeros. Since its appearance, ZIBR has been succesfully applied in several studies (Hu et al., 2022; D'Agata et al., 2019) as it is capable of treating the features above mentioned, explaining within-subject correlations and providing methods to conduct hypothesis tests on the significance of covariates.

As for other complex mixed-effects models, the Maximum Likelihood (ML) estimation method for ZIBR proposed by its authors relies on approximating the log-likelihood using Gauss-Hermite quadrature and numerical optimization of this expression. However, there is evidence that in certain scenarios the estimators obtained in this way may be biased and, in the case of generalized mixed models, even be outperformed by other techniques (Handayani et al., 2017). Additionally, the proposed estimation method can only be used for balanced data; that is, with the same number of observations per individual. In

*This chapter, with some changes, corresponds to an article which is a joint work with Cristian Meza and Ana Arribas-Gil submitted for review to *Statistical Modeling* and partially presented in the 32nd International Biometric Conference (Atlanta, USA, December 2024).

clinical studies this is not often the case, resulting in analysis with potentially misleading conclusions (Powney et al., 2014). Therefore, new strategies are required to address the challenges posed by missing data (Myers, 2000). The ZIBR model can also be thought as a particular case of the Generalized Additive Model for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005) and its parameters can be estimated in this framework. Although in this case the available maximum estimation strategy does allow to handle missing data, it still relies on a penalized local log-likelihood approximation which can cause problems in performing likelihood ratio tests, particularly in complex GLMMs such as the ZIBR (see Stasinopoulos et al., 2017).

All things considered, the versatile features of the ZIBR model make it a promising choice for the analysis of compositional microbiome data, despite possible drawbacks in existing estimation strategies. Therefore, for the final purpose of precisely identifying those taxa responsive to disease onset, changes in environmental conditions or specific interventions, any possible improvement in the estimation method which is able to provide more accurate estimates will amount to significant progress in the understanding of human microbiome and its relation to human health.

Following this aim, in this chapter we propose a new estimation framework for the ZIBR model on longitudinal compositional data based on the Stochastic Approximation EM (SAEM) algorithm (Delyon et al., 1999). This algorithm provides an exact maximum likelihood estimation strategy in missing data models for which the EM algorithm (Dempster et al., 1977) is not directly applicable because the complexity of the likelihood function does not allow for exact calculation of its conditional expectation. This would be the case of the ZIBR model. The SAEM algorithm not only preserves the good behaviour of the EM algorithm in terms of convergence, unbiasedness and monotonicity, but has also shown interesting properties in complex mixed models (Márquez et al., 2023; Meza et al., 2012) and includes procedures for statistical inference and hypothesis testing (Samson et al., 2007). Furthermore, it can be combined with MCMC techniques for improved performance (Kuhn and Lavielle, 2005) and can be extended to Restricted Maximum Likelihood (REML) estimation (Meza et al., 2007). We extend the algorithm to distributions not belonging to the exponential family and derive the explicit expressions at all its steps, for both parameter estimation and log-likelihood approximation, once the ML estimators have been obtained. We also obtain approximations of the standard errors of the estimators, by means of the stochastic approximation of the Fisher information matrix. This allows us to provide a comprehensive estimation approach that avoids downsides related to likelihood approximations, is able to incorporate unbalanced data, and facilitates the inference pipeline, from modeling to covariate effects testing, under the same framework.

The structure of the chapter is as follows: in Section 2.2, we introduce the ZIBR model and develop the SAEM based inference method to be used in our work. In Section 2.3 we present simulation studies on synthetic data generated under different settings, comparing the results obtained with our approach and those given by estimation based on likelihood approximation or penalization, and in Section 2.4 we assess the behaviour of the proposed routine on a dataset coming from clinical microbiome studies. Finally, Section 2.5 closes with the main conclusions, a discussion of the results, and possible limitations and future developments.

2.2 The ZIBR model and its SAEM-based estimation

2.2.1 Definition of the model

The ZIBR model describes the presence and abundance of a single bacterial taxon on different individuals over time, and can be subsequently applied to different bacteria. Let $y_{it} \in (0,1)$ be the relative abundance of a bacterial taxon in the individual i at time t, $1 \leq i \leq N$, $1 \leq t \leq T_i$. The model assumes that y_{it} follows the distribution:

$$y_{it} \sim \begin{cases} 0 & \text{with prob. } 1 - p_{it}, \\ \text{Beta}(u_{it}\phi, (1-u_{it})\phi) & \text{with prob. } p_{it} \end{cases} \tag{2.1}$$

with $\phi > 0$ and $0 < u_{it}, p_{it} < 1$. These two last components are characterized by

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = a_i + X_{it}^T\alpha, \qquad \log\left(\frac{u_{it}}{1-u_{it}}\right) = b_i + Z_{it}^T\beta, \tag{2.2}$$

where a_i and b_i are individual specific intercepts, α and β are vectors of regression coefficients and X_{it} and Z_{it} are covariates for each individual and time point. We further consider that each one of the random intercepts follows a normal distribution, independently from each other:

$$a_i \sim N(a, \sigma_1^2), \qquad b_i \sim N(b, \sigma_2^2).$$

From Equations 2.1 and 2.2, it can be seen that the ZIBR model explicitly includes a component that is responsible for the presence of zeros in the data. It is also clear that conveniently defined covariates X_{it} and Z_{it} can influence both the probability of presence or absence of a bacterial taxon (through the logistic regression that defines p_{it}) and the magnitude of its relative abundance (through the u_{it} component in the proposed Beta distribution). Furthermore, the inclusion of a random intercept allows modeling correlations in observations from the same individual. Even though it is easy to expand the definition to consider random slopes, in practice it is enough to consider just a random intercept (Min and Agresti, 2005).

The model parameter $\boldsymbol{\theta} = (\phi, a, b, \alpha, \beta, \sigma_1^2, \sigma_2^2)$ can be estimated by maximum likelihood. The likelihood function for data $\boldsymbol{y} = (y_{it}, 1 \leq i \leq N, 1 \leq t \leq T_i)$ is

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{N} \int_{\mathbb{R}} \int_{\mathbb{R}} \prod_{t=1}^{T_i} (1-p_{it})^{\mathbb{1}_{\{y_{it}=0\}}} \left[p_{it} f(y_{it}; u_{it}, \phi)\right]^{\mathbb{1}_{\{y_{it}>0\}}} g\left(a_i, b_i | a, \sigma_1^2, b, \sigma_2^2\right) da_i db_i \tag{2.3}$$

where $f(y_{it}; u_{it}, \phi)$ is the Beta density function with parameters u_{it} and ϕ on y_{it}:

$$f(y_{it}; u_{it}, \phi) = \frac{\Gamma(\phi)}{\Gamma(u_{it}\phi)\Gamma((1-u_{it})\phi)} y_{it}^{u_{it}\phi - 1} (1-y_{it})^{(1-u_{it})\phi - 1}, \tag{2.4}$$

and g is the product of the two univariate normal density functions of random effects a_i and b_i.

Given the impossibility of analytical calculation of the integral shown in Equation 2.3, an approximation can be achieved by means of the Gauss-Hermite quadrature (GHQ). With this approximation, and through numerical optimization, the maximum likelihood

estimators for θ can be found as proposed by Chen and Li (2016). Hypothesis tests for the significance of covariates can also be conducted, in particular the Likelihood Ratio Test (LRT). The implementation of this approach is available in the `ZIBR` package (Zhang Chen, 2023) developed for the R software. In addition to this alternative, the `gamlss` package (Rigby and Stasinopoulos, 2005) can also be used, which, in a similar manner to the aforementioned one, is based on a penalized quasilikelihood approximation and its optimization based on numerical algorithms (Rigby and Stasinopoulos, 1996). A notable advantage of this implementation, however, is its capacity to handle unbalanced data, which renders it a suitable option for a comparative analysis of the results obtained in this study.

2.2.2 The SAEM algorithm for ZIBR parameter estimation

As we have seen before, a mixed model can be considered as an unobserved data problem and therefore be addressed using the SAEM algorithm. Let us consider $\varphi_i = (a_i, b_i)$, $1 \leq i \leq N$, the non-observed data. By the definition of the ZIBR model φ_i follows the multivariate normal distribution $\varphi_i \sim N(\boldsymbol{\mu}, \mathbf{G})$ with $\boldsymbol{\mu} = (a, b)$ and $\mathbf{G} = diag(\sigma_1^2, \sigma_2^2)$. With the usual notation $\boldsymbol{y} = (y_{it} : \ 1 \leq i \leq N, 1 \leq t \leq T_i)$ and $\boldsymbol{\varphi} = (\varphi_i : \ 1 \leq i \leq N)$, the complete-data likelihood writes:

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\varphi}; \theta) =& p(\boldsymbol{y}|\boldsymbol{\varphi}; \alpha, \beta, \phi) p(\boldsymbol{\varphi}|\boldsymbol{\mu}, \mathbf{G}) \\ \propto& |\mathbf{G}|^{-\frac{N}{2}} \prod_i \exp\left(-\frac{(\varphi_i - \boldsymbol{\mu})^T \mathbf{G^{-1}} (\varphi_i - \boldsymbol{\mu})}{2}\right) \\ & \times \prod_{i,t} (1 - p_{it})^{\mathbb{1}_{\{y_{it}=0\}}} p_{it}^{\mathbb{1}_{\{y_{it}>0\}}} f(y_{it}; u_{it}, \phi)^{\mathbb{1}_{\{y_{it}>0\}}}. \end{aligned} \tag{2.5}$$

Like most zero-inflated models, the ZIBR model cannot be considered part of the exponential family (Eggers, 2015). However, the decomposition presented in Equation 2.5 allows us to propose a simplified structure for the SAEM algorithm (check Equation 1.14 in Section 1.3). For the multivariate normal part corresponding to the random effects, the actualization in the SA step is done on the respective sufficient statistics. For the mixture distribution corresponding to the observed data, $\boldsymbol{y}|\boldsymbol{\varphi}; \alpha, \beta, \phi$, maximization of the conditional log-likelihood is followed by estimates updates, as suggested for non-exponential family models (Comets et al., 2021). Then, the Maximum Likelihood iterative estimation algorithm for the parameters of the ZIBR model writes, for a given starting point $\theta^{(0)}$ and at iteration q, as:

1. **Simulation step:** draw $\varphi_i^{(q)}, i = 1, \cdots, N$ from the distribution $p(\cdot|\boldsymbol{y}; \theta^{(q-1)})$.
2. **Stochastic Approximation step:** update the summary data functions $F_1(\boldsymbol{y}, \boldsymbol{\varphi})$ and $F_2(\boldsymbol{y}, \boldsymbol{\varphi})$ with the scheme:

$$\begin{aligned} F_1^{(q)}(\boldsymbol{y}, \boldsymbol{\varphi}) &= F_1^{(q-1)}(\boldsymbol{y}, \boldsymbol{\varphi}) + \gamma_q \left(\sum_i \varphi_i^{(q)} - F_1^{(q-1)}(\boldsymbol{y}, \boldsymbol{\varphi}) \right) \\ F_2^{(q)}(\boldsymbol{y}, \boldsymbol{\varphi}) &= F_2^{(q-1)}(\boldsymbol{y}, \boldsymbol{\varphi}) + \gamma_q \left(\sum_i \varphi_i^{(q)} \varphi_i^{(q)^T} - F_2^{(q-1)}(\boldsymbol{y}, \boldsymbol{\varphi}) \right). \end{aligned} \tag{2.6}$$

where $\{\gamma_q\}_{q\in\mathbb{N}}$ is a decreasing sequence of stepsizes with $\gamma_1 = 1$.

3. **Maximization step:** update the parameters of the model with

$$\begin{aligned} \boldsymbol{\mu}^{(q)} &= \frac{F_1^{(q)}(\boldsymbol{y}, \boldsymbol{\varphi})}{N} \\ \mathbf{G}^{(q)} &= \frac{F_2^{(q)}(\boldsymbol{y}, \boldsymbol{\varphi})}{N} - \frac{F_1^{(q)}(\boldsymbol{y}, \boldsymbol{\varphi}) F_1^{(q)}(\boldsymbol{y}, \boldsymbol{\varphi})^T}{N^2} \end{aligned} \tag{2.7}$$

Given the form of the model definition in the Beta part, steps 2 and 3 are modified by first calculating

$$\left(\tilde{\beta}^{(q)}, \tilde{\phi}^{(q)}\right) = \tag{2.8}$$
$$\arg\max_{\beta,\phi} \sum_{i,t} \left[\mathbb{1}_{\{y_{it}>0\}} \left(\log \frac{\Gamma(\phi)}{\Gamma\left(u_{it}^{(q)}\phi\right)\Gamma((1-u_{it}(q))\phi)} + u_{it}^{(q)} \phi \log y_{it} + \phi \left(1 - u_{it}^{(q)}\right) \log\left(1 - y_{it}\right) \right) \right]$$

and

$$\tilde{\alpha}^{(q)} = \arg\max_{\alpha} \sum_{i,t} \left[\mathbb{1}_{\{y_{it}>0\}} \log\left(p_{it}^{(q)}\right) + \mathbb{1}_{\{y_{it}=0\}} \log\left(1 - p_{it}^{(q)}\right) \right]. \tag{2.9}$$

where $u_{it}^{(q)} = u_{it}^{(q)}(b_i, \beta)$ and $p_{it}^{(q)} = p_{it}^{(q)}(a_i, \alpha)$ are calculated using $\varphi_i^{(q)}$ and Equation 2.2. Maximization in (2.8) and (2.9) is achieved numerically. Finally, the values are updated by doing

$$\begin{aligned} \phi^{(q)} &= \phi^{(q-1)} + \gamma_q \left(\tilde{\phi}^{(q)} - \phi^{(q-1)}\right) \\ \alpha^{(q)} &= \alpha^{(q-1)} + \gamma_q \left(\tilde{\alpha}^{(q)} - \alpha^{(q-1)}\right) \\ \beta^{(q)} &= \beta^{(q-1)} + \gamma_q \left(\tilde{\beta}^{(q)} - \beta^{(q-1)}\right) \end{aligned} \tag{2.10}$$

Let us discuss the details of this implementation. As mentioned in Section 1.3, the choice of the starting point $\theta^{(0)}$ for SAEM does not affect its convergence; however, it is recommended to use values obtained in previous studies or with other estimation methods. Following the example of the existing implementation of the `saemix` package (Comets et al., 2017), we will use γ_q defined as follows:

$$\gamma_q = \begin{cases} 1 & \text{if } q \leq K_1, \\ \frac{1}{q-K_1} & \text{if } K_1 < q \leq K_1 + K_2. \end{cases}$$

where $K_1 + K_2$ is the total number of iterations.

We also discussed in Section 1.3 that it is possible to improve the performance of the algorithm by taking multiple sequences or Markov chains in the Simulation step, and using Monte Carlo in Equations 2.7 and 2.10. Furthermore, during the SA step, we obtain sequences that allow to estimate $\mathbb{E}\left(\varphi_i | y_i; \hat{\theta}\right)$ and $\text{Var}\left(\varphi_i | y_i; \hat{\theta}\right)$ to be calculated, values necessary to approximate the log-likelihood through Importance Sampling, with which the Likelihood Ratio Test (LRT) can be computed, as presented in Section 1.4.1. We will also use the stochastic approximation of the Fisher matrix to calculate the variances of the estimators and implement the Wald test, as mentioned in Section 1.4.2.

2.3 Simulation studies

To evaluate the behavior of the proposed estimation method, and to compare it with existing alternatives, we conducted several simulation studies. It is worth noticing that the GHQ approach does not allow to deal with a different number of observations per individual, which is possible with our SAEM-based approach and the `gamlss` package. Therefore, we present simulations with balanced data first. Additional simulations on unbalanced data are also provided in Section 2.3.2, in which the performance of SAEM on the unbalanced datasets is compared with the use of the GHQ algorithm on balanced datasets obtained from imputation, and with `gamlss` without imputation. Covariates significance analysis based on the LRT and the Wald test are also presented in Section 2.3.3

2.3.1 Balanced datasets

Setup

We use two different settings for generating synthetic data under the ZIBR model (Equations 2.1 and 2.2). The parameters for each configuration were chosen as follows:

- *Setting 1:* $a = b = -0.5$, $\alpha = \beta = 0.5$, $\sigma_1 = 3.2$, $\sigma_2 = 2.6$, $\phi = 6.4$.
- *Setting 2:* $a = b = -0.5$, $\alpha = \beta = 0.5$, $\sigma_1 = 0.7$, $\sigma_2 = 0.5$, $\phi = 6.4$.

In the balanced scenario, for both Settings 1 and 2 the number of individuals $N = 100$ will remain fixed, but the number of observations per individual T_i will change, making $T_i = T$ with $T = 3, 5, 10$. In addition, a variable X is defined that mimics the concept of treatment and control groups, making $X = 0$ for the first half of individuals and $X = 1$ for the other half. Furthermore, we consider the same variable as covariate in both parts of the models, making $Z = X$.

For both Settings 1 and 2, $R = 1000$ datasets were generated, and the SAEM estimation was implemented with $m = 5$ chains and $(K_1, K_2) = (750, 250)$, having therefore 1000 total iterations, with a starting point $\theta_0 = (\phi_0, a_0, b_0, \alpha_0, \beta_0, \sigma_{1,0}, \sigma_{2,0}) = (8, -0.3, -0.2, 0.7, 0.8, 0.38, 0.31)$. On the other hand, for the GHQ method, 30 quadrature points were considered, the default value in the `ZIBR` package.

Results

Table 2.1 shows the performance analysis of the two estimation methods for Settings 1 and 2 on balanced datasets, evaluated by bias $\left(\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^r - \theta\right)$, mean absolute error $\left(\text{MAE} = \frac{1}{R}\sum_{r=1}^{R}|\hat{\theta}^r - \theta|\right)$ and root mean square error $\left(\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^r - \theta)^2}\right)$.

Analyzing these results globally, and the global estimates distributions in Figures 2.1 and 2.2, we can see how SAEM estimates are always centered (except for a small bias for σ_2 in Setting 1), whereas GHQ and GAMLSS methods present strong biases or bimodality for different parameters: b, β, σ_2 and ϕ for GHQ in Setting 1 and σ_1 (only Setting 1), σ_2 and ϕ (both settings) for GAMLSS. A thorough examination of these results reveals that the

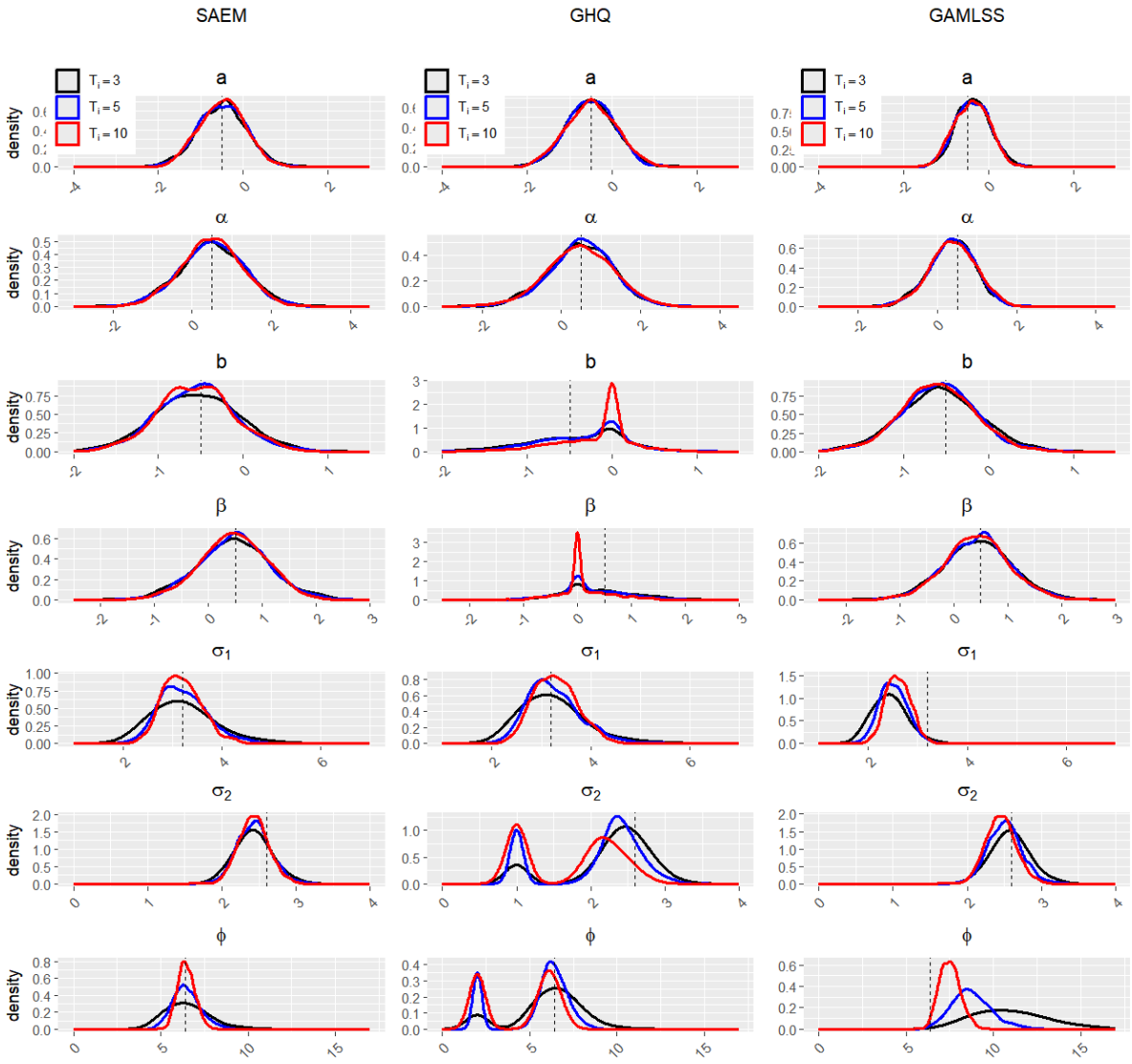


Figure 2.1: *Estimated density of the parameters obtained by the SAEM algorithm, the GHQ method and the GAMLSS procedure on artificial balanced datasets simulated under Setting 1. The dotted vertical line represents the true value of the parameter.*

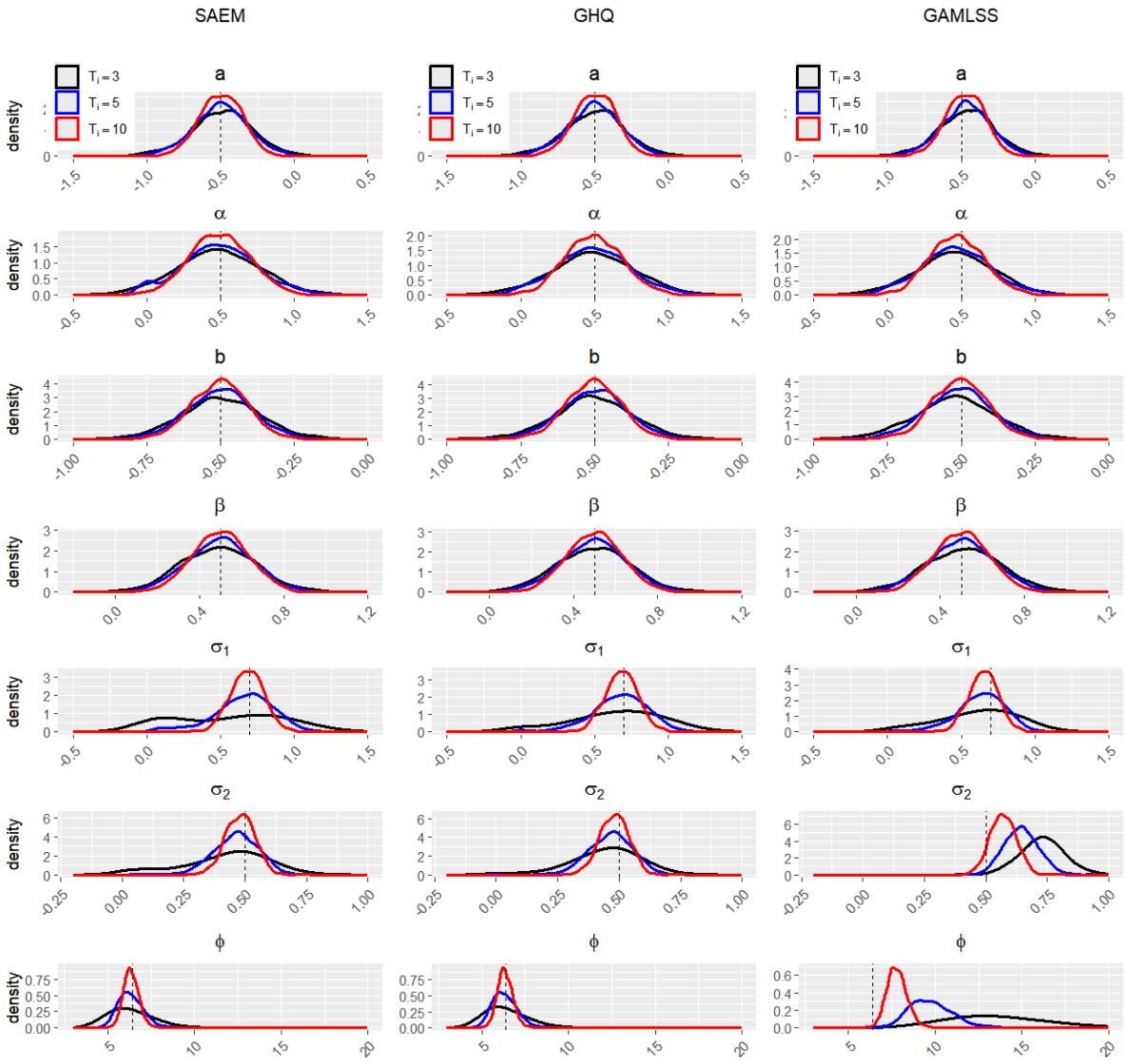


Figure 2.2: *Estimated density of the parameters obtained by the SAEM algorithm, the GHQ method and the GAMLSS procedure on artificial balanced datasets simulated under Setting 2. The dotted vertical line represents the true value of the parameter.*

Table 2.1: *Summary statistics of the results obtained by SAEM algorithm, the GHQ procedure and the GAMLSS method on balanced data sets over 1000 simulation runs. For each parameter value and number of observations per individual, T_i, bold numbers indicate the lowest (absolute) value for each of bias, RMSE and MAE.*

Parameter	**Value**		**Bias**	**RMSE**	**MAE**	**Bias**	**RMSE**	**MAE**	**Bias**	**RMSE**	**MAE**
				SAEM			**GHQ**			**GAMLSS**	
a	-0.5	$T_i = 3$	0.0064	0.5896	0.4629	**-0.0010**	0.6509	0.4734	0.1460	**0.4348**	**0.3449**
		$T_i = 5$	0.0081	0.5626	0.4526	**0.0044**	0.5634	0.4514	0.1302	**0.4281**	**0.3436**
		$T_i = 10$	**0.0154**	0.5369	0.4306	0.0221	0.5947	0.4749	0.1139	**0.4269**	**0.3469**
α	0.5	$T_i = 3$	**0.0091**	0.8277	0.6519	0.0215	0.9634	0.6587	-0.1343	**0.5846**	**0.4626**
		$T_i = 5$	**0.0046**	0.8001	0.6340	0.0138	0.7926	0.6186	-0.1203	**0.5822**	**0.4590**
		$T_i = 10$	**-0.0046**	0.7533	0.6007	-0.0253	0.8363	0.6625	-0.1026	**0.5827**	**0.4673**
b	-0.5	$T_i = 3$	**-0.0414**	0.4936	0.3970	0.0970	0.5570	0.4607	-0.0723	**0.4744**	**0.3767**
		$T_i = 5$	**-0.0532**	0.4511	0.3560	0.1752	0.5057	0.4317	-0.0829	**0.4416**	**0.3478**
		$T_i = 10$	**-0.0552**	0.4392	0.3512	0.3197	0.5155	0.4582	-0.0993	**0.4375**	**0.3481**
β	0.5	$T_i = 3$	**-0.0384**	0.6827	0.5372	-0.1442	0.7253	0.5788	-0.0493	**0.6529**	**0.5122**
		$T_i = 5$	**-0.0216**	0.6530	0.5092	-0.2532	0.6784	0.5544	-0.0526	**0.6160**	**0.4809**
		$T_i = 10$	**-0.0275**	0.6083	0.4825	-0.3828	**0.5972**	0.5202	-0.0660	0.5980	**0.4704**
σ_1	3.2	$T_i = 3$	**0.0284**	**0.6423**	**0.4961**	0.0693	1.1306	0.5277	-0.7702	0.8400	0.7754
		$T_i = 5$	**0.0259**	**0.4920**	**0.3904**	0.0631	0.5534	0.4238	-0.6914	0.7461	0.6932
		$T_i = 10$	**-0.0012**	**0.3951**	**0.3170**	0.0870	0.4433	0.3550	-0.5939	0.6428	0.5956
σ_2	2.6	$T_i = 3$	-0.1720	0.2979	0.2429	-0.3206	0.6612	0.4347	**-0.0142**	**0.2478**	**0.1957**
		$T_i = 5$	-0.1639	0.2779	0.2270	-0.4970	0.8065	0.5605	**-0.0899**	**0.2425**	**0.1932**
		$T_i = 10$	-0.1699	0.2635	0.2162	-0.8565	1.0651	0.8674	**-0.1457**	**0.2480**	**0.2018**
ϕ	6.4	$T_i = 3$	**0.1301**	**1.2251**	**0.9465**	-0.3696	2.0104	1.4482	4.6834	5.1779	4.6834
		$T_i = 5$	**0.0895**	**0.8059**	**0.6283**	-0.9741	2.1842	1.4390	2.3698	2.6248	2.3713
		$T_i = 10$	**0.0645**	**0.4940**	**0.3931**	-1.9106	2.8099	2.0202	1.1085	1.2564	1.1139
				SAEM			**GHQ**			**GAMLSS**	
a	-0.5	$T_i = 3$	0.0106	0.2101	0.1659	**0.0026**	0.2060	0.1623	0.0376	**0.1935**	**0.1540**
		$T_i = 5$	0.0017	0.1813	0.1432	**-0.0003**	0.1780	0.1404	0.0317	**0.1688**	**0.1340**
		$T_i = 10$	0.0033	0.1414	0.1144	**0.0023**	0.1380	0.1114	0.0276	**0.1337**	**0.1080**
α	0.5	$T_i = 3$	-0.0120	0.2912	0.2283	**-0.0040**	0.2859	0.2243	-0.0385	**0.2663**	**0.2102**
		$T_i = 5$	-0.0060	0.2551	0.2025	**-0.0032**	0.2482	0.1973	-0.0349	**0.2346**	**0.1870**
		$T_i = 10$	-0.0042	0.1997	0.1607	**-0.0029**	0.1910	0.1528	-0.0282	**0.1833**	**0.1468**
b	-0.5	$T_i = 3$	**0.0006**	0.1348	0.1063	-0.0025	**0.1331**	**0.1043**	-0.0212	0.1409	0.1102
		$T_i = 5$	-0.0031	0.1099	**0.0873**	**-0.0028**	**0.1096**	0.0874	-0.0094	0.1116	0.0885
		$T_i = 10$	-0.0032	**0.0922**	0.0732	**-0.0031**	**0.0922**	**0.0729**	-0.0040	0.0928	0.0737
β	0.5	$T_i = 3$	-0.0089	0.1792	0.1430	**-0.0049**	**0.1753**	**0.1403**	0.0122	0.1829	0.1463
		$T_i = 5$	**-0.0006**	0.1536	0.1223	-0.0011	**0.1516**	**0.1206**	0.0050	0.1542	0.1226
		$T_i = 10$	0.0015	0.1326	0.1060	**0.0012**	**0.1308**	**0.1044**	**0.0012**	0.1314	0.1052
σ_1	0.7	$T_i = 3$	-0.1448	0.3944	0.3239	**-0.0535**	0.3176	0.2501	-0.0845	**0.2768**	**0.2164**
		$T_i = 5$	-0.0394	0.2107	0.1605	**-0.0275**	0.1951	0.1499	-0.0643	**0.1808**	**0.1373**
		$T_i = 10$	-0.0192	0.1137	0.0910	**-0.0169**	0.1108	0.0883	-0.0489	**0.1106**	**0.0877**
σ_2	0.5	$T_i = 3$	-0.0812	0.1918	0.1407	**-0.0542**	**0.1529**	**0.1139**	0.2324	0.2481	0.2328
		$T_i = 5$	-0.0359	0.0991	0.0771	**-0.0337**	**0.0963**	**0.0749**	0.1453	0.1610	0.1460
		$T_i = 10$	-0.0193	0.0629	0.0499	**-0.0190**	**0.0624**	**0.0496**	0.0715	0.0891	0.0754
ϕ	6.4	$T_i = 3$	-0.0920	1.2570	1.0054	**0.0031**	**1.1890**	**0.9402**	7.4381	8.0589	7.4381
		$T_i = 5$	-0.0700	0.7283	0.5820	**-0.0620**	**0.7214**	**0.5767**	3.3718	3.6059	3.3718
		$T_i = 10$	**-0.0463**	0.4423	0.3518	-0.0472	**0.4409**	**0.3510**	1.3984	1.5107	1.4002

SAEM estimation achieves optimal bias performance in Setting 1. In Setting 2, the GHQ method exhibits slightly lower bias values. It merits attention, however, that GAMLSS attains the lowest root mean square error (RMSE) and mean absolute error (MAE) values in many parameters across both settings. It is worth noticing, nevertheless, that the GAMLSS estimates for ϕ are extremely biased in both settings. These two situations are common in estimators obtained by quasi-likelihood methods, as indicated by Nelder and Lee (1992).

A remarkable property of SAEM is its tendency to exhibit a consistent decrease in the error measures as the number of observations per individual increases, a phenomenon not consistently observed in the other analyzed alternatives. In Setting 1, where the values of the variance parameters are higher, it is evident that the worst GHQ results are obtained in the estimation of the parameters associated with the Beta part of the model (b, β, σ_2 and ϕ). This component has a complex functional form that could be incorrectly approximated when integrating by numerical methods. The SAEM approach is advantageous in this regard. In the context of GAMLSS, the estimation of the parameter controlling the overdispersion of the data, i.e., ϕ, is particularly challenging (as well as the estimation of σ_2 in Setting 2). According to Stasinopoulos et al. (2017), by defining ZIBR as a model with several random effects, GAMLSS estimates the variance components with a method prone to generating biases. Given that overdispersion is a common feature in real microbiota data, the GAMLSS procedure may not be suitable for modeling this phenomenon.

A more detailed analysis of the Figures shows that for Setting 1 the estimated densities of a and α are very similar in all methods, while the other parameters show marked differences. In the case of GHQ, the results for b and β are very skewed, while the distribution of σ_2 and ϕ shows bimodal behavior. This is due to the poor approximation of the log-likelihood, which leads to an erroneous estimation of these parameters. A review of the GAMLSS results reveals that the distribution of σ_1 is significantly deviated from the true parameter, while the distribution of ϕ exhibits considerable variability when observations per individual are limited. Figure 2.2, on the other hand, shows that for Setting 2 SAEM and GHQ are practically equivalent in the estimates densities, while GAMLSS differs only in that it performs poorly for σ_2 and ϕ. This indicates that the minor errors differences between the methods in favor of GHQ for this simulation scenario (Table 2.1) are not significant.

Regarding the execution times of the routines, GHQ and GAMLSS are faster in execution time than SAEM in the 3 cases considered in both Settings. On a computer with Intel Core i7-13700HX processor at 2.10 GHz, for $T_i = 3$ $(5, 10)$, on average, GAMLSS takes 0.94 $(1.03, 1.20)$ seconds, GHQ takes 1.77 $(3.39, 8.54)$ seconds, while SAEM takes 9.86 $(14.14, 25.30)$ seconds.

2.3.2 Unbalanced and interpolated datasets

Now we report some results obtained in the application of the ZIBR model to an unbalanced data scenario. In this context, some studies (Abe and Iwasaki, 2007) advise the use of imputation with the individual average or to remove some observations in the data for the use of longitudinal data methods that cannot cope with unbalanced designs. Therefore, this is the approach we consider here to be able to compare GHQ to SAEM on

this setting. However, as already mentioned, interpolation can create some inaccuracies in the estimations. In comparison, estimation using the SAEM algorithm does allow for its use directly on the original unbalanced data. We also compare the estimation obtained with the GAMLSS implementation, which does allow to work with unbalanced data.

Setup

In the context of an unbalanced data scenario, data generation will have two parts. First, we will use the parameters of Setting 2 ($a = b = -0.5$, $\alpha = \beta = 0.5$, $\sigma_1 = 0.7$, $\sigma_2 = 0.5$, $\phi = 6.4$.) to generate 1000 datasets, with $T_i = 10$ and different values for the number of individuals, $N \in \{50, 100, 200\}$. The variables X and Z are defined as usual. Then, we will randomly eliminate 20% of the observations from each data set; therefore, the median (interquartile range (IQR)) of the number of observations per individual in the three specifications is 8 (IQR: 7 to 9). Given the drop-out method we chose to simulate an unbalanced data situation, we can assume that we are in a case of MCAR (Missing Completely At Random) (Rubin, 1976). Finally, we will compare the performance of SAEM on unbalanced data with GHQ on interpolated data. For each individual, the interpolation process will be carried out as follows:

- if the missing value is between two known observations, linear interpolation will be performed; and
- if the missing value is at the end of the observations, it will be replaced with the last known value.

As in the main article, we compute the bias, MAE and RMSE of the estimations with all methods and the estimated densities of the parameter estimates.

Results

First of all, notice that, from the two simulations scenarios used in the analysis of balanced data (Section 2.3.1), the one used here (Setting 2) was the most favorable to the GHQ method. Table 2.2 shows the results of the estimation on unbalanced datasets for the SAEM algorithm and the GAMLSS method, and on interpolated datasets with the GHQ procedure. These results show that in most cases, the SAEM estimators outperform the GHQ and GAMLSS estimators by having a lower absolute bias. Furthermore, in the case of both SAEM and GAMLSS, the increase in the number of individuals reduces the MAE and RMSE values in all cases; something that cannot be said for the estimates obtained by GHQ. Although there are some scenarios in which GHQ obtains better results in error measures, this advantage is quite small, whereas when SAEM is superior, the differences in the values are much more noticeable, which can be clearly seen in Figure 2.3. In this scenario, the estimate using GAMLSS presents the best RMSE and MAE values of the three methods in most cases, although this advantage is generally quite small. Additionally, as the sample size increases, these estimates gradually approach those obtained by SAEM.

It is interesting to note that the worst results of GHQ are concentrated in the parameters related to the logistic part of the ZIBR model; that is, the one that controls zero

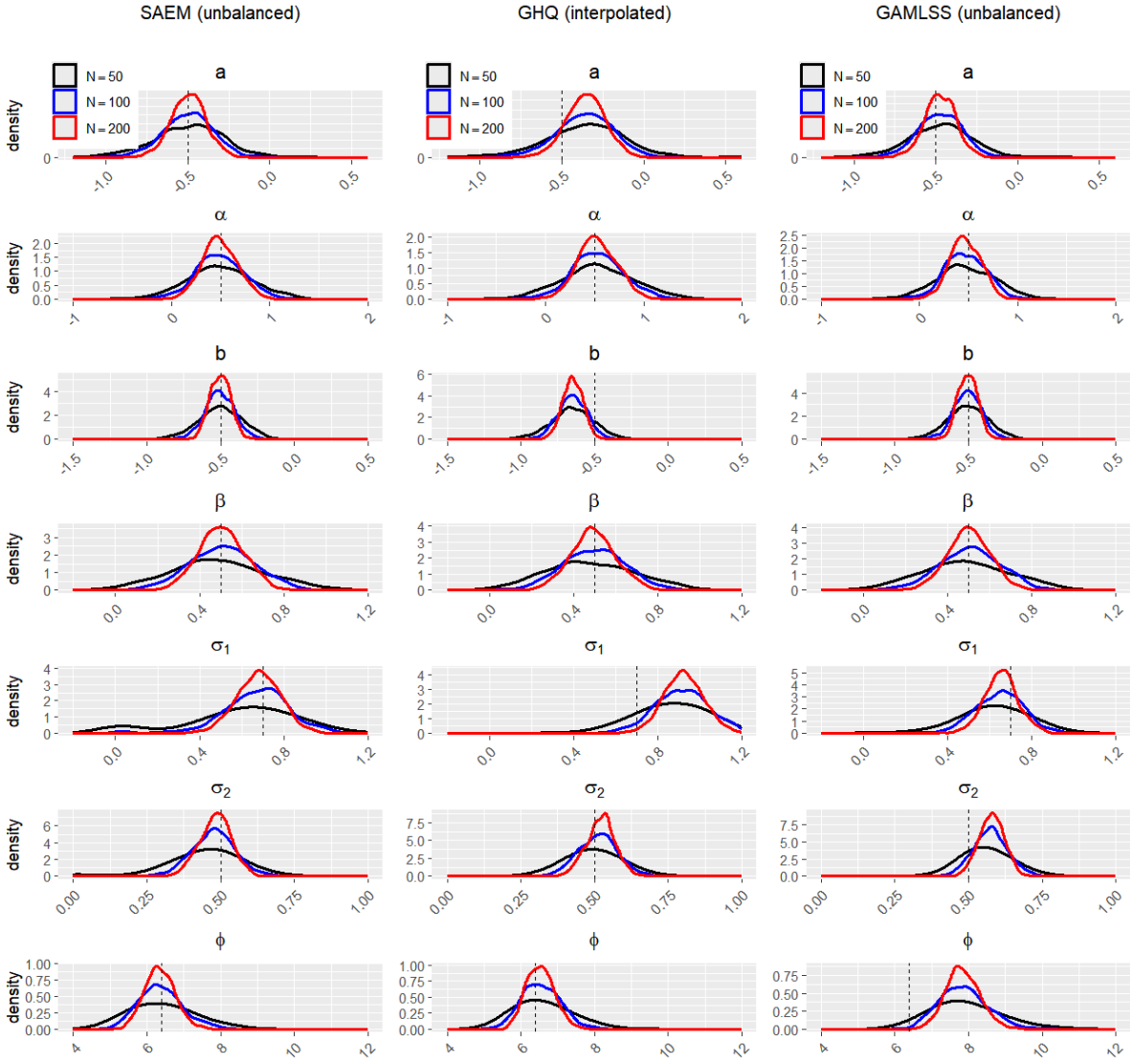


Figure 2.3: *Estimated density of the parameters obtained by the SAEM algorithm on unbalanced datasets and the GHQ procedure on interpolated datasets simulated under Setting 2. The dotted vertical line represents the true value of the parameter.*

Table 2.2: *Summary statistics of the results obtained by the SAEM algorithm and GAMLSS method on unbalanced and GHQ procedure on interpolated data sets over 1000 simulation runs. For each parameter value and number of individuals, N, bold numbers indicate the lowest (absolute) value for each of bias, RMSE and MAE.*

Parameter	Value		Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
			SAEM (unbalanced)			GHQ (interpolated)			GAMLSS (unbalanced)		
a	-0.5	$N = 50$	**0.0048**	0.2083	0.1662	0.1425	0.2726	0.2226	0.0264	**0.1874**	**0.1494**
		$N = 100$	**0.0015**	0.1508	0.1199	0.1408	0.2206	0.1826	0.0275	**0.1338**	**0.1081**
		$N = 200$	**0.0041**	0.1041	0.0830	0.1459	0.1869	0.1589	0.0283	**0.0971**	**0.0774**
α	0.5	$N = 50$	**-0.0107**	0.3419	0.2694	0.0329	0.3782	0.2965	-0.0279	**0.3075**	**0.2445**
		$N = 100$	**-0.0052**	0.2450	0.1949	0.0358	0.2680	0.2109	-0.0296	**0.2125**	**0.1714**
		$N = 200$	**-0.0069**	0.1801	0.1434	0.0296	0.2002	0.1584	-0.0314	**0.1610**	**0.1299**
b	-0.5	$N = 50$	**-0.0016**	0.1449	0.1156	-0.1495	0.2052	0.1689	-0.0028	**0.1340**	**0.1073**
		$N = 100$	**-0.0047**	0.0977	0.0777	-0.1489	0.1765	0.1532	-0.0048	**0.0928**	**0.0736**
		$N = 200$	**-0.0031**	0.0685	0.0555	-0.1476	0.1621	0.1481	-0.0045	**0.0654**	**0.0531**
β	0.5	$N = 50$	-0.0118	0.2283	0.1828	-0.0163	0.2171	0.1766	**-0.0096**	**0.2131**	**0.1714**
		$N = 100$	0.0073	0.1578	0.1257	**-0.0015**	0.1497	0.1203	0.0070	**0.1484**	**0.1177**
		$N = 200$	0.0042	0.1084	0.0864	-0.0054	0.1064	0.0842	**0.0022**	**0.1012**	**0.0801**
σ_1	0.7	$N = 50$	-0.1083	0.2935	0.2164	0.1867	0.2706	0.2205	**-0.0787**	**0.1859**	**0.1443**
		$N = 100$	**-0.0288**	0.1637	0.1226	0.2223	0.2606	0.2281	-0.0507	**0.1246**	**0.0984**
		$N = 200$	**-0.0180**	0.1063	0.0841	0.2248	0.2441	0.2253	-0.0447	**0.0912**	**0.0723**
σ_2	0.5	$N = 50$	-0.0532	0.1411	0.1034	**-0.0038**	**0.0960**	**0.0764**	0.0626	0.1059	0.0838
		$N = 100$	-0.0245	0.0797	0.0616	**0.0120**	**0.0685**	**0.0546**	0.0759	0.0961	0.0811
		$N = 200$	**-0.0173**	0.0554	0.0438	0.0210	**0.0526**	**0.0425**	0.0792	0.0897	0.0799
ϕ	6.4	$N = 50$	**0.0405**	0.9219	0.7313	0.2200	**0.8603**	**0.6580**	1.5476	1.8430	1.5695
		$N = 100$	**-0.0186**	0.6004	0.4772	0.1560	**0.5562**	**0.4419**	1.4776	1.6241	1.4802
		$N = 200$	**-0.0454**	0.4161	0.3334	0.1330	**0.4030**	**0.3204**	1.4547	1.5270	1.4547

inflation. This could be evidence that interpolation affects this component of the model much more than the other. In Figure 2.3, where the densities of the estimators obtained by the methods are shown, we see a behavior that confirms what was mentioned, being also clear the fact that the random components a and b are those that show a distribution much further from the theoretical values for the GHQ method with interpolation. Also, the variance components show greater deviation from the real value according to the density graph. For GAMLSS, its densities are comparable to those of SAEM in unbalanced datasets, with the exception of the σ_2 and ϕ parameters, which are specifically responsible for regulating the dispersion in the beta part of the ZIBR model.

Lastly, although a comparison with the results obtained on balanced data sets simulated under Setting 2 can not be directly established since the number of observations differs, notice that the estimation results obtained with SAEM in the two cases (Figure 2.3 here and Figure 2.2 in Section 2.3.1) are quite similar, whereas for GHQ there is a clear impact of interpolation on the estimation results.

Regarding the execution times of the routines, the same trends are confirmed that were already seen in the balanced data. On the same computer as Section 2.3.2, for $N = 50$ $(100, 200)$, on average, GAMLSS takes 1.07 $(1.38, 1.71)$ seconds, GHQ takes 1.85 $(3.66, 9.08)$ seconds, while SAEM takes 10.31 $(18.02, 32.15)$ seconds.

2.3.3 Hypothesis testing on covariates association

In the context of the ZIBR model, the effect of covariates on the presence or absence and abundance of a bacteria taxon must be studied. Therefore, we need procedures to test

the null hypothesis $H_0 : \alpha = \beta = 0$, $H_0 : \alpha = 0$ and $H_0 : \beta = 0$. A common approach in mixed models is to use the Likelihood Ratio Test (LRT) in this context, as a way of comparing two nested models. An alternative way to test both fixed and random effects parameters significance is to use the Wald test, for which standard errors estimates of each parameter are required.

Likelihood ratio test (LRT)

The Likelihood Ratio Test is performed to test the null hypothesis $H_0 : \alpha = \beta = 0$. We will now analyze its Type I error with the SAEM estimation method. As for parameter estimation, we are interested in the performance on both balanced and unbalanced data, and we will compare it with the results of the LRT based on the GHQ procedure, only in the balanced data scenario. However, as can be seen in Stasinopoulos et al. (2017), the calculation of the log-likelihood of the ZIBR model using GAMLSS is not the same as in the other two procedures, since it uses estimators based on a penalized quasi-likelihood method (Breslow and Clayton, 1993). For this reason, the values obtained for the log-likelihood in GAMLSS are not comparable with the other methods and the Likelihood Ratio Test cannot be applied in the same way. Therefore, we will not report it in this section.

The parameter values to simulate 1000 datasets are set as follows:

- $a = -0.5, b = 0.5$,
- $\alpha = \beta = 0$,
- $\sigma_1 = 0.7$, $\sigma_2 = 0.5$,
- $\phi = 6.4$.

The number of individuals in each dataset takes two values $N \in \{50, 100\}$, and we keep the number of observations per individual fixed $T_i = 10$. The procedure will be carried out on both 1000 balanced datasets and 1000 datasets from which 20% of their observations are dropped out following the MCAR process already described. The SAEM estimation settings were the same as above, while the log-likelihood estimation by Importance Sampling was carried out by simulating $K = 500$ values.

Table 2.3: *Type I error for testing $H_0 : \alpha = \beta = 0$ in balanced and unbalanced data with the SAEM algorithm and the GHQ procedure for nominal significance level of 0.05 and 0.01.*

		SAEM		**GHQ**	
		Significance level		**Significance level**	
Data type		0.05	0.01	0.05	0.01
Balanced	$N = 50$	0.050	0.007	0.059	0.009
	$N = 100$	0.048	0.009	0.050	0.012
Unbalanced	$N = 50$	0.064	0.009		
	$N = 100$	0.045	0.010		

The results shown in Table 2.3 make evident the similar behavior in balanced and unbalanced data with regard to the LRT procedure. In no case is there a significant difference between the expected and obtained values, and this behavior is not affected by the number of individuals or the type of data being worked on. Both SAEM and GHQ obtain good results in the balanced case, while in the unbalanced case SAEM approximates the theoretical level quite well.

Wald test for individual parameters

Finally, we will check the results of the Wald test using the standard errors of the estimators computed by stochastic approximation of the Fisher observed matrix. It is important to mention that the GHQ method implemented in the `ZIBR` package does not offer any method for calculating the standard errors of the estimates, so the results obtained using SAEM cannot be compared with those provided by the GHQ method. However, GAMLSS does produce estimates for the standard errors of the estimators, so comparisons can be made between SAEM and GAMLSS.

Fixed effects

The null hypotheses to be tested are $H_0 : \alpha = 0$ and $H_0 : \beta = 0$, for which the test statistics are defined by

$$\frac{\hat{\alpha}^2}{Var(\hat{\alpha})} \quad \text{and} \quad \frac{\hat{\beta}^2}{Var(\hat{\beta})}$$

where $Var(\hat{\alpha})$ and $Var(\hat{\beta})$ are estimated by the procedure described in Section 2.3 of the main document. Under the null hypothesis, these variables follow an asymptotic χ^2 distribution with one degree of freedom.

We keep the simulation settings used for the LRT. To improve the convergence properties of the SAEM algorithm, we use $m = 10$ Markov chains in the execution. The results are summarized in Table 2.4. We can see that the values obtained through simulation are not far from the theoretically expected values, although slightly above the ones obtained with the Likelihood Ratio Test. As shown in the table, SAEM achieves Type I errors close to the theoretical values of the test, while GAMLSS does not achieve optimal results. Furthermore, the inaccuracies of GAMLSS are more pronounced for testing $\beta = 0$ than testing $\alpha = 0$, which confirms the tendency of GAMLSS to produce poor results in the Beta component of the ZIBR model.

Table 2.4: *Type I error of the Wald test for $H_0 : \alpha = 0$ and $H_0 : \beta = 0$ using the SAEM algorithm for nominal significance level of 0.05 and 0.01.*

	SAEM		**GAMLSS**	
	Significance level		**Significance level**	
	0.05	0.01	0.05	0.01
$H_0 : \alpha = 0$	0.069	0.022	0.120	0.038
$H_0 : \beta = 0$	0.062	0.013	0.296	0.158

Random effects

The null hypotheses are now $H_0 : a = 0$ and $H_0 : b = 0$, and the corresponding test statistics

$$\frac{\hat{a}^2}{Var(\hat{a})} \quad \text{and} \quad \frac{\hat{b}^2}{Var(\hat{b})},$$

follow each a χ^2 distribution with one degree of freedom, if the null hypothesis are valid. The simulation settings now are those of Setting 2 with $a = 0$ and $b = 0$. As for the fixed-effects test, we use $m = 10$ Markov chains to accelerate convergence. The results of the test are presented in Table 2.5.

Table 2.5: *Type I error of the Wald test for $H_0 : a = 0$ and $H_0 : b = 0$ using the SAEM algorithm for nominal α-level of 0.05 and 0.01.*

	SAEM		**GAMLSS**	
	Significance level		**Significance level**	
	0.05	0.01	0.05	0.01
$H_0 : a = 0$	0.051	0.014	0.092	0.029
$H_0 : b = 0$	0.049	0.011	0.250	0.126

The results of this section, in the same way as those of the previous one, are indicative that the calculation of the standard errors given by SAEM meets the expected statistical properties, while GAMLSS does not obtain similar Type I errors. One of the advantages of this result is that it allows the development of hypothesis tests in a less computationally demanding way than the Likelihood Ratio Test, since this requires the estimation of two different models, while the Wald test only needs to estimate one.

2.4 Case studies

In this section we demonstrate the use of the proposed inference framework on two publicly available microbiome study and we will focus on the capabilities of the SAEM algorithm to detect changes in the presence of bacterial taxa in response to treatments. As we mentioned in the previous section, given that the ZIBR model has more than one random effect, GAMLSS uses an estimation method that can cause a bias in the estimation of the variance parameters, and also does not allow the evaluation of the tests based on the calculation of the log-likelihood that we will use in the following section. For this reason, the GAMLSS approach will not be applied to the real data considered below.

2.4.1 Inflammatory bowel disorder pediatric study

The data used in this section come from a study to verify the effectiveness of treatments in pediatric inflammatory bowel disorder (IBD) patients (Lewis et al., 2015). This study includes information from 90 children subjected to three different types of therapy: anti-TNF treatment (TNF: tumor necrosis factor), exclusive enteral nutrition (EEN) and

partial enteral nutrition with an ad lib diet (PEN). After filtering the data to discard low sequencing depth samples, low abundant genus and taxa with a proportion of zeros higher than 0.9 or lower than 0.1, the information from 18 bacterial genera measured at 4 different time points for each one of the 59 individuals (47 anti-TNF and 12 EEN) remained for the analysis. Figure 2.4 shows the average composition of the intestinal microbiome of the subjects in both groups and its evolution over time.

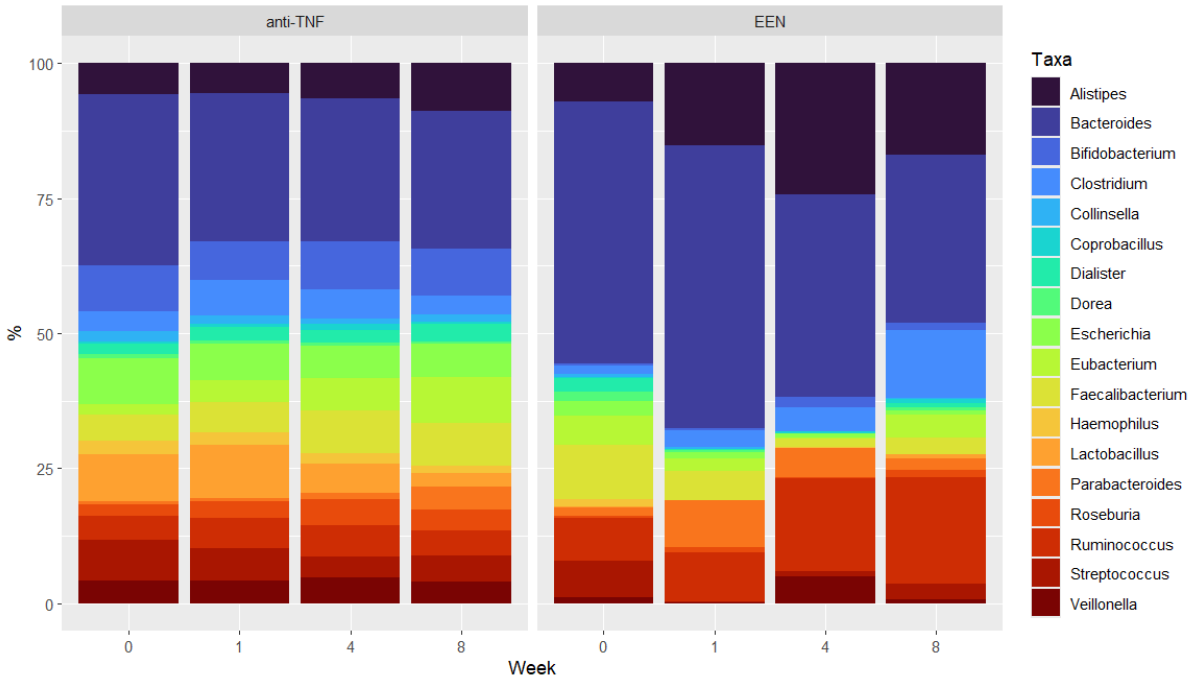


Figure 2.4: *Average gut microbiome composition of the treatment groups (anti-TNF and EEN) over observation week*

The objective of the study is to verify if the different treatments influence the presence of the different bacterial taxa in the samples, controlling for time and initial abundance. In addition, we want to compare if the results obtained by the SAEM algorithm differ from those obtained through the GHQ procedure implemented in the `ZIBR` package. The initial values for SAEM were the estimates found by the GHQ method, for each model corresponding to each bacterial taxon. Given this choice of initial values, we used $m = 5$ Markov chains and $(K_1, K_2) = (375, 125)$ iterations, as well as 500 simulated values for the log-likelihood calculation by Importance Sampling. The p-values obtained through the LRT were adjusted using the Benjamini-Hochberg process (Benjamini and Hochberg, 1995) to decrease the false discovery rate (FDR) and the full values are presented in Table A.1, Appendix A.

After model fitting, the GHQ method detected 11 bacterial taxa in which the treatment influenced the abundances, while SAEM managed to identify 12 taxa, all those identified by GHQ in addition to *Escherichia* with FDR=5%, as shown in Figure 2.5.

A more detailed analysis of the *Escherichia* data shows that the influence of treatment is greater on the frequency of presence of this bacterium in individuals than on the level of abundance once its presence is confirmed (Figure 2.6). This is confirmed by the LRT results for the significance of the treatment in the calculation of the probability p_{it} (FDR

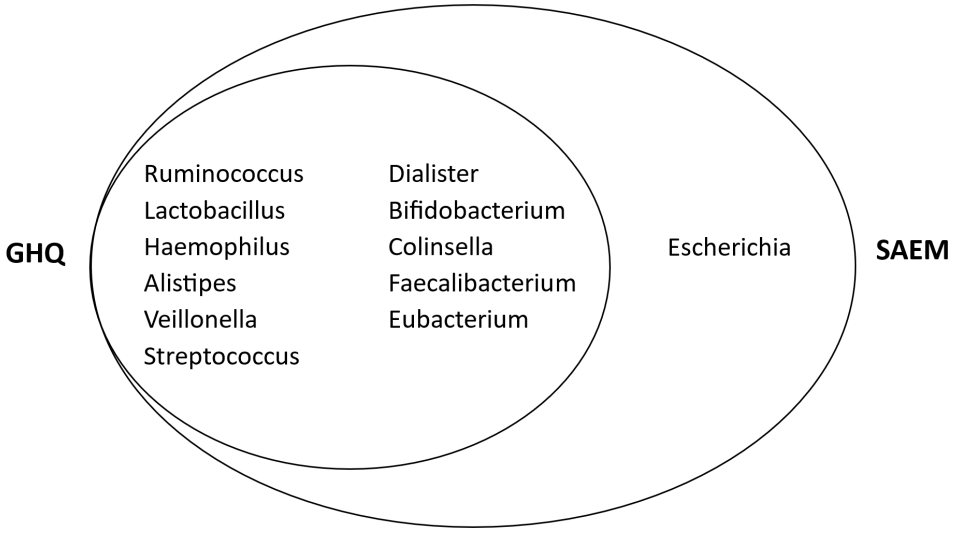


Figure 2.5: *Bacterial taxa in which the treatment (anti-TNF vs. EEN) have a statistical effect in abundance identified by SAEM and GHQ*

Table 2.6: *Estimated effects with the SAEM algorithm of the variables in the ZIBR model for Escherichia.*

Variable[1]	**Beta part**	**Logistic part**
Baseline	2.5521***	323.66***
	(0.3656)	(89.4662)
Time	-0.0356	0.1979***
	(0.0261)	(0.0412)
Treat	0.0746	3.0186***
	(0.1576)	(0.8815)

Note: Standard errors of the respective coefficients in parentheses. Symbol * (**,***) represents significance at 10% (5%, 1%) level.

[1] Statistical significance is calculated with the Wald test.

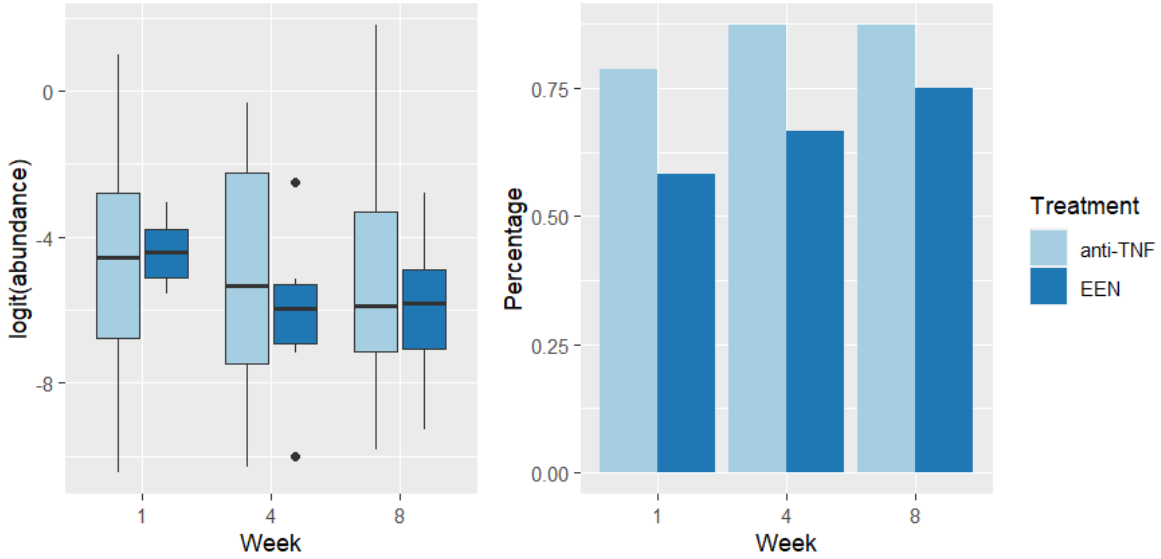


Figure 2.6: *Logit of the non-zero abundance (left) and percentage of samples with presence (right) for Escherichia in each treatment group (anti-TNF and EEN) across observation week.*

p-value 0.03) compared to those of the Beta component of the abundance u_{it} (FDR p-value 0.80), and by the Wald test (Table 2.6). These results show that at the 5% significance level the treatment is significant in the logistic part but not in the Beta part, proving that the definition of the ZIBR model and the combination with the SAEM estimation allows the increase of the ability to detect the influence of a given treatment defined. A detailed figure with the convergence behavior of the estimators across iterations is shown in Figure A.1, Appendix A.

The role of *Escherichia* in IBD is well documented. There is evidence (Baldelli et al., 2021) that the accumulation of *Escherichia coli* and other strains of *Escherichia* in the intestine is related to inflammatory processes, and other works (Mirsepasi-Lauridsen et al., 2019) suggest that a combination of antibiotic and dietary treatments is capable of controlling overproliferation of *E. coli* in the digestive system and also reducing the symptoms of IBD, allowing to infer a correlation between these two events.

2.4.2 Pregnancy effect in vaginal microbiome

In this case we apply the ZIBR model to the analysis of longitudinal data from a study (Romero et al., 2014) describing the vaginal microbiome of a group of 22 pregnant and 32 non-pregnant women. In this case we try to verify the effect of pregnancy on the distribution of the different bacterial taxa observed, in a similar way to what is done in a recent work (Zhang et al., 2020). However, unlike the cited study, where the analysis is performed on count data, we will analyse the data as proportions using the ZIBR model with SAEM estimation. Furthermore, a comparison of the results with the GHQ method can not be established since the number of time points is different between individuals; that is, the data is unbalanced.

A preliminary review of the data (Table 2.7) allows us to notice that there are large differences between the characteristics of pregnant and non-pregnant women. The time

Table 2.7: *Characteristics of the two groups of women, separated by pregnancy status*

Variable[1]	Non-pregnant $N = 32$	Pregnant $N = 22$
Age (years)	37 (31-43)	24 (20-29)
Time (months)	3.40 (3.52-3.67)	8.13 (8.00-8.43)
N. of observations	24 (21-29)	6 (6-7)

[1] Mean(Q1-Q3)

span of observations is much longer for pregnant women, and thus they have many more readings than non-pregnant women. In addition, the average age of pregnant women is much younger than the non-pregnant group. This is why we decided to include age as a covariate in the models to be used, according to the following specifications:

- **Model 1:** pregnancy, time and age as covariates, taking pregnancy as a factor of interest for testing.
- **Model 2:** pregnancy, time, age and interaction between time and pregnancy as covariates, testing the effect of pregnancy and the interaction.

Once we filter the bacterial taxa with a proportion of zeros between 0.1 and 0.9 and those that are absent in either of the two groups of women, we have 897 observations from 54 individuals and 57 taxa. With these data, we developed the LRT for the proposed variables in each model using SAEM at a significance level of 0.05.

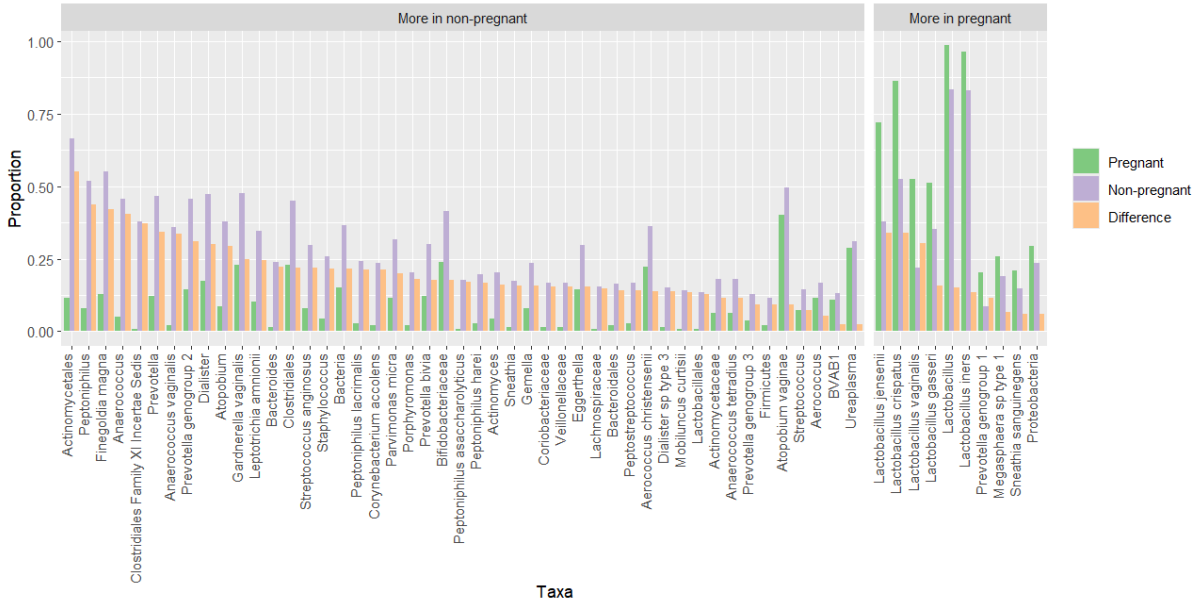


Figure 2.7: *Proportion of presence of the taxa in the observations of the two groups of women (pregnant and non-pregnant) and the difference between these values*

Figures 2.7 and 2.8 show the differences in presence of the taxa considered and the distribution of the non-zero abundance data. From these figures we can see that, on

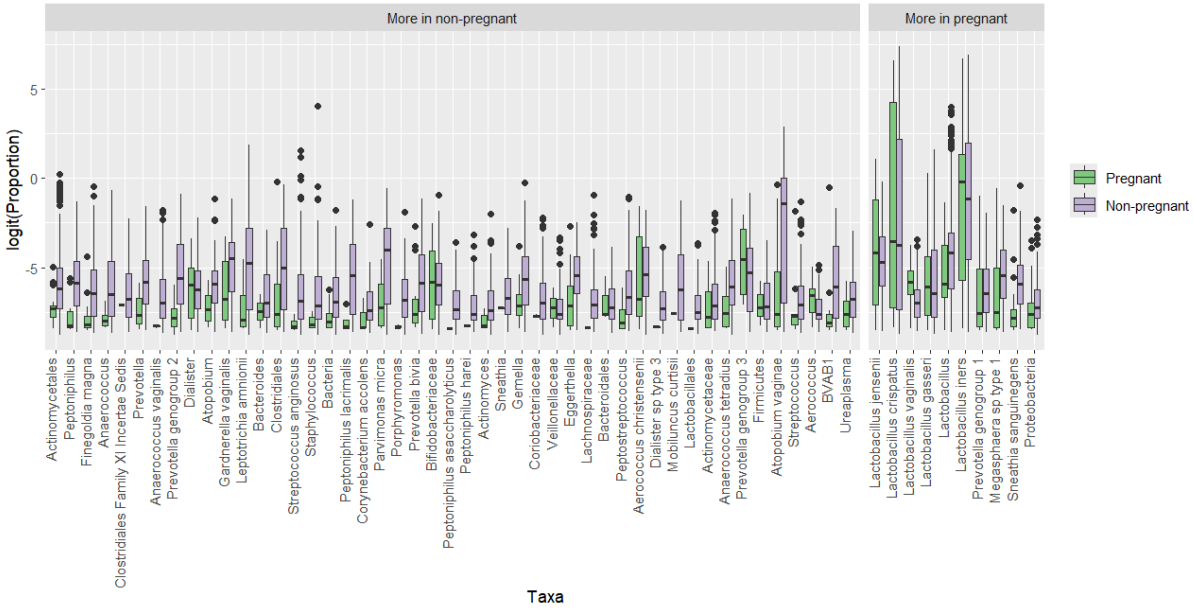


Figure 2.8: *Logit of the non-zero abundance of the taxa in the observations of the two groups of women (pregnant and non-pregnant)*

average, there are more bacteria with a higher abundance in non-pregnant women. In pregnant women, however, the dominance of the *Lactobacillus* genus in the most abundant bacteria is very evident, which is consistent with previous findings (Walther-António et al., 2014), which also suggest that low variety and high stability is another characteristic of the vaginal microbiome in pregnant women.

The LRT indicates that Model 1 was able to detect a higher number of bacteria (51% of all taxa) affected by pregnancy than Model 2 (16%), for which the interaction of time with pregnancy is statistically significant for a higher number of bacteria (26%) compared to pregnancy alone. Figure 2.9a details these results further. In particular, for Model 1 the bacteria for which an influence of pregnancy is detected are more commonly found among those that are more abundant in non-pregnant women. In comparison, only 2 of the most present bacteria in pregnant women show a statistical significance of pregnancy: *Lactobacillus crispatus* and *Sneathia sanguinegens*. The information in Table A.2 also confirms these findings, showing that the coefficients associated with pregnancy for these taxa in the abundance part have different sign. In a previous work (Romero et al., 2014) it is found that bacteria of the genus *Sneathia*, potentially pathogenic, reduce their presence during pregnancy, while *Lactobacillus crispatus* abundance in pregnancy is associated with a lower risk of preterm delivery (Veščičík et al., 2020).

However, Figure 2.9b shows that adding the time-pregnancy interaction to the model specification changes the results. First, only a small number of the bacteria that predominate in non-pregnant women show significant dependence on pregnancy or the interaction between time and pregnancy. But in the other group of bacteria, nine show significance for pregnancy or the interaction, and three of them for both: *Lactobacillus jensenii*, *Prevotella genogroup 1* and *Megasphaera sp type 1*. Previous studies (Severgnini et al., 2022) report that both *Lactobacillus* bacteria and those associated with bacterial vaginosis (*Prevotella*, *Sneathia*) change their abundance between pregnant and non-pregnant women and also

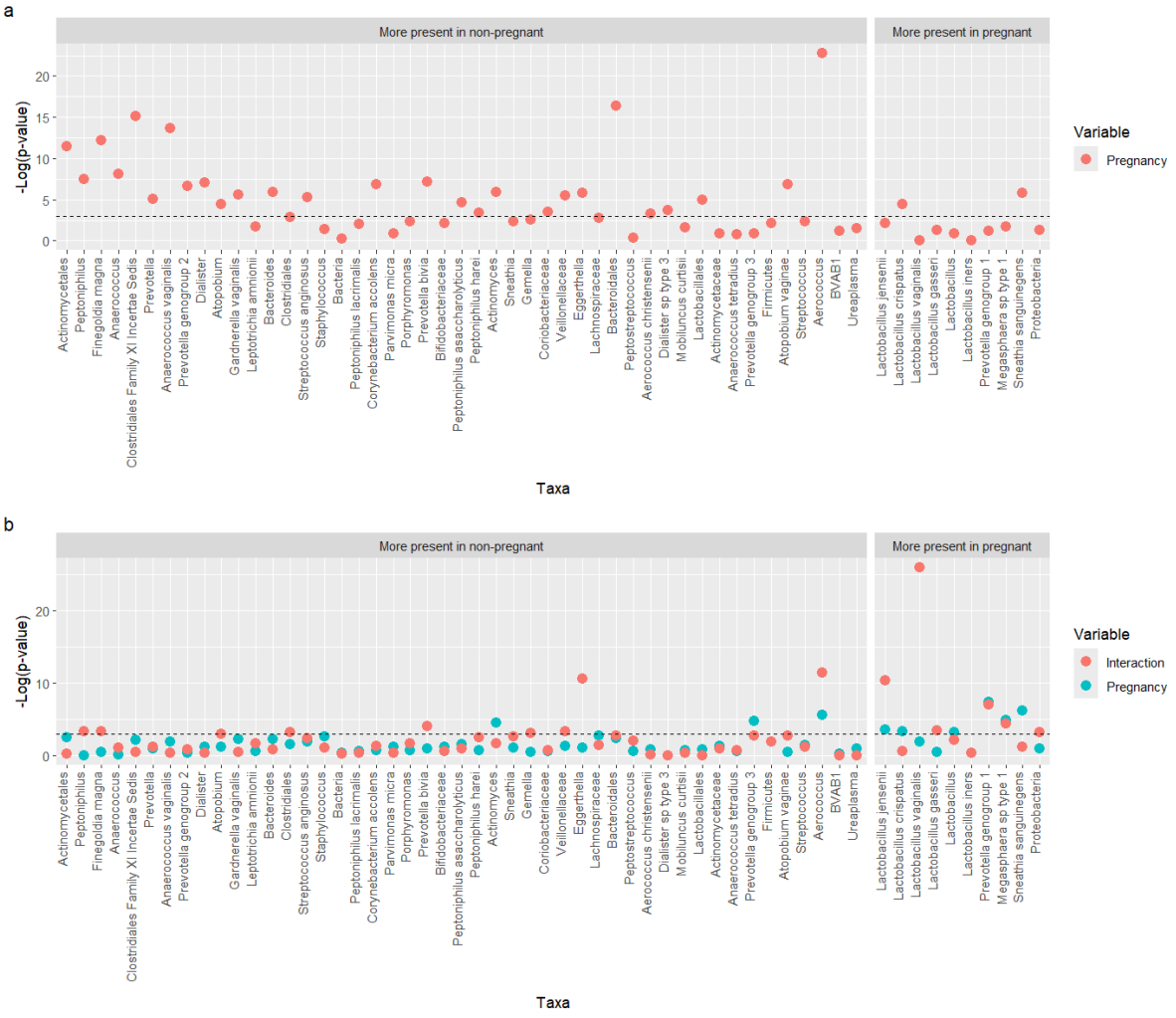


Figure 2.9: *Negative of log transformed p-value of the LRT for the interest variables in Model 1 (a) and Model 2 (b) for the bacterial taxa. The horizontal line represents the threshold* $\alpha = 0.05$

along time in case of pregnancy. The coefficients associated with the variables can be consulted in Table A.3. In view of these results, we can assert that the ZIBR model and the SAEM estimation for relative abundance data obtain similar conclusions as both previous research results and the mixed models defined for log-transformed count data.

2.5 Conclusions

In this chapter we have developed an exact maximum likelihood estimation strategy for the ZIBR model for the analysis of longitudinal compositional microbiome data using the SAEM algorithm. We have also proposed a method for calculating the log-likelihood of the model which allows to obtain information criteria for the model, and approximations of the estimators standard errors, which is not possible under the alternative estimation method based on Gauss-Hermite quadrature likelihood approximation. Moreover, despite the capacity of the GAMLSS approach to estimate the ZIBR model parameters and calculate the standard errors of these estimates, the SAEM method exhibits superior performance controlling the Type I error of the Wald test for the significance of parameters. Although the results obtained by SAEM conform to the expected theoretical properties of standard errors, it must be noted that the method can still be improved, since, as it depends on a stochastic approximation, convergence towards coherent values for the Fisher information matrix is not guaranteed, which could introduce bias in the estimation of the standard errors. We are confident that these details can be enhanced in further developments.

Another advantage of the proposed estimation method is its ability on handling unbalanced data, a scenario that can occur both due to the design of the experiment itself and due to external factors, such as to individuals dropping out during the study. This aspect was not considered in the development of the original estimation method for ZIBR, so comparisons of performance with our method cannot be established in this scenario, unless data interpolation is performed before using the GHQ method. The GAMLSS method, another option that allows for handling unbalanced data, has been found to show errors in the estimation of certain important parameters of the model, even when its performance in the rest is adequate. It should be emphasized that unbalanced data is a fairly common situation in medical experiments, in which multiple factors influence patients abandoning the follow-up. This could be one of the reasons contributing to the high non-publication rate in many medical studies, which according to certain sources could be close to 50% (Chan et al., 2014). Therefore, developing analysis methods that can deal accurately with unbalanced data is of great interest.

The definition of the ZIBR model used throughout this work corresponds to the one originally proposed by Chen and Li (2016) and implemented in the `ZIBR` package for R statistical software. However, there are possibilities for modification of this definition that have been discussed. One of them is the use of random effects for more than one covariate, an aspect that has been already incorporated in the implementation used in this article. Another possibility is the inclusion of cross correlations in the random effects, proposing a different structure in the variance of these effects. Liu et al. (2019) mention that this inclusion could alter the results for tests on covariates, detecting significance where a simpler structure would not detect it. Although this approach has not been implemented here, the SAEM algorithm could be easily modified to serve this purpose.

In the field of human microbiome analysis, other models have been proposed in addition to ZIBR. Among the most important are ZIBR-SRE (Han et al., 2021), an extension of ZIBR which considers the compositional nature of microbiota data; zero-inflated Gaussian mixed models (ZIGMM) (Zhang et al., 2020), which in addition to managing the overabundance of zeros can work with both proportion data and counts; and the negative binomial mixed model (NBMM) (Zhang et al., 2018), which allows the specification of more general variance structures and also be modified to deal with zero inflation. It seems interesting to implement the SAEM algorithm to these models and study its potential benefits in estimation.

Finally, an interesting extension of this work would be to obtain the Restricted Maximum Likelihood (REML) estimates, a known method for reducing bias in the estimation of variance components in mixed effects models (Meza et al., 2007), using the Harville's approach, i.e., integrating out the fixed effects, via the SAEM algorithm. We expect that, in the context of longitudinal models on microbiome data, this could improve the results obtained through ML estimation.

CHAPTER 3

ZERO INFLATED BETA-BINOMIAL MIXED REGRESSION (ZIBBMR) FOR LONGITUDINAL COUNT DATA AND A SAEM-BASED ESTIMATION*

3.1 Introduction

The study of the human microbiota has generated a substantial amount of information in the form of sequence counts, which represent the relative abundance of different taxa present in a sample. These counts, derived from technologies such as 16S rRNA gene sequencing or shotgun metagenomics, serve as the foundation for numerous research endeavors in microbial ecology, human health, and biotechnology. However, the discrete, non-negative, and overdispersed nature of these data poses significant statistical challenges for their analysis and modeling.

One of the earliest approaches involves the treatment of data as realizations of count-type random variables. The Poisson model, in which the variance is equivalent to the mean, is the simplest model for this type of data. The Poisson distribution is frequently employed in circumstances characterized by low dispersion; however, it often falls short in adequately capturing the variability present in microbiota data, which frequently manifest as overdispersion (Love et al., 2014). In such cases, a more flexible alternative is the negative binomial distribution, which introduces an additional parameter to model variance, allowing it to be greater than the mean (Cameron and Trivedi, 2013). However, when the data originate from proportions of counts (e.g., number of readings of a species given a total number of readings per sample), it may be more appropriate to consider binomial models or related distributions.

In this context, the beta-binomial distribution (Skellam, 1948) is a particularly useful extension, as it posits that the probability of success in each binomial trial is a random

*This chapter corresponds to an article which is a joint work with Cristian Meza, Dae-Jin Lee and Ana Arribas-Gil under development to be submitted for review in a scientific journal.

variable with a beta distribution. This model introduces correlation between trials, allowing for the capture of additional overdispersion with respect to the traditional binomial model (Williams, 1975). The beta-binomial distribution has been employed in a variety of contexts, including those where data exhibit a natural hierarchical structure or where proportional rates with variability between experimental units need to be modeled.

With regard to practical applications, models based on the beta-binomial distribution have been employed in a wide variety of fields. In the domain of health services research, beta-binomial regression models have been utilized to evaluate the appropriateness of hospital stays, accounting for within-cluster homogeneity and policy modifications (Gange et al., 1996). A distribution directly related to the beta-binomial model, the log-Lindley binomial distribution, has been proposed for modeling clustered binary outcomes in developmental toxicology experiments (Razzaghi, 2022). In the context of biometric identification, the beta-binomial distribution has been employed to estimate matching performance and assess variability in false match and false non-match rates. This approach accounts for extraneous variability and facilitates the creation of confidence intervals (Schuckers, 2003). In the context of microbiota data analysis, Hu et al. (2018) proposed a model based on the beta-binomial distribution, incorporating a term designed to address the zero-inflation frequently observed in microbiota data. However, this model did not consider the possibility of repeated data from the same individual. A comparable approach was adopted by Luo and Paul (2018), who focused on estimation under missing data using the EM algorithm (Dempster et al., 1977). On the other hand, Najera-Zuloaga et al. (2019) and Wu et al. (2017) defined models that incorporated random effects into the beta-binomial model but did not incorporate a specific component for inflation in zeros.

In this chapter, motivated by the study of microbiota as sequence counts, we propose a novel model derived from the beta-binomial distribution, which was developed for the specific purpose and characteristics of human microbiota data. The model incorporates a component for zero-inflation, as well as individual random effects that reflect the autocorrelation inherent in measurements from the same individual, and also describes the overdispersion that escapes traditional models, such as the Poisson distribution. In addition to defining the model, we implement methods for estimating its parameters and for statistical inference of the significance of covariates. These methods are based on the SAEM algorithm (Delyon et al., 1999), which yielded promising results in the previous chapter. The efficacy of this method is evaluated through a comparative analysis with other existing alternatives, employing both artificially generated datasets and a study with real data.

3.2 Motivating data

The human vaginal microbiota is a complex ecosystem that plays a crucial role in women's reproductive health. The composition and stability of the substance during pregnancy are of particular significance, as they have the capacity to exert a substantial influence on perinatal outcomes. These outcomes may include preterm birth and the development of infections. The application of 16S rRNA gene sequencing techniques has facilitated a more profound characterization of bacterial diversity and its temporal dynamics, thereby circumventing the constraints imposed by conventional culture methodologies.

In the study by Romero et al. (2014), the vaginal microbiota of 22 healthy pregnant women who underwent full-term deliveries was longitudinally compared with that of 32 non-pregnant women. Utilizing an extensive sequencing approach and statistical models appropriate for longitudinal count data, it was determined that the vaginal microbiota during pregnancy is distinguished by enhanced stability and a heightened abundance of species belonging to the genus *Lactobacillus* (including *L. crispatus*, *L. gasseri*, *L. vaginalis*, and *L. jensenii*). This is in contrast to the increased diversity and presence of microorganisms associated with dysbiosis observed in non-pregnant women, such as *Gardnerella*, *Atopobium*, and *Prevotella*.

The research mentioned above was a pioneering study in the longitudinal analysis of the vaginal microbiome in pregnant women. Prior to this study, the majority of research focused on the vaginal microbiome of young, non-pregnant women of reproductive age with no known diseases (Srinivasan et al., 2010; Ravel et al., 2011). Moreover, the data collection methodologies were predominantly culture-based, as opposed to the in situ sequencing of samples. Consequently, no results were available concerning the stability of the vaginal microbiota composition over time. In light of the aforementioned points, this study has been cited in numerous other works (Florova et al., 2021; Kroon et al., 2018; Zhu et al., 2022) due to the novelty of its methodology and conclusions. A recent study by the same author (Romero et al., 2023) built upon the findings of the previous work, yet the original data are publicly available in statistical software packages (Yi, 2024).

Table 3.1: *Demographic description of the pregnant and non pregnant women from the study of Romero et al. (2014).*

Characteristic[1]	**Non-pregnant** N = 32	**Pregnant** N = 22
Age	37(31-43)	24(20-29)
Ethnic origin		
Black	16 (50%)	19 (86%)
Hispanic, others	3 (9.4%)	0 (0%)
White	13 (41%)	3 (14%)
Nugent score[2]		
< 7	16 (50%)	19 (86%)
≥ 7	16 (50%)	3 (14%)

[1] Age: mean(Q1-Q3); ethnic group and Nugent score: n (%).

[2] If at least one sample has a Nugent score over 7. According to Nugent et al. (1991), this corresponds to a diagnosis of bacterial vaginosis.

However, despite the numerous advantages of the study, it is important to note that it is not without its limitations. A notable limitation is evident in Table 3.1, which indicates that the majority of the subjects were of African American descent. As previously stated in Serrano et al. (2019), the composition of the vaginal microbiota is closely linked to ethnic origin due to genetic factors. This, in conjunction with the limited sample size employed in the study, precludes the generalization of the findings. In a separate context, mixed models were employed to ascertain the impact of pregnancy on the prevalence of each bacterial taxon. However, these models were constrained to include only pregnancy as a covariate, thereby excluding other potentially influential variables, such as age. In

addition, for those taxa that had a large number of null observations, zero-inflated models that did not have a random effect for this component were used, using only fixed effects.

With the data mentioned above, in Zhang et al. (2020) several mixed model specifications are used to detect the influence of pregnancy on the composition of the vaginal microbiota. To be precise, linear and negative binomial models without zero-inflation are used, in addition to a Gaussian model that includes zero-inflation defined for log-transformed data. This transformation is one of the main weaknesses of the design, since according to McMurdie and Holmes (2014), it will always be preferable to use the data as recorded (in this case sequence counts). On the other hand, although it includes demographic covariates as controls in the specifications, it only verifies the influence of pregnancy on abundance using the Wald test, ruling out the possible influence on inflation at zero. Therefore, with the model proposed below, we propose to verify the influence of pregnancy on both non-zero abundance and zero inflation, controlling for demographic covariates. We will also improve the inference method, using the likelihood ratio test, and compare it with the results obtained in Section 2.4.2 of the previous chapter to verify whether there are substantial differences between our new model and the ZIBR model (Chen and Li, 2016), which uses proportion data.

3.3 Model definition, estimation and inference

3.3.1 Definition of the Zero Inflated Beta-Binomial Mixed Regression (ZIBBMR)

The ZIBBMR model shares notable parallels with the ZIBR model, drawing from its framework to delineate its components, as will be subsequently discussed. Let Y_{it} be the count of sequences of a bacterial taxon for an individual i at time t, $1 \leq i \leq N$, $1 \leq t \leq T_i$. The model assumes that Y_{it} follows the distribution:

$$Y_{it} \sim \begin{cases} 0 & \text{with probability } 1 - p_{it}, \\ \text{BetaBin}(S_{it}, u_{it}\phi, (1-u_{it})\phi) & \text{with probability } p_{it}. \end{cases} \tag{3.1}$$

with $\phi > 0$ and $0 < u_{it}, p_{it} < 1$. These two last components are characterized by

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = a_i + X_{it}^T\alpha, \qquad \log\left(\frac{u_{it}}{1-u_{it}}\right) = b_i + Z_{it}^T\beta, \tag{3.2}$$

where a_i and b_i are individual-specific intercepts, α and β are vectors of regression coefficients and X_{it} and Z_{it} are covariates for each individual and time point. We further consider that each one of the random intercepts follows a normal distribution, independently from each other:

$$a_i \sim N(a, \sigma_1^2), \quad b_i \sim N(b, \sigma_2^2).$$

The mass function for the beta-binomial distribution is defined by:

$$P\left(Y = Y_{it} | S_{it}, u_{it}, \phi\right) = f(Y_{it}; S_{it}, u_{it}, \phi) = \binom{S_{it}}{Y_{it}} \frac{B\left(Y_{it} + u_{it}\phi, S_{it} - Y_{it} + (1-u_{it})\phi\right)}{B\left(u_{it}\phi, (1-u_{it})\phi\right)}, \tag{3.3}$$

where $B(\cdot,\cdot)$ is the beta function and S_{it} represents the total number of trials, which in the case of microbiota data can be identified as the total number of sequences for individual i at time t. From this definition, it is easy to calculate that $E\left(Y_{it}|S_{it},u_{it},\phi\right)=S_{it}u_{it}$ and $Var\left(Y_{it}|S_{it},u_{it},\phi\right)=S_{it}u_{it}(1-u_{it})\left[1+\frac{S_{it}-1}{\phi+1}\right]$. With this, we can see that the variance of a beta-binomial variable contains a factor that cannot be explained in the case of a classical binomial distribution, and if $S_{it}\gg\phi$ (which is the case almost surely) makes this variance bigger than the binomial distribution with fixed success rates. This makes it suitable for modeling data with overdispersion.

An equivalent way to express the ZIBBMR model is by using a hierarchical structure. This is due to the fact that the beta-binomial distribution can be expressed as a binomial distribution $Bin(n,p)$ where p follows a beta distribution. Using this, ZIBBMR can be defined as

$$\begin{aligned} Y_{it}|S_{it},p_{it},w_{it},\phi &\sim \begin{cases} 0 & \text{with probability } 1-p_{it}, \\ \text{Bin}(S_{it},w_{it}) & \text{with probability } p_{it}. \end{cases} \\ w_{it}|u_{it},\phi &\sim \text{Beta}(u_{it},\phi) \\ \text{logit}(p_{it}) &\sim N\left(a+X_{it}^{T}\alpha,\sigma_1^2\right) \\ \text{logit}(u_{it}) &\sim N\left(b+Z_{it}^{T}\beta,\sigma_2^2\right) \end{aligned} \tag{3.4}$$

where the beta distribution has the density defined in Equation 2.4. As can be seen, this expression introduces a latent term w_{it} that is not observable, which can be used in estimation tasks as part of a novel calculation scheme, which we will discuss later.

Regardless of how the model is expressed, the maximum likelihood estimate of $\theta=(\phi,a,b,\alpha,\beta,\sigma_1^2,\sigma_2^2)$ is performed in the same way. From Equations 3.1 and 3.2, the likelihood function for data $\boldsymbol{Y}=(Y_{it},1\leq i\leq N,1\leq t\leq T_i)$ is

$$L(\theta;\boldsymbol{Y})=\prod_{i=1}^{N}\int_{\mathbb{R}}\int_{\mathbb{R}}\prod_{t=1}^{T_i}(1-p_{it})^{\mathbb{1}_{\{Y_{it}=0\}}}\left[p_{it}f(Y_{it};u_{it},\phi)\right]^{\mathbb{1}_{\{Y_{it}>0\}}}g\left(a_i,b_i|a,\sigma_1^2,b,\sigma_2^2\right)da_idb_i \tag{3.5}$$

where $f(Y_{it};u_{it},\phi)$ is the beta-binomial mass function defined in Equation 3.3 and g is the product of the two univariate normal density functions of random effects a_i and b_i. This integral is analytically impossible to calculate, so alternative methods must be used for estimation. In addition to the SAEM algorithm, which we expose in more detail in the following section, there are other routines based on approximating the definite integral in 3.5. One of these is the `glmmTMB` package (Brooks et al., 2017), which is based on approximating the integral using the Laplace technique and its numerical optimization. There is also the `gamlss` package (Rigby and Stasinopoulos, 2005), which we discussed in detail in Section 2.2.1 and which, although it allows for penalized maximum likelihood estimation, does not provide consistent log-likelihood values for inference procedures.

3.3.2 Estimation of ZIBBMR parameters using SAEM

Let us consider $\varphi_i=(a_i,b_i)$, $1\leq i\leq N$, the non-observed data. By the definition of the ZIBBMR model φ_i follows the multivariate normal distribution $\varphi_i\sim N(\boldsymbol{\mu},\mathbf{G})$ with $\boldsymbol{\mu}=$

(a, b) and $\mathbf{G} = diag(\sigma_1^2, \sigma_2^2)$. With the usual notation $\boldsymbol{Y} = (Y_{it} : \ 1 \leq i \leq N, 1 \leq t \leq T_i)$ and $\boldsymbol{\varphi} = (\varphi_i : \ 1 \leq i \leq N)$, the complete-data likelihood writes:

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{\varphi}; \theta) =& p(\boldsymbol{Y}|\boldsymbol{\varphi}; \alpha, \beta, \phi) p(\boldsymbol{\varphi}|\boldsymbol{\mu}, \mathbf{G}) \\ \propto& |\mathbf{G}|^{-\frac{N}{2}} \prod_i \exp\left(-\frac{(\varphi_i - \boldsymbol{\mu})^T \mathbf{G}^{-1} (\varphi_i - \boldsymbol{\mu})}{2}\right) \\ & \times \prod_{i,t} (1 - p_{it})^{\mathbb{1}_{\{Y_{it}=0\}}} p_{it}^{\mathbb{1}_{\{Y_{it}>0\}}} f(Y_{it}; S_{it}, u_{it}, \phi)^{\mathbb{1}_{\{Y_{it}\geq 0\}}}. \end{aligned} \tag{3.6}$$

Then, similar to what was explained in Section 2.2.2, the SAEM algorithm estimate for the ZIBBMR model follows the following scheme, for a given starting point $\theta^{(0)}$ and at iteration q:

1. **Simulation step:** draw $\varphi_i^{(q)}, i = 1, \cdots, N$ from the distribution $p(\cdot|\boldsymbol{Y}; \theta^{(q-1)})$.

2. **Stochastic Approximation step:** update the summary data functions $F_1(\boldsymbol{Y}, \boldsymbol{\varphi})$ and $F_2(\boldsymbol{Y}, \boldsymbol{\varphi})$ with the scheme:

$$\begin{aligned} F_1^{(q)}(\boldsymbol{Y}, \boldsymbol{\varphi}) &= F_1^{(q-1)}(\boldsymbol{Y}, \boldsymbol{\varphi}) + \gamma_q \left(\sum_i \varphi_i^{(q)} - F_1^{(q-1)}(\boldsymbol{Y}, \boldsymbol{\varphi})\right) \\ F_2^{(q)}(\boldsymbol{Y}, \boldsymbol{\varphi}) &= F_2^{(q-1)}(\boldsymbol{Y}, \boldsymbol{\varphi}) + \gamma_q \left(\sum_i \varphi_i^{(q)} \varphi_i^{(q)^T} - F_2^{(q-1)}(\boldsymbol{Y}, \boldsymbol{\varphi})\right). \end{aligned} \tag{3.7}$$

   where $\{\gamma_q\}_{q \in \mathbb{N}}$ is a decreasing sequence of stepsizes with $\gamma_1 = 1$.

3. **Maximization step:** update the parameters of the model with

$$\begin{aligned} \boldsymbol{\mu}^{(q)} &= \frac{F_1^{(q)}(\boldsymbol{Y}, \boldsymbol{\varphi})}{N} \\ \mathbf{G}^{(q)} &= \frac{F_2^{(q)}(\boldsymbol{Y}, \boldsymbol{\varphi})}{N} - \frac{F_1^{(q)}(\boldsymbol{Y}, \boldsymbol{\varphi}) F_1^{(q)}(\boldsymbol{y}, \boldsymbol{\varphi})^T}{N^2} \end{aligned} \tag{3.8}$$

   Given the form of the model definition in the beta-binomial part, steps 2 and 3 are modified by first calculating

$$\left(\tilde{\beta}^{(q)}, \tilde{\phi}^{(q)}\right) = \arg\max_{\beta, \phi} \sum_{i,t} \left[\mathbb{1}_{\{Y_{it}>0\}} \left(\log \frac{B\left(Y_{it} + u_{it}^{(q)}\phi, S_{it} - Y_{it} + (1 - u_{it}^{(q)})\phi\right)}{B\left(u_{it}^{(q)}\phi, (1 - u_{it}^{(q)})\phi\right)}\right)\right] \tag{3.9}$$

   and

$$\tilde{\alpha}^{(q)} = \arg\max_{\alpha} \sum_{i,t} \left[\mathbb{1}_{\{Y_{it}>0\}} \log\left(p_{it}^{(q)}\right) + \mathbb{1}_{\{Y_{it}=0\}} \log\left(1 - p_{it}^{(q)}\right)\right]. \tag{3.10}$$

   where $u_{it}^{(q)} = u_{it}^{(q)}(b_i, \beta)$ and $p_{it}^{(q)} = p_{it}^{(q)}(a_i, \alpha)$ are calculated using $\varphi_i^{(q)}$ and Equation 3.2. Maximization in (3.9) and (3.10) is achieved numerically. Finally, the values

are updated by doing

$$\begin{aligned}\phi^{(q)} &= \phi^{(q-1)} + \gamma_q\left(\tilde{\phi}^{(q)} - \phi^{(q-1)}\right)\\ \alpha^{(q)} &= \alpha^{(q-1)} + \gamma_q\left(\tilde{\alpha}^{(q)} - \alpha^{(q-1)}\right)\\ \beta^{(q)} &= \beta^{(q-1)} + \gamma_q\left(\tilde{\beta}^{(q)} - \beta^{(q-1)}\right)\end{aligned} \tag{3.11}$$

A closer look to the Simulation step

At this point, we have a clear outline of the structure followed by the SAEM algorithm for estimating parameters of the ZIBBMR model by maximum likelihood. However, if we use the hierarchical structure defined in Equation 3.4, an alternative simulation method can be proposed. In particular, if we consider the unobserved data $\boldsymbol{w} = (w_{it} : 1 \leq i \leq N, 1 \leq t \leq T_i)$ in the hierarchical definition of the model, Equation 3.6 becomes:

$$\begin{aligned}p(\boldsymbol{Y}, \boldsymbol{w}, \boldsymbol{\varphi}; \theta) =& p(\boldsymbol{Y}|\boldsymbol{w}, \boldsymbol{\varphi}; \alpha)p(\boldsymbol{w}|\boldsymbol{\varphi}; \beta, \phi)p(\boldsymbol{\varphi}|\boldsymbol{\mu}, \mathbf{G})\\ \propto& \prod_{i,t}(1-p_{it})^{\mathbb{1}_{\{Y_{it}=0\}}} p_{it}^{\mathbb{1}_{\{Y_{it}>0\}}} f_1(Y_{it}; S_{it}, w_{it})^{\mathbb{1}_{\{Y_{it}>0\}}}\\ &\times \prod_{i,t} f_2\left(w_{it}; u_{it}, \phi\right)^{\mathbb{1}_{\{Y_{it}>0\}}}\\ &\times |\mathbf{G}|^{-\frac{N}{2}} \prod_i \exp\left(-\frac{(\varphi_i - \boldsymbol{\mu})^T \mathbf{G^{-1}} (\varphi_i - \boldsymbol{\mu})}{2}\right).\end{aligned} \tag{3.12}$$

where f_1 is the binomial mass function and f_2 is the beta density function defined in Equation 2.4. Therefore, and now denoting $\boldsymbol{z} = (\boldsymbol{\varphi}, \boldsymbol{w})$ as the non-observed data under this specification, we have that the complete data will be $(\boldsymbol{Y}, \boldsymbol{z})$ and the Simulation step of the SAEM algorithm must be modified as follows:

1'. **Simulation step:** draw $\boldsymbol{z}^{(q)}$ from the distribution $p(\cdot|\boldsymbol{Y}; \theta^{(q-1)})$.

For the original specification, the Simulation step required knowledge of the conditional distribution $\boldsymbol{\varphi}|\boldsymbol{Y}; \theta$, which is approximated using the Metropolis-Hastings algorithm (Metropolis et al., 1953), implemented using m iterations at each step q of the SAEM algorithm as follows, for $i = 1, \ldots, N$:

1. Set $\varphi_{i,0} = \varphi_i^{(q-1)} = (a_i^{(q-1)}, b_i^{(q-1)})$.
2. For $p = 1, \ldots, m$:
   - select $\hat{\varphi}_{i,p}$ from a defined kernel $Q_{\theta^{(q-1)}}(\varphi_{i,p-1}, \cdot)$ and $d_{i,p} \sim U[0,1]$,
   - calculate $\Delta_{i,p}$ defined by

$$\Delta_{i,p} = \log\left(\frac{p(\hat{\varphi}_{i,p}|Y_i; \theta^{(q-1)})Q_{\theta^{(q-1)}}(\hat{\varphi}_{i,p}, \varphi_{i,p-1})}{p(\varphi_{i,p-1}|Y_i; \theta^{(q-1)})Q_{\theta^{(q-1)}}(\varphi_{i,p-1}, \hat{\varphi}_{i,p})}\right),$$

   - set

$$\varphi_{i,p} = \begin{cases}\hat{\varphi}_{i,p} & \text{if } \Delta_{i,p} \leq \log(d_{i,p}),\\ \varphi_{i,p-1} & \text{otherwise.}\end{cases}$$

3. Finally, define $\varphi_i^{(q)} = \varphi_{i,m}$.

In the current case, three kernels are proposed for implementing the algorithm:

1. the prior distribution $N(\boldsymbol{\mu}^{(q)}, \mathbf{G}^{(q)})$,
2. the multidimensional random walk $N(\varphi_{i,p-1}, \Omega^{(q)})$ with $\Omega^{(q)}$ a positive definite matrix adjusted to produce an optimal acceptance rate,
3. and the one-dimensional random walk by random components of $\varphi_{i,p-1}$; that is, only one element picked at random is altered at a time with a noise $N(0,1)$.

In the first case, we have

$$\Delta_{i,p}^{(1)} = \sum_{t=1}^{T_i} \mathbb{1}_{\{Y_{it}>0\}} \left(\log \frac{B\left(Y_{it} + \hat{u}_{it}\phi^{(q)}, S_{it} - Y_{it} + (1-\hat{u}_{it})\phi^{(q)}\right)}{B\left(Y_{it} + u_{it}\phi^{(q)}, S_{it} - Y_{it} + (1-u_{it})\phi^{(q)}\right)} - \log \frac{B\left(\hat{u}_{it}\phi^{(q)}, (1-\hat{u}_{it})\phi^{(q)}\right)}{B\left(u_{it}\phi^{(q)}, (1-u_{it})\phi^{(q)}\right)} \right)$$
$$+ \mathbb{1}_{\{Y_{it}>0\}} \log \frac{\hat{p}_{it}}{p_{it}} + \mathbb{1}_{\{Y_{it}=0\}} \log \frac{1-\hat{p}_{it}}{1-p_{it}},$$

and in the other two kernels, since they are symmetric kernels, we have

$$\Delta_{i,p}^{(2)} = \Delta_{i,p}^{(1)} + \frac{1}{2}\left((\varphi_{i,p-1} - \boldsymbol{\mu}^{(q)})^\top {\mathbf{G}^{(q)}}^{-1} (\varphi_{i,p-1} - \boldsymbol{\mu}^{(q)}) - (\hat{\varphi}_{i,p} - \boldsymbol{\mu}^{(q)})^\top {\mathbf{G}^{(q)}}^{-1} (\hat{\varphi}_{i,p} - \boldsymbol{\mu}^{(q)} \right)$$

where $\hat{u}_{it}$ and u_{it} are calculated using $\hat{\alpha}_{i,p}$ and $\alpha_{i,p-1}$ respectively. Similarly, $\hat{p}_{it}$ and p_{it} are obtained given $\hat{\beta}_{i,p}$ and $\beta_{i,p-1}$. In this way, m_1 iterations are implemented with kernel 1, m_2 iterations with kernel 2, and m_3 iterations with kernel 3.

But now, using the specification in Equation 3.12, the Simulation step requires knowledge of the conditional distribution $\boldsymbol{\varphi}, \boldsymbol{w}|\boldsymbol{Y}; \theta$. In this case, and in the same context as the SAEM algorithm, Meza et al. (2007) recommends first extracting $\boldsymbol{\varphi}^{(q+1)}$ from the conditional distribution $p(\cdot|\boldsymbol{Y}, \boldsymbol{w}^{(q)}; \theta^{(q)})$ and then extracting $\boldsymbol{w}^{(q+1)}$ from distribution $p(\cdot|\boldsymbol{Y}, \boldsymbol{\varphi}^{(q+1)}; \theta^{(q)})$; that is, using a Gibbs scheme (Geman and Geman, 1984). The second conditional distribution has a simple closed form for calculation. Indeed, upon consideration of the classical parameterization of the beta distribution $Beta(\alpha, \beta)$, it becomes evident that the expression shown in Equation 2.4 is equivalent to a distribution $Beta(\alpha = u_{it}\phi, \beta = (1-u_{it})\phi)$. Following this, is easy to show that

$$w_{it}^{(q+1)}|\boldsymbol{Y}, \boldsymbol{\varphi}^{(q+1)}; \theta^{(q)} \sim Beta(Y_{it} + u_{it}\phi, S_{it} - Y_{it} + (1-u_{it})\phi),$$

so this simulation process is fairly straightforward. However, for the first conditional distribution, which is more complex, we can again resort to the Metropolis-Hastings algorithm, or what Gamerman and Lopes (2006) calls *Metropolis-within-Gibbs*. This process is analogous to that explained in the definition of the algorithm, so we will implement it with the same details and kernels mentioned above. The only change will be the definition of the values $\Delta_{i,p}^{(1)}$, which will now become

$$\Delta_{i,p}^{(1)} = \sum_{t=1}^{T_i} \mathbb{1}_{\{Y_{it}>0\}} \left(\log \frac{\Gamma(u_{it}\phi^{(q)})\Gamma((1-u_{it})\phi^{(q)})\hat{p}_{it}}{\Gamma(\hat{u}_{it}\phi^{(q)})\Gamma((1-\hat{u}_{it})\phi^{(q)})p_{it}} + \phi^{(q)}(\hat{u}_{it} - u_{it}) \log \frac{w_{it}^{(q)}}{1-w_{it}^{(q)}} \right)$$
$$+ \mathbb{1}_{\{Y_{it}=0\}} \log \frac{1-\hat{p}_{it}}{1-p_{it}}.$$

$\Delta_{i,p}^{(2)}$ is defined in the same way as in the previous section.

This specification is novel and delivers results similar to the original, but takes longer to compute (approx 1.5x the time of the first approach). For this reason, throughout this work, we will use the initial specification and scheme.

3.4 Numerical examples

Once the core of the estimation method has been defined using the SAEM algorithm for the ZIBBMR model, we will test its performance in various scenarios. The implementation of the method for calculating approximations of the standard errors of the estimates and the likelihood are the same as those described in Sections 1.4.1 and 1.4.2, respectively.

3.4.1 Artificially generated datasets

Data generation method

Given that this is the first application of our model, and because the case of unbalanced data offers no advantage for SAEM since it is also considered in the other packages we will use, we will work with balanced data generated under two different settings that describe plausible situations for microbiota data. These Settings are defined as:

- *Setting 1:* $a = b = -0.5$, $\alpha = \beta = 0.5$, $\sigma_1 = 0.7$, $\sigma_2 = 0.5$, $\phi = 6.4$.
- *Setting 2:* $a = -0.5$, $\alpha = 0.5$, $b = 0.5$, $\beta = -0.5$, $\sigma_1 = 1.4$, $\sigma_2 = 0.8$, $\phi = 10.4$.

Thus, Setting 1 describes a situation in which the covariate of interest has the same sign and value on both sides of the model (inflation at zero and abundance), and the variances of the random effects are small, as is ϕ, which also controls the dispersion of the observations; while Setting 2 describes a covariate with different effects on the two sides of the model, as well as larger variances and larger dispersion parameter ϕ. Similar to Chapter 2, we will generate 1000 data sets under each Setting and with the same total number of individuals $N = 100$ with different numbers of observations per individual, setting $T_i \in \{5, 10, 15\}$, giving us a total of six simulation scenarios. In the same way, a covariable X is defined, making $X = 0$ for the first half of individuals and $X = 1$ for the other half, and we make $Z = X$, having the same covariable in both parts of the model. For Settings 1 and 2, the total number of sequences S_{it} is a random value that follows a discrete uniform distribution between 200 and 800.

For type I error performance for Wald tests and Likelihood Ratio Tests (LRT), we will define another Setting:

- *Setting 3:* $a = 0.5$,$b = -0.5$, $\alpha = (\alpha_1, \alpha_2) = (0, -0.5)$, $\beta = (\beta_1, \beta_2) = (0, 0.5)$, $\sigma_1 = 0.7$, $\sigma_2 = 0.5$, $\phi \sim U[2, 10]$.

In this part, we will utilize small variances and assume that the parameter ϕ is random, as it is not of interest to calculate its exact value, also allowing for more robust conclusions to be drawn from hypothesis testing. Furthermore, two covariates are hereby defined: $X_1 = Z_1$, which simulates treatment and control, defined in the same way as in the previous section, and $X_2 = Z_2$, which will be a continuous variable following an $N(1, 1)$

distribution. In this section, we will maintain a constant number of observations per individual $T_i = 10$ while systematically adjusting the total number of individuals, with $N \in \{30, 50\}$. For this Setting S_{it} is defined in the same way as in Settings 1 and 2.

From the point of view of statistical inference, the following biologically relevant null hypotheses are important in the ZIBBMR model:

i. the covariates are associated with the bacterial taxon by affecting its presence or absence ($H_0 : \alpha_j = 0$);

ii. the taxon is associated with the covariates by showing different abundances ($H_0 : \beta_j = 0$);

iii. the covariates affect the taxon both in terms of presence/absence and its abundance ($H_0 : \alpha_j = \beta_j = 0$);

so, for this simulation study, we will focus on α_1 and β_1, and the hypotheses (i) and (ii) will be tested using the Wald procedure, using LRT for (iii).

The SAEM estimation is performed with 5 chains and 1000 iterations, except in the Wald test, where 10 chains were used. This is done in order to improve the convergence of the approximate Fisher matrix and better calculate the standard errors. For Settings 1 and 2, the initial values were $\theta_0 = (\phi_0, a_0, b_0, \alpha_0, \beta_0, \sigma_{1,0}, \sigma_{2,0}) = (18, -0.3, 0.2, 0.8, 0.1, 0.48, 0.72)$. For Setting 3,

$$\theta_0 = (\phi_0, a_0, b_0, \alpha_{1,0}, \alpha_{2,0}, \beta_{1,0}, \beta_{2,0}, \sigma_{1,0}, \sigma_{2,0}) = (6, 0.4, -0.7, 0.3, -0.2, 0.2, 0.1, 0.28, 0.61).$$

Results

Figures 3.1 and 3.2 and Table 3.2 offer a synopsis of the results pertaining to the estimated parameters subsequent to the execution of the simulations. A close examination reveals no significant disparities between the estimation methods proposed for fixed and random effects in Setting 1. However, a closer look at Setting 2 reveals some inaccuracies in the estimation of a and α for `gamlss`. Upon further examination of the dispersion parameters, including the variances of the random effects and the ϕ parameter, both SAEM and `glmmTMB` yielded results that were highly analogous to the anticipated values in both contexts, but `gamlss` exhibited a substantial deviation from the theoretical values. However, this deviation diminished as T_i increased, though it never attained the levels observed in the other two methods.

Proceeding to a more detailed examination of Table 3.2, in which the bias $\left(\frac{1}{R}\sum_{r=1}^{R}\hat{\theta}^r - \theta\right)$ and two measures of dispersion: mean absolute error $\left(\text{MAE} = \frac{1}{R}\sum_{r=1}^{R}|\hat{\theta}^r - \theta|\right)$ and root mean square error $\left(\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^r - \theta)^2}\right)$ are shown for each Setting, parameter, and number of observations are reported, additional intriguing findings emerge. Let us consider a division of the parameters into three parts: the zero-inflation parameters (a and α), the abundance parameters (b and β), and the dispersion parameters (σ_1, σ_2, and ϕ). For Setting 1, in the zero-inflation parameters, the lowest bias is achieved between SAEM and `glmmTMB`, but `gamlss` has the smallest error measures of all the methods. Moving on to the abundance parameters, *glmmTMB* concentrates the results with the

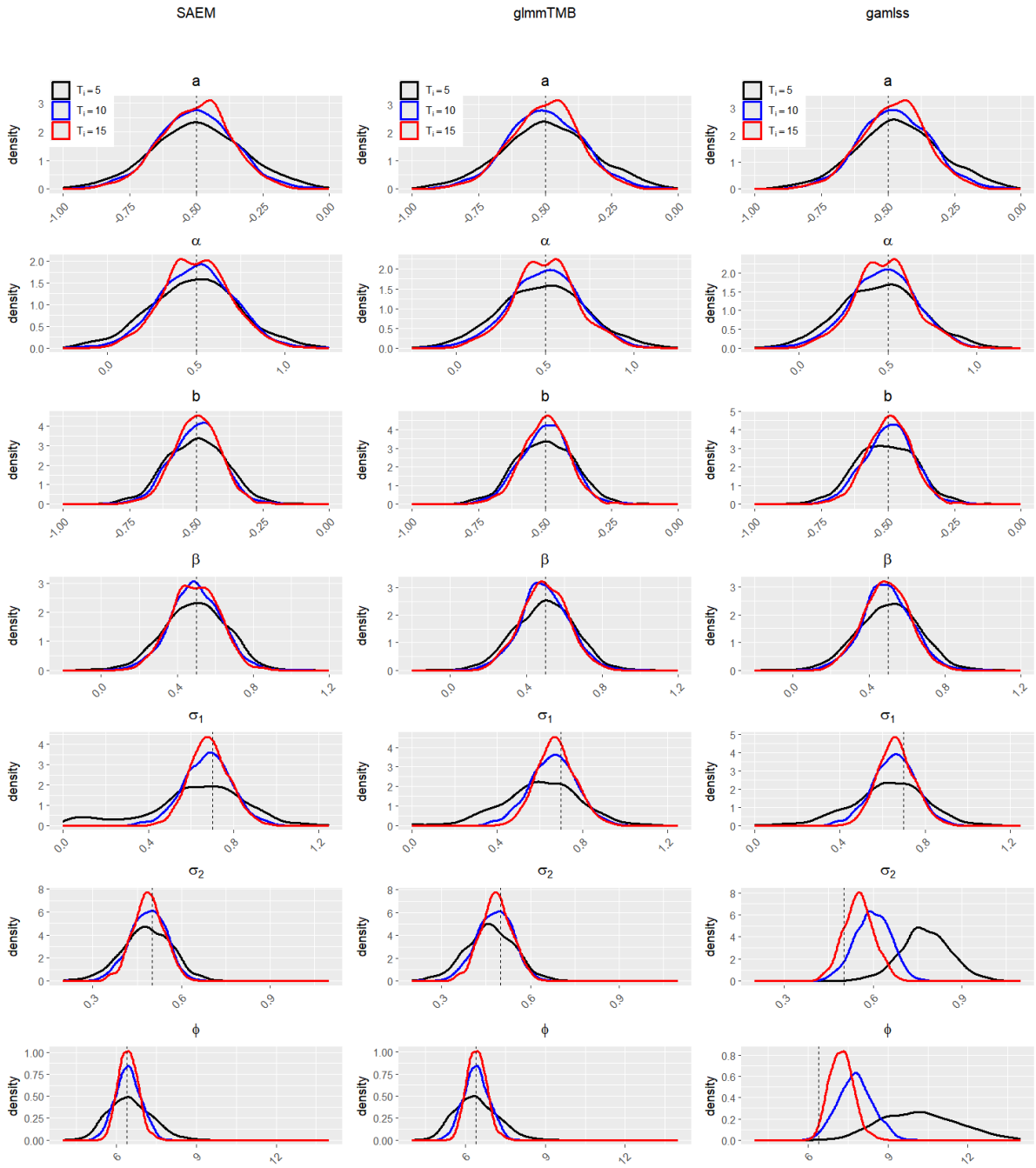


Figure 3.1: *Estimated density of the parameters obtained by the SAEM algorithm and the packages* `glmmTMB` *and* `gamlss` *under Setting 1. The dotted vertical line represents the true value of the parameter.*

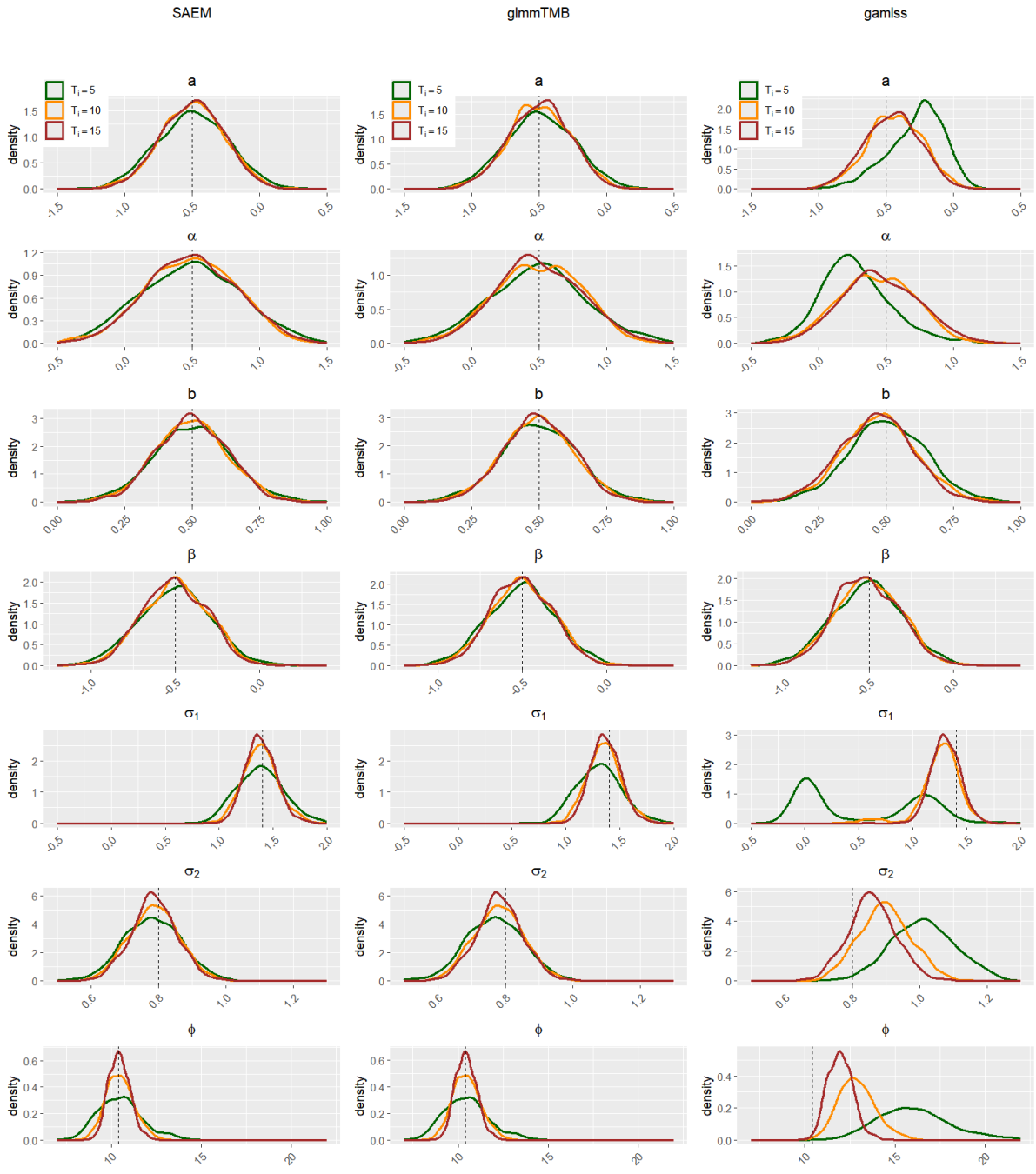


Figure 3.2: *Estimated density of the parameters obtained by the SAEM algorithm and the packages* `glmmTMB` *and* `gamlss` *under Setting 2. The dotted vertical line represents the true value of the parameter.*

Table 3.2: *Summary statistics of the results obtained by the SAEM algorithm and the packages* `glmmTMB` *and* `gamlss` *over 1000 simulation runs. For each parameter value and number of observations per individual, T_i, bold numbers indicate the lowest (absolute) value for each of bias, RMSE and MAE.*

Parameter	Value		Bias	RMSE	MAE	Bias	RMSE	MAE	Bias	RMSE	MAE
				SAEM			`glmmTMB`			`gamlss`	
a	-0.5	$T_i = 5$	**0.0076**	0.1743	0.1381	0.0093	0.1700	0.1351	0.0394	**0.1644**	**0.1301**
		$T_i = 10$	**-0.0001**	0.1419	0.1129	0.0011	0.1389	0.1109	0.0272	**0.1340**	**0.1066**
		$T_i = 15$	-0.0009	0.1296	0.1039	**0.0002**	0.1263	0.1009	0.0223	**0.1224**	**0.0984**
α	0.5	$T_i = 5$	**-0.0093**	0.2474	0.1968	-0.0104	0.2389	0.1921	-0.0399	**0.2273**	**0.1825**
		$T_i = 10$	0.0024	0.2031	0.1628	**0.0013**	0.1953	0.1560	-0.0251	**0.1863**	**0.1483**
		$T_i = 15$	0.0049	0.1837	0.1480	**0.0043**	0.1752	0.1395	-0.0181	**0.1682**	**0.1337**
b	-0.5	$T_i = 5$	0.0005	0.1136	0.0910	**-0.0003**	**0.1103**	**0.0888**	-0.0090	0.1176	0.0932
		$T_i = 10$	0.0012	0.0955	0.0758	**0.0006**	0.0939	0.0743	0.0033	**0.0939**	**0.0743**
		$T_i = 15$	0.0026	0.0857	0.0681	**0.0015**	0.0844	0.0669	0.0053	**0.0837**	**0.0664**
β	0.5	$T_i = 5$	-0.0027	0.1622	0.1303	**-0.0019**	**0.1550**	**0.1238**	0.0053	0.1607	0.1285
		$T_i = 10$	-0.0012	0.1336	0.1054	**-0.0009**	**0.1281**	**0.1011**	-0.0035	0.1287	0.1018
		$T_i = 15$	-0.0044	0.1233	0.1002	**-0.0036**	0.1191	0.0958	-0.0072	**0.1188**	**0.0955**
σ_1	0.7	$T_i = 5$	-0.0771	0.2466	0.1828	-0.0931	0.2046	0.1589	**-0.0718**	**0.1895**	**0.1454**
		$T_i = 10$	**-0.0210**	0.1135	0.0896	-0.0377	0.1147	0.0909	-0.0457	**0.1106**	**0.0877**
		$T_i = 15$	**-0.0135**	0.0928	**0.0743**	-0.0235	0.0933	0.0749	-0.0376	**0.0925**	0.0750
σ_2	0.5	$T_i = 5$	**-0.0199**	**0.0925**	**0.0706**	-0.0375	0.0933	0.0730	0.2804	0.2967	0.2804
		$T_i = 10$	**-0.0136**	**0.0632**	**0.0500**	-0.0207	0.0647	0.0511	0.0987	0.1156	0.1016
		$T_i = 15$	**-0.0113**	**0.0532**	**0.0426**	-0.0150	0.0545	0.0436	0.0518	0.0728	0.0598
ϕ	6.4	$T_i = 5$	**0.1208**	0.8169	0.6489	0.0258	**0.7992**	**0.6377**	3.8619	4.1377	3.8619
		$T_i = 10$	0.0431	0.4741	0.3795	**0.0157**	**0.4731**	**0.3791**	1.3843	1.5273	1.3877
		$T_i = 15$	0.0456	0.3700	0.2961	**0.0309**	**0.3689**	**0.2957**	0.8584	0.9680	0.8618
				SAEM			`glmmTMB`			`gamlss`	
a	-0.5	$T_i = 5$	**0.0056**	0.2603	**0.2094**	0.0084	**0.2548**	0.2039	0.2250	0.3043	0.2653
		$T_i = 10$	0.0043	0.2331	0.1867	**0.0034**	0.2255	0.1819	0.0678	**0.2137**	**0.1720**
		$T_i = 15$	**0.0029**	0.2252	0.1820	0.0031	0.2183	0.1766	0.0476	**0.2054**	**0.1671**
α	0.5	$T_i = 5$	**-0.0057**	0.3674	0.2953	-0.0073	0.3543	**0.2802**	-0.2220	**0.3408**	0.2889
		$T_i = 10$	-0.0100	0.3370	0.2706	**-0.0063**	0.3175	0.2592	-0.0719	**0.2910**	**0.2354**
		$T_i = 15$	**0.0001**	0.3299	0.2653	0.0010	0.3105	0.2503	-0.0433	**0.2865**	**0.2329**
b	0.5	$T_i = 5$	0.0015	0.1466	0.1157	**0.0010**	**0.1425**	**0.1127**	0.0059	0.1457	0.1150
		$T_i = 10$	**-0.0029**	0.1358	0.1075	-0.0031	**0.1330**	**0.1050**	-0.0235	0.1403	0.1106
		$T_i = 15$	0.0008	0.1290	0.1028	**-0.0007**	**0.1254**	**0.1005**	-0.0371	0.1402	0.1106
β	-0.5	$T_i = 5$	**-0.0017**	0.2084	0.1667	-0.0024	**0.1976**	**0.1583**	-0.0099	0.2038	0.1631
		$T_i = 10$	0.0020	0.1925	0.1539	**0.0013**	**0.1847**	**0.1478**	-0.0051	0.1910	0.1537
		$T_i = 15$	-0.0050	0.1851	0.1492	**-0.0031**	**0.1754**	**0.1416**	-0.0127	0.1863	0.1505
σ_1	1.4	$T_i = 5$	**-0.0137**	**0.2173**	**0.1737**	-0.0895	0.2238	0.1807	-0.8534	1.0247	0.8755
		$T_i = 10$	**-0.0158**	**0.1601**	**0.1269**	-0.0467	0.1617	0.1289	-0.1458	0.2412	0.1739
		$T_i = 15$	**-0.0144**	**0.1425**	**0.1135**	-0.0333	0.1429	0.1147	-0.0967	0.1637	0.1322
σ_2	0.8	$T_i = 5$	**-0.0182**	**0.0884**	**0.0704**	-0.0310	0.0919	0.0732	0.2122	0.2339	0.2133
		$T_i = 10$	**-0.0122**	**0.0741**	**0.0594**	-0.0185	0.0753	0.0605	0.0980	0.1245	0.1046
		$T_i = 15$	**-0.0122**	**0.0688**	**0.0544**	-0.0163	0.0699	0.0553	0.0599	0.0926	0.0751
ϕ	10.4	$T_i = 5$	0.1150	**1.2357**	**0.9758**	**0.0891**	1.2472	0.9846	5.7136	6.0776	5.7136
		$T_i = 10$	0.0933	**0.7848**	**0.6240**	**0.0917**	0.7878	0.6267	2.4983	2.6995	2.4997
		$T_i = 15$	0.0537	**0.5907**	**0.4705**	**0.0516**	0.5909	0.4708	1.5690	1.7177	1.5714

lowest bias, and `gamlss` has the smallest error measures, although it is true that the difference between the three methods is relatively small in all measures. Considering now the dispersion parameters, the performance of `gamlss` begins to deteriorate in both bias and error, with SAEM and `glmmTMB` sharing the scenarios with the lowest biases and error measures, confirming what we already knew from the graphical analysis.

We shall proceed to the subsequent setting, Setting 2. As in the previous case, SAEM and `glmmTMB` demonstrate optimal performance in terms of bias, yet alterations emerge in the error measures, wherein `gamlss` ceases to accumulate the lowest error measures, being eclipsed in certain instances by SAEM when $T_i = 5$, i.e., with a small number of observations per individual. This is for the zero-inflation parameters. In the context of abundance parameters, `glmmTMB` has been identified as a superior alternative, as evidenced by its superior performance in terms of error and bias measures when compared to the other methodologies. However, it is noteworthy that SAEM has demonstrated competitive edge in terms of bias measures in a few instances again, specially in scenarios with a few observations per individual. Turning to the dispersion parameters, SAEM demonstrates superior outcomes across all metrics when compared to the other two methods. Its sole inferiority, evidenced by a marginal difference, is observed in the bias of the estimates of ϕ when benchmarked against `glmmTMB`.

A number of conclusions may be drawn from these results. The zero-inflation component is the simplest component to estimate, and the three estimation methods yield analogous results. `gamlss` is distinguished by its superior performance in terms of error measures, while SAEM and `glmmTMB` are notable for their accuracy in addressing bias. Nonetheless, when transitioning to the abundance component involving the beta-binomial distribution, the superiority of `gamlss` in error measures is scenario-dependent; as the dispersion defined in the data increases, the superiority of `gamlss` in these measures diminishes, thereby being surpassed by `glmmTMB`. A closer examination of SAEM reveals two notable aspects: its exceptional performance with a limited number of observations and its superiority in variance component estimation, a field in which it consistently outperforms the other methods. Its sole disadvantage is observed in the bias of the ϕ estimate when this value is large, a shortcoming that is only surpassed by `glmmTMB`. Consequently, a comprehensive evaluation reveals that SAEM and `glmmTMB` demonstrate a notable advantage in terms of parameter estimation. Regarding the execution times of the routines, `glmmTMB` is much faster than `gamlss` and SAEM, but `gamlss` has the interesting property of not increasing the processing time when the number of observations per individual increases. On a computer with Intel Core i7-13700HX processor at 2.10 GHz, for $T_i = 5$ $(10, 15)$, on average, `glmmTMB` takes 1.05 $(1.56, 2.03)$ seconds, `gamlss` takes 13.16 $(13.22, 13.79)$ seconds, while SAEM takes 13.29 $(23.14, 36.01)$ seconds.

The following part will address the methodology of evaluating the statistical significance of covariates. As previously stated, the null hypotheses to be examined are $\alpha_1 = 0, \beta_1 = 0$, and $\alpha_1 = \beta_1 = 0$, which, by virtue of the definition of Setting 3, are all true. The initial two scenarios will be addressed through the implementation of the Wald test, while the final scenario will be approached via the LRT method. Notably, the LRT approach is not applicable in the context of `gamlss` due to its inherent inconsistencies in the estimation process based on penalized maximum likelihood (Breslow and Clayton, 1993), leading to unreliable results when calculating the log-likelihood of the models. Consequently, we will abstain from reporting its results. As illustrated in Table 3.3, the

Table 3.3: *Type I error for testing $H_0 : \alpha_1 = 0$, $H_0 : \beta_1 = 0$ and $H_0 : \alpha_1 = \beta_1 = 0$ with the SAEM algorithm and the packages* `glmmTMB` *and* `gamlss` *for nominal significance level of 0.05 and 0.01.*

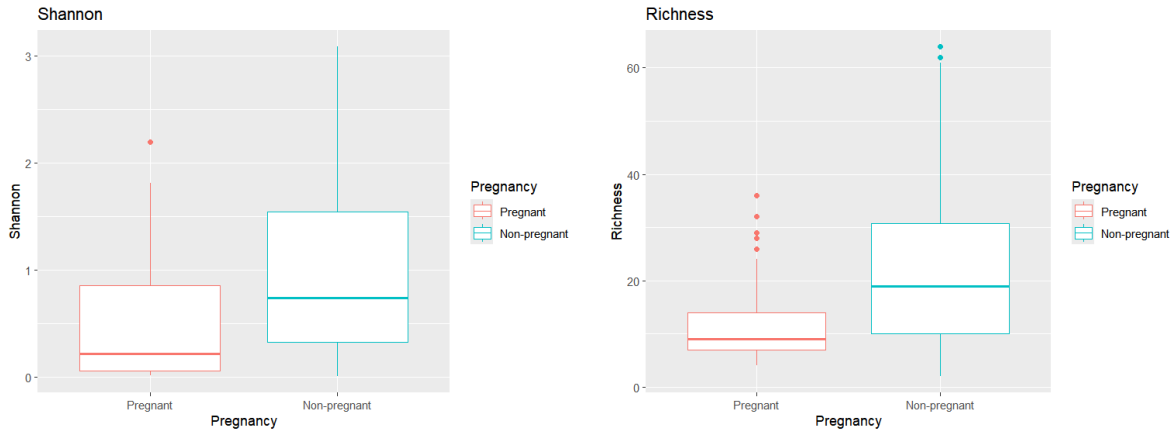
		SAEM		`glmmTMB`		`gamlss`	
		Significance level		**Significance level**		**Significance level**	
Null hypotheses		0.05	0.01	0.05	0.01	0.05	0.01
$H_0 : \alpha_1 = 0$	$N = 30$	0.078	0.025	0.076	0.015	0.284	0.170
	$N = 50$	0.060	0.019	0.061	0.018	0.328	0.195
$H_0 : \beta_1 = 0$	$N = 30$	0.077	0.022	0.081	0.019	0.168	0.067
	$N = 50$	0.080	0.022	0.085	0.023	0.159	0.058
$H_0 : \alpha_1 = \beta_1 = 0$	$N = 30$	0.063	0.012	0.069	0.014		
	$N = 50$	0.063	0.017	0.070	0.018		

Type I error has been calculated for the Wald and LRT tests on the null hypotheses under consideration. A close examination of the results obtained from the SAEM and `glmmTMB` procedures reveals a high degree of similarity between them, with both models demonstrating a strong resemblance to the nominal theoretical values. Furthermore, it is not possible to assume that there is any advantage in taking a larger number of individuals. It is evident that the performance of `gamlss` in this scenario is suboptimal, with observed values significantly deviating from the expected outcomes. This suggests potential flaws in the implementation of the Wald test procedure.

3.4.2 Pregnant women microbiome data

We will now analyze the dataset from Romero et al. (2014), which was also examined in Section 2.4.2. Since Tables 2.7 and 3.1 presented some of the important covariate characteristics in this dataset, Figure 3.3 now shows alpha-diversity indices for observations separated by pregnant and non-pregnant women. As is well known, alpha-diversity is an ecological measure that quantifies species diversity within a single sample or biological community, and can be expressed using indices that favor more diverse communities (Richness, Chao1) as well as those that focus on more balanced communities (Shannon, Simpson) (Thukral, 2017). As can be seen in this figure, all indices show greater diversity in non-pregnant women in terms of both parity and species richness. This is consistent with Ravel et al. (2011), which mentions that stability in the vaginal microbiota is a common characteristic among pregnant women, regardless of their ethnic origin.

To have a basis for comparison, the specifications of the models and taxon selection filters we use are the same as those in the previous chapter. As before, we verify statistical significance using the Likelihood Ratio Test calculated with the SAEM algorithm. As mentioned earlier, `gamlss` cannot be used for this task since it does not use the same estimation method for the log-likelihood function, and in the case of `glmmTMB`, of the 57 taxa proposed for study in the dataset, it only reached convergence in three. In all of these cases, it showed no evidence of the significance of the covariates of interest; and for the remaining 54 taxa, `glmmTMB` was unable to provide a log-likelihood value. Therefore, we cannot present results for this method in this section either. Let us recall that the models were defined by:



(a) Shannon

(b) Richness

(c) Pielou

(d) Simpson

Figure 3.3: *Alpha-diversity indices for pregnant and non-pregnant women.*

- **Model 1:** pregnancy, time and age as covariates, taking pregnancy as a factor of interest for testing.
- **Model 2:** pregnancy, time, age and interaction between time and pregnancy as covariates, testing the effect of pregnancy and the interaction.

Once the respective LRTs were developed with a significance value of 0.05, SAEM-ZIBR detected 29 taxa in Model 1 in which pregnancy is significant, the same number as SAEM-ZIBBMR. For Model 2, SAEM-ZIBR detected nine and 15 taxa for pregnancy and interaction, respectively, while SAEM-ZIBBMR detected nine and 11 taxa, respectively. While the results appear similar, it should be noted that the bacterial species detected in both models are not necessarily the same, as summarized in Table 3.4.

Table 3.4: *Bacterial taxa detected exclusively by ZIBR or ZIBBMR using the SAEM algorithm, for the covariates of interest in the two specifications considered.*

Only detected by ZIBBMR			**Only detected by ZIBR**		
Model 1	**Model 2**		**Model 1**	**Model 2**	
Pregnancy	Pregnancy	Interaction	Pregnancy	Pregnancy	Interaction
Lachnospiraceae *Lactobacillus vaginalis*	*Lachnospiraceae* *Staphylococcus*	*Prevotella genogroup 3* *Sneathia sanguinegens*	*Aerococcus* *Bacteroidales*	*Aerococcus* *Lactobacillus*	*Aerococcus* *Clostridiales* *Finegoldia magna* *Peptoniphilus* *Atopobium* *Proteobacteria*

The results indicate that the majority of the detected bacteria are the same for both methods and specifications. Therefore, it is worth analyzing the bacteria unique to each proposed model. For ZIBR, the most notable unique bacterium is *Aerococcus*, which was detected in all three covariates of interest. Wang et al. (2020) mentions that, when *Lactobacillus* increases in abundance in the vaginal microbiota for various reasons (e.g., aerobic vaginitis or pregnancy), *Aerococcus* decreases as a compensatory response. However, Huang et al. (2023) disputes this claim, pointing out that the evidence of pregnancy's influence on *Aerococcus* is limited to the data from Romero, which we examined here, and is not significant in other similar studies.

In the case of *Lachnospiraceae*, the most notable taxon detected by ZIBBMR, Yang et al. (2023) compared the gut microbiota of women with preterm and normal births. The study found that women with normal births had higher abundances of *Lachnospiraceae*, showing a notable correlation. According to Sorbara et al. (2020), this may be due to their role as producers of anti-inflammatory metabolites in the human body. However, this role is fulfilled by a different microbial colony than the one we are analyzing here, but this is a finding that deserves our attention. ZIBBMR also detects *Prevotella genogroup 3* and *Sneathia sanguinegens*, which were discussed at length in Section 2.4.2 regarding their role in pregnancy (Severgnini et al., 2022).

3.5 Conclusions

In this chapter, we proposed a new model for microbiota data expressed as sequence counts and an estimation and inference method based on the SAEM algorithm, which produced

good results in the previous chapter. Due to the hierarchical definition of the new model, we also suggested modifying the SAEM simulation phase using a Gibbs scheme. While this modification is not difficult to implement, it requires additional processing time. Thus, only the traditional approach was considered in the work carried out in this chapter, leaving comparisons between developments for future work. Aditionally, a comparative analysis was conducted between the outcomes of parameter estimation and statistical significance tests, and those of two alternative methods that have been documented in the extant literature. While both SAEM and `glmmTMB` demonstrated comparable performance and significant superiority over `gamlss`, SAEM exhibited optimal performance for small datasets with a limited number of observations per individual. This is a critical aspect, as it facilitates the extraction of information under highly constrained conditions, a common vulnerability observed in numerous healthcare research studies (Pais et al., 2024).

Applying the ZIBBMR model and comparing it with the ZIBR results on Romero's dataset revealed key findings. While there are differences in the results of each model for the same specification, it is not possible to clearly state the supremacy of one model over the other. It may be interesting to use both models as complementary phases of the same study, employing their respective advantages to draw better conclusions from the same experiment. In the study of the human microbiota, few models can address the challenges posed by the special characteristics of data extracted by genetic sequencing. It is also important to note that in this particular example, the log-likelihood calculation provided by the `glmmTMB` package failed for almost all bacterial taxa considered, as did the calculation of standard errors, both of which are key components for statistically determining the significance of the model's covariates. According to Shi et al. (2016), when it comes to the application of models in the context of microbiota data, the accurate estimation of parameters is not the sole concern (a domain in which `glmmTMB` demonstrates proficiency). Instead, the emphasis shifts towards the efficacy of statistical inference, a domain in which SAEM emerges as a preeminent solution.

Finally, a notable strength of the ZIBBMR model is its capacity to utilize data collected through genetic sequencing techniques. As previously stated, McMurdie and Holmes (2014) has indicated that working with the original data is consistently the most recommended approach. In addition, Gloor et al. (2016) asserts that the utilization of calculated proportion data invariably renders the findings susceptible to the presence of spurious correlations, which have the potential to modify the conclusions derived from models based on this data. Consequently, by incorporating this advantage into the utilization of the beta-binomial distribution, a methodology that has demonstrated efficacy in other studies (Martin et al., 2020; Dolzhenko and Smith, 2014) employing overdispersed, correlated, and variable proportion count data, it is anticipated that our model will facilitate the extraction of conclusions from existing experiments and, more optimally, the conceptualization of novel experiments that will yield significant findings within the domain of human microbiota analysis.

CHAPTER 4

APPLICATIONS TO REAL MICROBIOME DATASETS: THE COBRA-ENV AND MODUL-CF EXPERIMENTS*

In this chapter, we will discuss two studies conducted on real data collected in France, in which the models and methods developed in previous chapters were applied. Through these studies, we tested the usefulness of these developments in real-world situations, particularly in associating changes in the microbiota with disease progression or the onset of clinical symptoms. Section 4.1 focuses on the analysis of indoor microbiota for asthmatic and non-asthmatic patients and its correlations based on inference networks, while in Section 4.2, we study the differential abundance of bacteria in the lungs of children affected by cystic fibrosis undergoing CFTR modulator treatment.

4.1 Differential abundance study between bacteria and fungi on indoor microbiome of asthmatic and non-asthmatic patients

Asthma is a widespread chronic respiratory disease affecting both children and adults across the globe. While the prevalence and severity vary between regions, its impact on public health remains substantial. Asthma symptoms often emerge in childhood and are characterized by episodes of wheezing, coughing, chest tightness, and shortness of breath. The pathophysiology of asthma is complex, involving exaggerated immune responses to otherwise harmless environmental antigens such as dust mites or mold. These responses lead to the activation of inflammatory pathways, which can result in chronic airway remodeling (Fu et al., 2021, 2022). Recent research suggests that the composition of indoor

*This chapter corresponds to two contributions made within the framework of the Associated Team INRIA-Universidad de Valparaíso, *Valid statistical Analysis of Longitudinal compositional and high-dimensional microbiome data to Predict health Outcomes (VALPO)* and the project MATH-AmSud AMSUD230032 SMILE.

microbial communities may be linked to asthma risk and could serve as a modifiable factor for disease prevention (Salaun-Ferron et al., 2023). Given that individuals spend a significant portion of their time indoors, particularly in early life, understanding the relationship between indoor microbial exposures and asthma is crucial (Vandenborght et al., 2021). However, characterizing these associations is statistically and biologically challenging due to the nature of microbiome data, which are typically sparse, compositional, and high-dimensional.

To identify microbial taxa associated with disease status, researchers often employ Differential Abundance Analysis (DAA). This method seeks to detect significant differences in microbial abundance between predefined groups, such as individuals with asthma and healthy controls. DAA serves as a foundational step in discovering potential microbial biomarkers that may inform prevention, diagnosis, or therapeutic strategies. However, existing DAA methods face several statistical hurdles: the data often include a large number of zeros (either due to absence or detection limits), show high variability in abundance across samples, and reflect only relative rather than absolute microbial quantities (Yang and Chen, 2022). Numerous DAA methods have been developed to tackle these challenges. Some rely on over-dispersed count models, such as the negative binomial or beta-binomial distributions, while others use zero-inflated or hurdle models to distinguish between true absences and undetected low-abundance taxa. Normalization techniques also play a critical role in managing the compositional nature of the data. Despite these methodological advances, results can vary widely between tools, and choosing the appropriate method for a specific dataset remains a nontrivial task (Cappellato et al., 2022).

Beyond traditional DAA, network-based approaches offer an additional layer of insight (Hossine et al., 2025). These methods consider the dependencies and interactions among microbial features, rather than evaluating each taxon in isolation. By constructing microbial co-occurrence networks, researchers can identify taxa that are directly influenced by external factors—such as disease state—versus those affected indirectly through microbial interactions. This holistic perspective allows for a more nuanced understanding of microbial community dynamics in disease contexts.

Considering this, in this section we investigate differences in indoor microbiome composition between individuals with and without asthma. We apply statistical and network-based approaches specifically designed for microbiome data, accounting for key features such as compositionality, zero inflation, spatial clustering (handled via random effects or adjustment for environmental covariates), and high dimensionality. Differential abundance analysis is used to evaluate associations between specific taxa and asthma status, while network analysis provides a broader view of the interdependencies between bacterial and fungal communities. Both analyses adjust for potential confounders.

4.1.1 Data and models

Data

Observations come from 42 asthma patients (cases) from the national Cohort of Bronchial Obstruction and Asthma (COBRA) cohort and 20 healthy individuals living in non-asthmatic, non-respiratory allergic households (controls). We analyzed the indoor microbial flora using electrostatic dust collectors (EDCs) under standardized conditions, over

10 weeks in spring 2019 in the southwest region of France. Each EDC consisted of a textile surface mounted in a plastic folder, placed horizontally to passively collect settling dust. All EDCs were deployed simultaneously in the bedrooms of patients and controls over a 10-week period in spring 2019, following an identical protocol.

Microbial communities in the collected dust were characterized using amplicon-targeted metagenomics. The bacterial community was profiled via the V3–V4 region of the 16S rRNA gene, and the fungal community via the ITS2 region of the fungal rDNA, using optimized and standardized library preparation protocols from Metabiote (GenoScreen, Lille, France). Raw sequences were quality-filtered, assembled, and processed using a standardized bioinformatics pipeline. Amplicon Sequence Variants (ASVs) were first inferred to achieve high-resolution identification of unique sequences. These ASVs were subsequently clustered into Operational Taxonomic Units (OTUs) at 97% sequence similarity to facilitate taxonomic assignment and comparison. Taxonomic classification was performed using reference databases, with bacterial taxa identified at the genus level and fungal taxa at the species level. Preprocessing involved first identifying bacteria and fungi in our samples that are well-known or suspected to be related to asthma. An umbrella review was performed to synthesize previously published systematic reviews and meta-analyses. These bacteria and fungi were considered throughout the analysis without being excluded due to low representation or statistical significance. Individual Operational Taxonomic Units (OTUs) abundances were converted to relative abundances by dividing by the total counts, separately for bacteria and fungi. Only microorganisms present at $\geq 1\%$ in at least 3 households were considered.

Additionally, demographic data (i.e. sex, age, municipality) and indoor environmental characteristics (i.e. presence of mold) were collected. Environmental characteristics (i.e. rural/urban residence, humidity, proximity to coastal wetlands, pollen index, air quality) were inferred from the municipality linked to publicly available databases. A total of 16 clusters were manually defined by grouping geographically close municipalities with similar environmental characteristics.

Clinical characteristics were recorded for asthma patients, including *FeNO* (fractional exhaled nitric oxide, an indicator of airway inflammation), *VEMS%* (percentage of predicted forced expiratory volume in 1 second, a measure of lung function), *Tiff* (Tiffeneau index, which assesses airflow obstruction), *eosinophil count* (a type of white blood cell linked to allergic responses), *BMI* (body mass index), and asthma severity level (based on symptom control and exacerbation frequency). These data were available at two time points: prior to the 10-week exposure period in spring 2019, and at a single fixed time point in the preceding winter, identical for all participants.

Models

After collecting abundance data for bacteria and fungi, along with geographical and demographic covariates, a statistical analysis is performed with two objectives in mind: first, to identify bacterial and fungal taxa that demonstrate differences in abundance or presence between asthmatics and non-asthmatics that have not been previously detected, and second, to infer possible correlations between these taxa, distinguishing between those that have been previously identified and those detected in the preceding step.

For the first objective, we will use the Zero Inflated Beta Regression (ZIBR) model

(Chen and Li, 2016), which we worked with in Chapter 2. Let Y_{ij} denote the relative abundance of a given taxon in household j within geographical cluster i, with $i = 1, \ldots, 16$ and $j = 1, \ldots, n_i$ ($\sum_{i=1}^{16} n_i = 63$ total households). The distribution of Y_{ij} is modeled according to Equation 2.1:

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } 1 - p_{ij}, \\ \text{Beta}(u_{ij}\phi, (1-u_{ij})\phi) & \text{with probability } p_{ij}, \end{cases}$$

where $0 \leq Y_{ij} < 1$, $0 < p_{ij}, u_{ij} < 1$, and $\phi > 0$ is the precision parameter.

We specified a general model to estimate the association between asthma status and microbial relative abundance, while adjusting for potential confounders: k demographic and indoor environmental covariates (e.g., sex, age group, mold exposure) D_{jk}; and l environmental covariates (e.g., forestation rate, humidity, pollen and air quality indicators) E_{jl}.

The parameters of the zero-inflated Beta distribution are modeled as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a_i + \alpha_1 \cdot \text{Asthma}_{ij} + \sum_k \gamma_k D_{jk} + \sum_l \eta_l E_{jl},$$
$$\log\left(\frac{u_{ij}}{1-u_{ij}}\right) = b_i + \beta_1 \cdot \text{Asthma}_{ij} + \sum_k \delta_k D_{jk} + \sum_l \zeta_l E_{jl},$$

where $a_i \sim N(0, \sigma_a^2)$, $b_i \sim N(0, \sigma_b^2)$ are random intercepts at the cluster level. We considered two modeling strategies for controlling the environmental effects, both including Asthma as the main variable, and all D_{jk} as fixed effects:

- Fully adjusted fixed-effects models including all E_{jl} and no random effects ($\sigma_a^2 = \sigma_b^2 = 0$).
- Random-effect on geographical clusters models including random intercepts a_i, b_i to account for within-cluster correlation and excluding environmental covariates E_{jl} ($\eta_l = \zeta_l = 0$).

We tested for associations between asthma and the relative abundance of each taxon using Wald test on the asthma coefficients through the `glmmTMB` R package (Brooks et al., 2017). The decision to employ this package over the SAEM algorithm-based estimation developed in Chapter 2 was guided by two primary considerations. The primary concern pertains to the substantial quantity of bacteria and fungi encompassed within the dataset, that renders the SAEM-based modeling approach excessively time-consuming; in comparison, `glmmTMB` requires slightly less time, although it is subject to convergence problems in certain models.

The second reason pertains to the utilization of the ZIBR model. For those taxa that did not record a high number of zeros, the implementation of the ZIBR model would not be recommended. Rather, the zero-inflation component is discarded for those taxa that had less than 20% of null abundance observations. Adjusting this simple beta regression is a straightforward process in `glmmTMB`; however, adapting the work carried out in Chapter 2 to serve this purpose is still ongoing, since although theoretically it is a simple development, its implementation in code is not so straightforward.

Once the bacteria and fungi sensitive to the asthma covariate have been selected, the next step is to investigate microbial co-occurrence patterns, for which we employed a multivariate Zero-Inflated Poisson Lognormal (ZIPLN) model (Batardière et al., 2024), which extends the Poisson lognormal framework (Chiquet et al., 2019). Let $Y'_j = (Y'_{j1}, \ldots, Y'_{jp})$ denote the vector of microbial counts (bacteria and fungi) observed in household j, with p taxa. The ZIPLN model introduces two levels of latent variables:

- A latent Gaussian vector $Z_j = (Z_{j1}, \ldots, Z_{jp}) \in \mathbb{R}^p$, which captures correlations among taxa and serves as the log-intensity of the Poisson model,
- A binary vector $W_j = (W_{j1}, \ldots, W_{jp}) \in \{0,1\}^p$, which models the presence of structural zeros via independent Bernoulli trials with probability $\pi_{jk} = \mathbb{P}(W_{jk} = 1)$.

Formally, the model is specified as:

$$Z_j \sim N(X_j B, \Omega^{-1}), \quad W_j \sim \text{Bernoulli}(\pi_j),$$
$$Y'_{jk} \,|\, W_{jk}, Z_{jk} \sim \begin{cases} 0 & \text{if } W_{jk} = 1, \\ \text{Poisson}(\exp(o_{jk} + Z_{jk})) & \text{if } W_{jk} = 0, \end{cases}$$

where $X_j \in \mathbb{R}^d$ is the vector of covariates, o_{jk} is the offset accounting for sequencing depth normalization (which allows to convert Y'_{jk} in relative abundance Y_{jk}), $B \in \mathbb{R}^{d \times p}$ is the matrix of regression coefficients, and Ω is the precision matrix whose inverse $\Sigma = \Omega^{-1}$ represents the latent covariance between taxa. Network inference focuses on estimating the conditional dependencies between taxa, encoded in the precision matrix Ω. A non-zero off-diagonal entry Ω_{gh} implies a direct interaction (conditional dependence) between taxa g and h, adjusting for all other taxa and adjustment covariates.

The ZIPLN network model was fitted separately for asthmatic and control groups using the `PLNmodels` R package (Chiquet et al., 2021). To estimate sparse networks, a graphical LASSO penalty is applied to the precision matrix. The optimal sparsity level was selected using the Stability Approach to Regularization Selection (StARS), which retains edges that are consistently identified across subsampled datasets.

4.1.2 Results

Data description

Table 4.1 summarizes the characteristics of the dataset, particularly the variables we will use in the modeling. It shows that the two groups differ in their distributions, with cases including more older individuals, more rural residents, and greater exposure to cleaner air and lower pollen concentrations. Air quality and pollen concentration were correlated with geographical zone. The variable mold had missing values for two controls, which were imputed as "No", and a sensitivity analysis confirmed that this imputation had minimal impact on the results.

We initially considered 967 bacterial taxa (from genus level upward) and 3,709 fungal taxa (from species level upward). After applying prevalence and abundance filtering criteria, we retained 101 bacterial and 139 fungal taxa. Figure 4.1 displays bacterial alpha-diversity indices (which we discuss in more detail in Section 3.4.2) for asthma cases

Table 4.1: *Participant characteristics by asthma status. Values are presented as n (%).*

Characteristic		Overall (N=62)	No (N=20)	Yes (N=42)	p-value
Sex	Female	40 (65%)	13 (65%)	27 (64%)	> 0.9
	Male	22 (35%)	7 (35%)	15 (36%)	
Age Group	< 60 years	35 (56%)	17 (85%)	18 (43%)	0.002
	≥ 60 years	27 (44%)	3 (15%)	24 (57%)	
Forestation Rate	High	14 (23%)	3 (15%)	11 (26%)	0.5
	Low	48 (77%)	17 (85%)	31 (74%)	
Zone	Rural	26 (42%)	4 (20%)	22 (52%)	0.016
	Urban	36 (58%)	16 (80%)	20 (48%)	
Humidity	High	12 (19%)	4 (20%)	8 (19%)	> 0.9
	Medium	9 (15%)	2 (10%)	7 (17%)	
	Low	24 (39%)	8 (40%)	16 (38%)	
	Weak	17 (27%)	6 (30%)	11 (26%)	
Air Quality	Good	39 (63%)	8 (40%)	31 (74%)	0.010
	Moderate or unhealthy	23 (37%)	12 (60%)	11 (26%)	
Pollen Concentration	High	52 (84%)	20 (100%)	32 (76%)	0.023
	Low	10 (16%)	0 (0%)	10 (24%)	
Mold Presence	Yes	8 (13%)	4 (22%)	4 (9.5%)	0.2
	No	52 (87%)	14 (78%)	38 (90.5%)	

Note: p-values computed using Pearson's Chi-squared or Fisher's exact test.

and controls. The richness index shows similar median values between groups, though greater variability is observed among cases. Other indices balance richness and evenness, with those on the top of Figure 4.1 placing more emphasis on richness and those on the bottom more on evenness. When accounting for evenness, asthma cases tend to exhibit higher overall diversity.

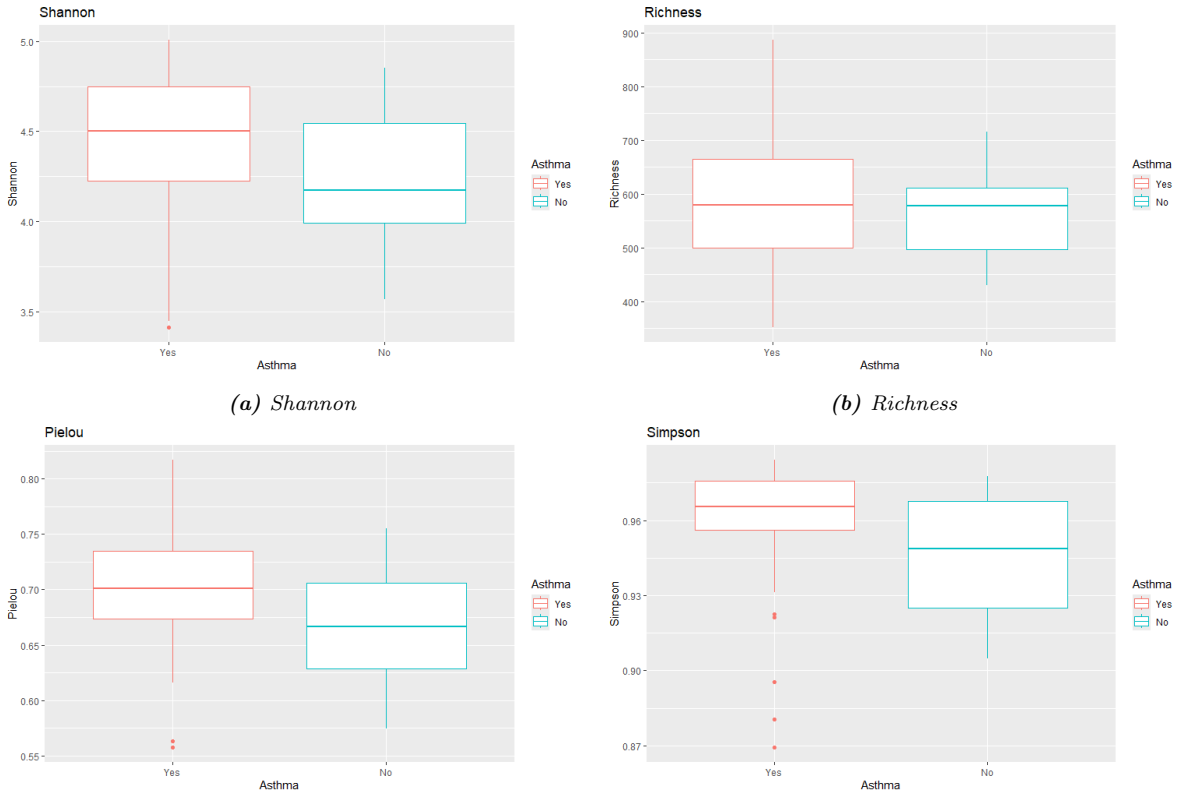
Differential abundance

For each of the bacteria and fungi that met the previously mentioned filters, we used the ZIBR model and its zero inflation variant to identify those taxa associated with asthma. Table 4.2 shows results for the most relevant taxa previously reported in the human respiratory microbiome literature as being positively or negatively associated with asthma or its severity. Only one model per taxon—the one with the lowest p-value for the asthma effect—is presented. In addition, 24 bacterial and 41 fungal taxa not previously reported were significantly associated with asthma ($p < 0.05$). In random-effects models, adjustment covariates were rarely significant. In fully adjusted models, age group, rural vs. urban setting, and humidity were frequently associated with microbial outcomes.

Inference networks

A total of 33 bacterial and 67 fungal taxa were included in network models, which were adjusted for age, rural vs. urban location, and humidity, and run separately for asthmatic and non-asthmatic households. Taxa previously reported in the literature were included regardless of their significance in prior models, while novel taxa were included only if significantly associated with asthma ($p < 0.05$). The networks obtained for asthmatic and non-asthmatic groups are shown in Figure 4.2.

The role of *Enterobacteriaceae* in the environment of asthmatic households appears to be supported by the network analysis, showing a notable interaction with *Escherichia*. Asthma households show reduced interactions both within and between kingdoms compared to controls while adjusting for age, zone, and humidity. Overall, taxa presence is



(a) Shannon *(b) Richness*

(c) Pielou *(d) Simpson*

Figure 4.1: *Bacterial alpha-diversity indices for cases and controls.*

Table 4.2: *Taxa associated with asthma, according to covariate and taxa modeling approaches.*

Kingdom	**Taxa**	**Beta Estimate**	**p-value**	**Covariate modeling**	**Taxa modeling**
Bacteria	*Enterobacteriaceae*	0.54	0.0332	Random-effect	Abundance
Bacteria	*Enterococcus*	0.53	0.0832	Random-effect	Abundance
Fungi	*Alternaria dactylidicola*	0.46	0.0357	Fully adjusted	Abundance
Fungi	*Aspergillus niger*	-1.94	0.0980	Fully adjusted	Presence
Fungi	*Aspergillus puulaauensis*	-1.39	0.0325	Random-effect	Presence
Fungi	*Blumeria graminis*	1.17	<0.0001	Fully adjusted	Abundance
Fungi	*Malassezia*	3.36	0.0146	Fully adjusted	Presence
Fungi	*Malassezia restricta*	-0.46	0.0147	Random-effect	Abundance
Fungi	*Malassezia slooffiae*	1.72	0.0678	Random-effect	Presence
Fungi	*Penicillium aethiopicum*	-1.02	0.0185	Fully adjusted	Abundance
Fungi	*Penicillium commune*	-1.43	0.0491	Fully adjusted	Presence
Fungi	*Penicillium salamorum*	1.31	0.0609	Random-effect	Presence
Fungi	*Trichoderma paraviridescens*	8.39	<0.0001	Fully adjusted	Abundance
Fungi	*Wallemia canadensis*	-1.64	0.0072	Fully adjusted	Abundance

Note: only taxa previously reported in the respiratory literature as being associated with asthma are displayed. Positive estimates indicate greater abundance or presence in asthmatic households.

similar between groups, but networks in asthmatic households show more interactions within small groups—often within the same family—whereas non-asthmatic households exhibit a greater number of interactions per taxon.

4.1.3 Conclusions

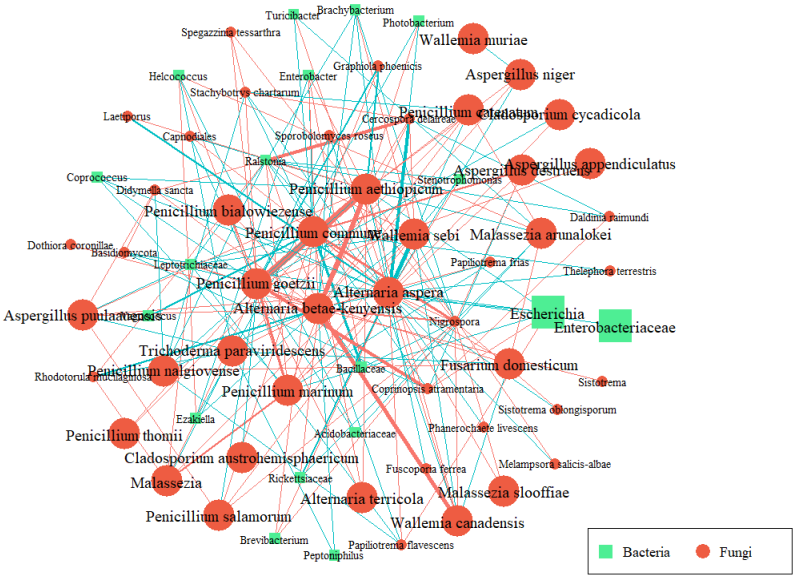
In microbiome studies of respiratory diseases, statistical modeling is inherently complex. Robust analysis pipelines—incorporating methods such as ZIBR and Poisson-lognormal graphical inference—are essential. A notable strength of the ZIBR model lies in its capacity to incorporate random effects, a feature that facilitates the detection of microorganisms exhibiting sensitivity to therapeutic interventions. This capacity is particularly advantageous in scenarios where observations exhibit similarities, such as in the case of repeated measurements on the same individual or measurements from individuals belonging to the same geographical region. Moreover, incorporating the inflation component in zeros introduces a novel dimension to the analysis, enabling the identification of effects not only in terms of abundance but also in the context of a specific taxon within the data. However, it is worth asking whether a model based on data such as sequence counts could yield similar or even better results, given that data of the same nature would be used in both steps of the modeling process.

The construction of networks according to the ZIPLN model in the microbiota is a particularly useful result, as it not only identifies new species of microorganisms related to asthma, but also reveals relationships between them and already known species. These networks facilitate access to their study through the examination of measures of influence or centrality, such as the Louvain algorithm (Blondel et al., 2008), or spectral clustering techniques (Von Luxburg, 2007), in the first case; as well as calculations of degree centrality, intermediation, and proximity indices (Bonacich, 1987; Freeman, 1977), in the second. These measures have already been utilized in the fields of microbiology (Faust and Raes, 2012) and medicine (Barabási et al., 2011); therefore, their application to our work may provide novel insights into the study of asthma as a multifactorial disease.

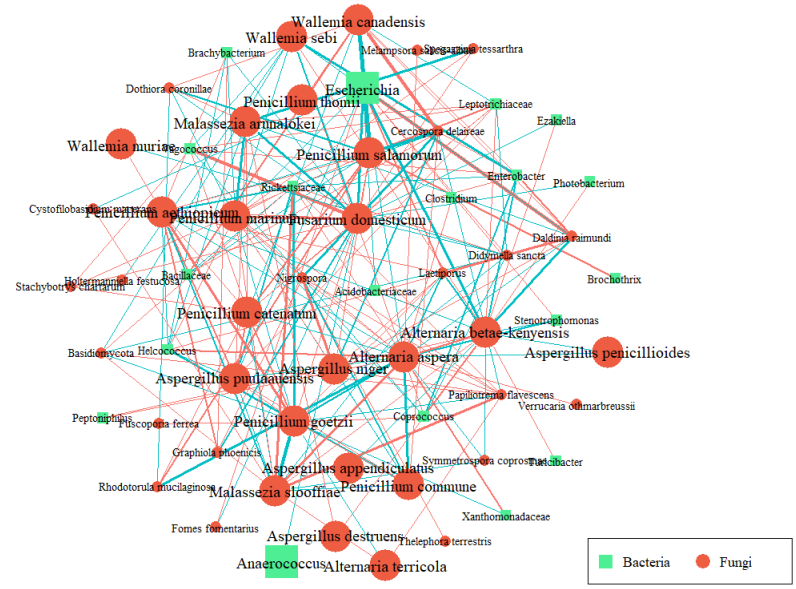
4.2 Identification of bacteria associated with clinical criteria in pediatric patients affected by cystic fibrosis

Cystic fibrosis (CF) is an autosomal recessive genetic disease caused by mutations in the CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) gene, which encodes an AMPc-dependent chloride channel present on epithelial surfaces. CFTR dysfunction disrupts ion and water transport in tissues, leading to the accumulation of thick, dehydrated secretions in various organs, especially the lungs, pancreas, intestine, and reproductive system (Rowe et al., 2005).

In the respiratory system, progressive airway obstruction by viscous mucus creates an environment conducive to microbial colonization and chronic inflammation. Respiratory symptoms, such as persistent cough, recurrent lung infections, dyspnea, and decreased lung function, are the leading causes of morbidity and mortality in people with CF (Elborn



(a) Patients



(b) Controls

Figure 4.2: *Inferred networks for cases and controls. Bacteria are represented by green squares, and fungi by red circles. Larger nodes indicate taxa previously reported in the literature.*

et al., 1991). In addition, the patient's immune system, although functional, responds in an exacerbated and prolonged manner to microbial agents, contributing to the destruction of lung tissue through inflammatory mechanisms.

In the last two decades, the role of the lung microbiome in the pathophysiology of cystic fibrosis has been the subject of increasing attention. Metagenomic studies have shown that the lungs of CF patients harbor complex, dynamic, and highly personalized microbial communities (Zhao et al., 2012). Throughout the course of the disease, this microbiota undergoes alterations that reflect the clinical status of the host, exposure to antibiotics, and the progression of lung damage. One of the most relevant microorganisms in this context is *Pseudomonas aeruginosa*, an opportunistic Gram-negative bacterium that chronically colonizes the lungs of most adult CF patients. This colonization is associated with accelerated lung function decline, increased respiratory exacerbations, and higher mortality risk (Folkesson et al., 2012). *P. aeruginosa* is capable of forming biofilms that are resistant to the immune system and to multiple antibiotics, making it difficult to eradicate and promoting long-term persistence.

Another pathogen of growing importance is *Aspergillus fumigatus*, a filamentous fungus that can colonize the lungs of CF patients and trigger an exacerbated immune response known as allergic bronchopulmonary aspergillosis (ABPA). This condition can aggravate bronchial obstruction, induce eosinophilia, and increase mucus production, thereby contributing to clinical deterioration (Amin et al., 2010).

Over the last decade, the treatment of CF has evolved significantly with the introduction of CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) modulators, drugs designed to correct the defective function of the CFTR protein, whose dysfunction is the primary cause of this hereditary disease. Enhancers, such as ivacaftor, improve the opening of the channel already present in the membrane, while correctors, such as lumacaftor, tezacaftor, and elexacaftor, facilitate the proper folding and transport of CFTR. Combination therapies, especially the triple combination elexacaftor/tezacaftor/ivacaftor (ETI), have shown substantial clinical benefits, including sustained improvements in lung function, weight gain, reduction in respiratory exacerbations, and improved quality of life in people with at least one copy of the F508del mutation (Heijerman et al., 2019; Middleton et al., 2019).

In pediatric patients, CFTR modulators have also been shown to be effective and well tolerated. Early use has been associated with improvements in nutritional and respiratory markers, as well as a reduction in colonization by common lung pathogens. Recent studies have reported that treatment with ETI in children over 6 years of age significantly improves forced expiratory volume (FEV1), nutritional status, and CFTR function biomarkers, even in those with mild or early disease. This is especially relevant, as intervening before irreversible lung damage occurs could modify the natural course of the disease, opening up the possibility of preventing clinical progression in future generations of CF patients.

Considering this, in this section the focus will be on identifying bacteria that demonstrate abundance or presence in response to specific significant clinical indicators in pediatric patients with cystic fibrosis undergoing CFTR therapy. To this end, the SAEM-ZIBR approach, as delineated in Chapter 2, will be employed, with demographic variables being controlled for. Furthermore, a comparison will be made between the results obtained from this study and those previously reported by the `glmmTMB` package. This will allow for the

verification of any significant differences between the two methods.

4.2.1 Data and methods

Data

The dataset under consideration is derived from 96 patients participating in the MODUL-CF experiment (Response to CFTR Modulators in Cystic Fibrosis Patients Under 18 Years, France). The experimental participants are undergoing treatment with CFTR modulators, as indicated by the title of the experiment. The study participants were limited to those who had undergone a baseline measurement and at least one additional measurement at 3, 6, and/or 12 months. In addition to the documentation of age and weight, several key markers were considered in order to assess the patient's clinical status. The following metrics were included:

- $ppFEV_1$ (percent predicted forced expiratory volume in 1 second), a measure of lung function that is also used to monitor the effectiveness of treatments;
- the *BMI z-score*, a measure of a child's weight and height development as part of a standardized population; and
- indicators of colonization by *Pseudomonas* and *Aspergillus*, which, as previously mentioned, represent risk markers for the patient's condition.

Table 4.3: *Description of some variables of the working dataset.*

Characteristic	**N = 96**[1]
Sex	
Female	47 (49%)
Male	49 (51%)
Age	
Under 5	19 (20%)
5 to 8	20 (21%)
Over 8	57 (59%)
N. of observations	
2	24 (25%)
3	44 (46%)
4	28 (29%)

[1] n (%).

With regard to the bacteria present in the samples, sequence count information was collected for 64 different taxa, which, after being subjected to abundance and importance filters (having values other than zero in at least 5% of the observations and exceeding 1% relative abundance in any observation), were reduced to 52 taxa at the genus level, of which 292 observations were obtained for the 96 individuals.

Model

Based on these relative abundance data, our objective will be to verify whether clinical variables are significant in explaining differences in the abundance or presence of the bacteria recorded in the lung microbiota of pediatric patients. To do this, we will use the ZIBR model (Chen and Li, 2016), which, given Y_{ij} the relative abundance of a taxon at time t for the individual i, with $i = 1, \ldots, 96$ and $j = 1, \ldots, t_i$ ($\sum_{i=1}^{96} t_i = 292$ total observations), models the distribution of Y_{ij} according to Equation 2.1:

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } 1 - p_{ij}, \\ \text{Beta}(u_{ij}\phi, (1-u_{ij})\phi) & \text{with probability } p_{ij}, \end{cases}$$

where $0 \leq Y_{ij} < 1$, $0 < p_{ij}, u_{ij} < 1$, and $\phi > 0$ is the precision parameter. In a similar way to Section 4.1.1, the parameters of the zero-inflated Beta distribution are modeled as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = a_i + \sum_k \gamma_k C_{jk} + \sum_l \eta_l D_{jl},$$
$$\log\left(\frac{u_{ij}}{1-u_{ij}}\right) = b_i + \sum_k \delta_k C_{jk} + \sum_l \zeta_l D_{jl},$$

where $a_i \sim N(0, \sigma_a^2)$, $b_i \sim N(0, \sigma_b^2)$ are random intercepts at the individual level, C_{jk} are control covariables (measurement time, age and sex) and D_{jl} are the clinical covariables of interest. Using this definition, we will implement the Likelihood Ratio Test to verify the null hypothesis $H_0 : \eta_l = \zeta_l = 0, \forall l$, calculating the log-likelihood of the models for each taxon using the `glmmTMB` package (Brooks et al., 2017) as well as the estimation process based on the SAEM algorithm developed in Chapter 2. We will then compare the taxa that show significance under both estimation methods.

4.2.2 Results

Data description

Figure 4.3 shows the evolution of the BMI z-score over the 12 months of sampling for individuals, as well as its distribution for each point in time. Given the age of the patients, it is expected that this index will increase, which is evident in the graph on average, even though there are some patients who show significant declines that could be attributed, among other things, to the progression of the disease. Unfortunately, there is insufficient evidence to confirm or rule out the latter.

On another note, Figure 4.4 shows the progression of colonization by *Aspergillus fumigatus* and *Pseudomonas aeruginosa* in patients and how these percentages evolve throughout the measurements. Although there are large variations, the percentages are not very high at any time, with *Aspergillus fumigatus* having the highest incidence at almost every time point in the patients studied. In contrast, although *Pseudomonas aeruginosa* initially registers high incidence values, after 3 months it decreases radically and never reaches more than 10% of patients.

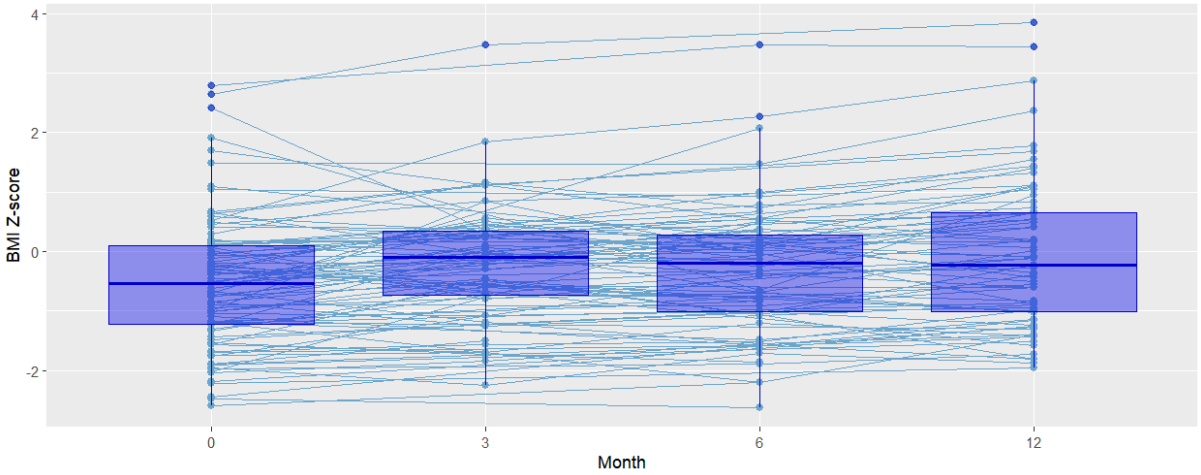


Figure 4.3: *Evolution and distribution of the BMI z-score in the patients over time.*

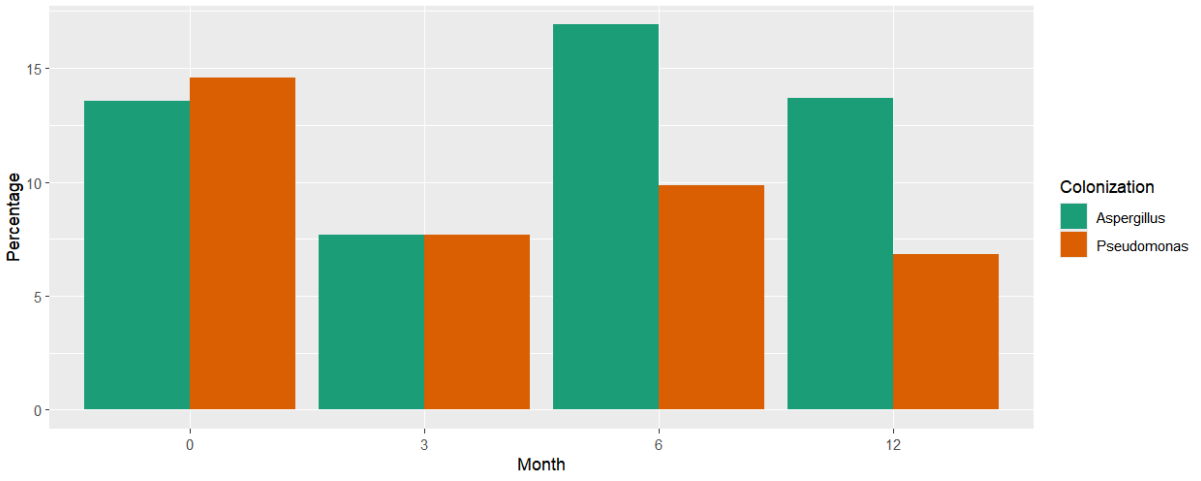


Figure 4.4: *Percentage of patients who show colonization by Aspergillus fumigatus and Pseudomonas aeruginosa over time.*

Differential abundance in taxa

After analyzing the important clinical covariates, we moved on to the results of the models. The ZIBR model was used, implemented both through the `glmmTMB` package and the SAEM algorithm, to detect the influence of the aforementioned covariates on the evolution of the abundance and presence of the 52 bacterial taxa considered in the study. Using a value of $\alpha = 0.05$ for the LRT, `glmmTMB` detected seven taxa, while SAEM detected eight, with the two methods having five taxa in common. The details of these results are shown in Figure 4.5.

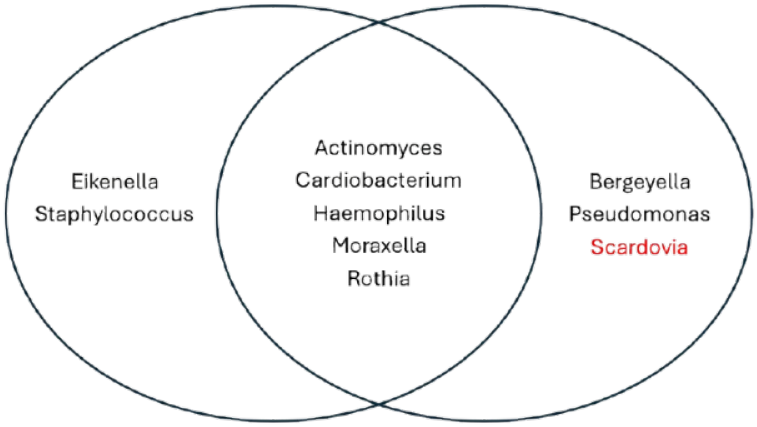


Figure 4.5: *Bacterial taxa detected by `glmmTMB` (left) and SAEM (right) using the ZIBR model.*

We have highlighted the *Scardovia* genus for an important reason. As previously demonstrated in Section 3.4.2, `glmmTMB` encounters convergence issues in specific instances, resulting in an absence of values for the log-likelihood or standard errors of the model. An example of these instances is the model for *Scardovia*. In contrast, SAEM has been demonstrated to be significantly more reliable in this regard. This is primarily due to the implementation of log-likelihood calculation using Importance Sampling (defined in Section 1.4.2), which has proven to be substantially more robust and has never failed throughout the procedures carried out in this thesis. As illustrated in Figure 4.6, the convergence graphs of the ZIBR model estimators for *Scardovia* demonstrate a characteristic evolution that is free of abrupt jumps, a property that is indicative of the SAEM algorithm, and the problem-free convergence to the estimators obtained. It is important to note that the initial values employed for the implementation of ZIBR using SAEM are those determined by `glmmTMB`. Consequently, there is no justification for the methods to exhibit substantial disparities in the likelihood values, or, as in this instance, for one method to encounter failure in the calculation while the other does not.

4.2.3 Conclusions

The findings from this experiment underscore the necessity of employing robust statistical methodologies that are adapted to the intricacies of microbiota data in cystic fibrosis patients. The employment of mixed models through the `glmmTMB` package enabled the identification of five bacterial taxa whose abundance exhibited a significant association

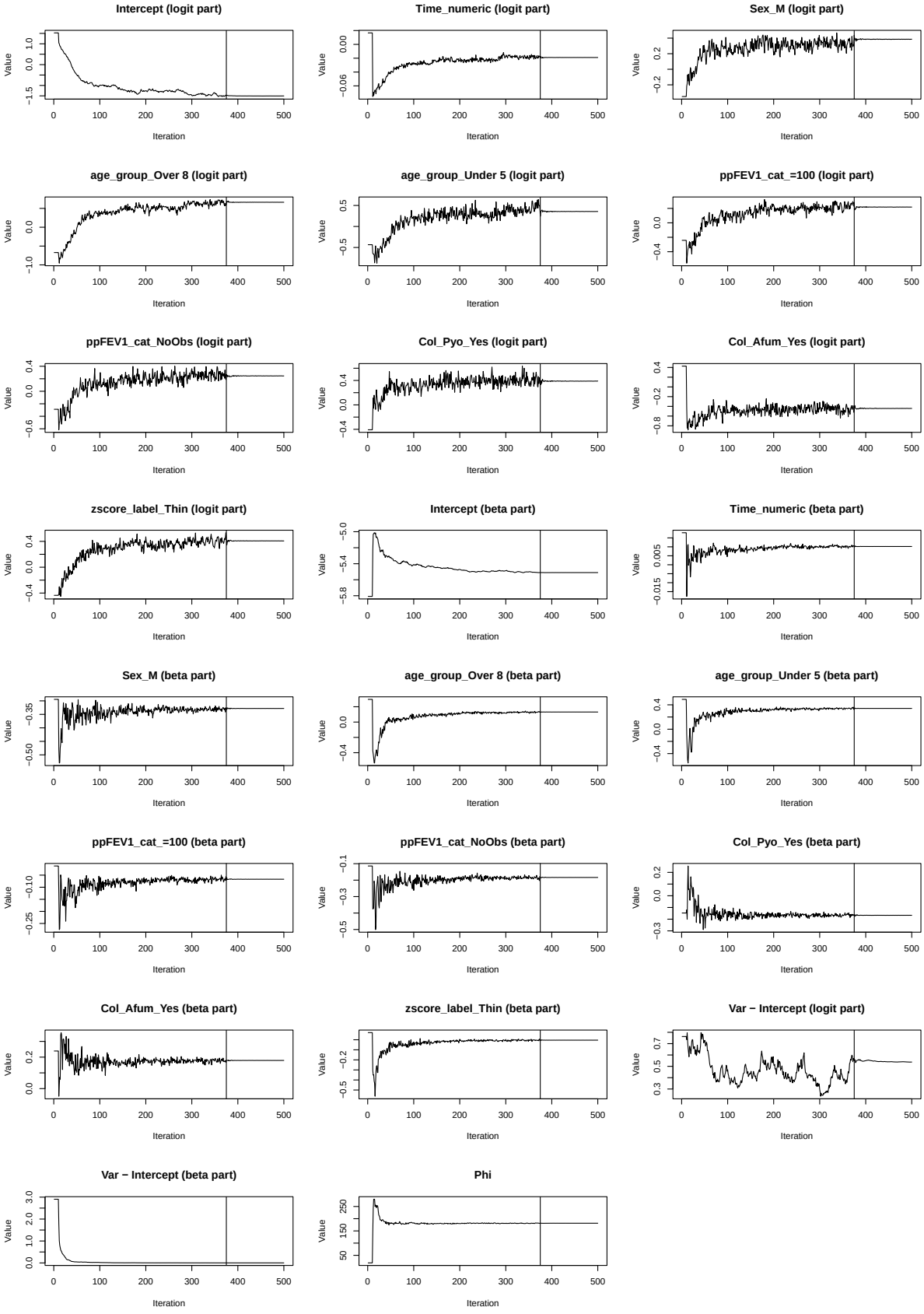


Figure 4.6: *Convergence plots for the ZIBR parameter estimates for the bacterial genus Scardovia.*

with pertinent clinical variables, thereby substantiating the efficacy of these methodologies for the analysis of longitudinal data characterized by overdispersion and inflation of zeros. However, the approach based on the SAEM algorithm seems to demonstrate a superior capacity to explore the parametric space, achieving convergence in scenarios where `glmmTMB` failed. Specifically, SAEM emerged as the sole method capable of detecting the genus *Scardovia*, thereby underscoring the potential for overlooked biological signals when numerical or fitting limitations are present. These findings underscore that the choice of estimation method is not only a technical issue but can directly influence the biological interpretation of the results. Consequently, there is a need to combine advanced statistical tools with biological criteria in order to generate more complete and reliable inferences that may eventually inform monitoring or clinical intervention strategies in cystic fibrosis.

In light of the findings, a number of avenues for future exploration have been identified, which, if pursued, could further refine our understanding of the role of the microbiome in cystic fibrosis. First, it would be valuable to extend current statistical models toward multivariate approaches that allow for the simultaneous capture of dependencies between multiple taxa, while also integrating longitudinal clinical and environmental variables. The incorporation of hierarchical Bayesian models has the potential to enhance the accuracy of inferences when confronted with intricate correlation structures or minimal sample sizes. It is also promising to explore the temporal dynamics of the microbiota using time series models or causal inference approaches that allow distinguishing transition patterns associated with clinical changes. Future research endeavors should consider the integration of multi-omic data, encompassing functional metagenomics, transcriptomics, and immunophenotyping, to construct a more comprehensive framework. This framework would facilitate not only the description of clinical behavior but also the prediction of outcomes based on the microbiome in patients with cystic fibrosis. This approach has already been used with proven results in diseases such as Parkinson's (La Cognata et al., 2021) and cancer (Dimitrakopoulos et al., 2021), so we are confident that its application to cystic fibrosis will also be of great benefit to medical research.

CHAPTER 5

FUTURE DEVELOPMENTS: NEW MODELS AND FRAMEWORKS

This chapter will describe certain models that are considered interesting and, based on the work carried out in this thesis, could be studied in the short term. Specifically, the analysis focuses on two models based on zero-inflated count data, as well as a model that summarizes longitudinal and time-to-event data. These models are motivated by the analysis of human microbiota datasets.

5.1 Models for longitudinal zero-inflated count data

5.1.1 Zero-Inflated Poisson-Gamma (ZIPG) model

This model was proposed by Jiang et al. (2023) as a flexible approach for modeling bacterial count data obtained from standard genetic sequencing methods. It has been specifically designed to address the main characteristics that make this type of data difficult to analyze, including the high proportion of zeros, over-dispersion, and inter-individual variability.

Keeping a bacterial taxon fixed, we define Y_{it} as the count of that taxon for individual i at time t, with $i = 1, \cdots, N$ and $t = 1, \cdots, T_i$. The ZIPG model is then defined as follows:

$$\begin{aligned} Y_{it}|u_{it} &\sim \begin{cases} 0 & \text{with probability } p, \\ Poisson(\lambda_{it} \cdot u_{it}) & \text{with probability } 1-p. \end{cases} \\ u_{it} &\sim Gamma(W_i^{-1}, W_i). \end{aligned} \tag{5.1}$$

where the parameter $0 < p < 1$ is the probability of observing a zero value and is considered constant between individuals and observations, $\lambda_{it} > 0$ is the mean component that models the expected non-zero bacterial abundance, and the overdispersion is modeled through the latent variables u_{it} that depend on the component $W_i > 0$, that explains the

extra dispersion in the data. From this, we can see that the ZIPG model is a hierarchical model that allows the breakdown of the data generation process into three different components to better encapsulate the complexity of the data, unlike traditional approaches that only model the mean of the counts. Furthermore, the authors have permitted the incorporation of covariates in the key parameters of mean and dispersion according to the following schemes:

$$\begin{aligned} \log \lambda_{it} &= \boldsymbol{X}_{it}^{\top} \boldsymbol{\beta} \\ \log W_i &= \boldsymbol{Z}_i^{\top} \boldsymbol{\gamma}, \end{aligned} \tag{5.2}$$

where $\boldsymbol{X}_{it}$ and $\boldsymbol{Z}_i$ are clinical covariates of interest for the mean abundance and the dispersion of the taxon counts, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are its respective regression coefficients. It is worth noting that this model can be seen as a generalization of other models that have also been proposed for the same task, such as the Zero-Inflated Negative Binomial Mixed Model (ZINBMM), which was introduced by Zhang and Yi (2020) and is a special case of the Poisson-Gamma mixture. Nevertheless, it should be clarified that this model offers the possibility of introducing random effects into its approach, something that ZIPG did not consider in its definition and therefore does not allow it to model one of the most important characteristics of microbiota data, which is intra-individual correlation. With regard to the parameter estimation method implemented for this model, the authors have chosen to use the EM algorithm (Dempster et al., 1977). However, other similar models for count data have shown the advantages of incorporating a Bayesian scheme for estimation, such as the zero-inflated discrete Weibull model (Burger et al., 2020) and the zero-inflated Poisson model (Rodrigues-Motta et al., 2010).

5.1.2 Zero Inflated Bell Regression (ZIBell)

Similar to the previous one, this model, proposed by Lemonte et al. (2020), is defined for count data that exhibit overdispersion and zero-inflation. Let Y_i be count data, $i = 1, \cdots, n$. The ZIBell model is then defined by the following equations:

$$Y_i \sim \begin{cases} 0 & \text{with probability } p_i, \\ Bell(\mu_i) & \text{with probability } 1 - p_i. \end{cases} \tag{5.3}$$

where Bell(μ) denotes the Bell distribution * with parameter $\mu > 0$ (Castellares et al., 2018).

One of the properties of the Bell distribution is that its mean (μ) is always smaller than its variance ($\mu(1+W(\mu))$), so it is suitable for use when working with overdispersed data. Furthermore, this distribution depends solely on one parameter, making its estimation more parsimonious than other distributions designed for the same task. Originally, and similarly to ZIPG, the ZIBell model allows the inclusion of covariates to explain the parameters of importance p_i and μ_i:

*The probability density function of a Bell random variable is $P(Y = y) = \exp(1 - \exp W(\mu)) \frac{W(\mu)^y \cdot B_y}{y!}$, where $W(\mu)$ is the Lambert function (Corless et al., 1996) and B_y denotes the Bell numbers (Bell, 1934a,b), defined by $B_y = e^{-1} \sum_{k=0}^{\infty} \frac{k^y}{k!}$.

$$\begin{aligned} \log \mu_i &= \boldsymbol{X}_i^\top \boldsymbol{\beta} \\ \log\left(\frac{p_i}{1-p_i}\right) &= \boldsymbol{Z}_i^\top \boldsymbol{\gamma}, \end{aligned} \tag{5.4}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are clinical covariates of interest and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are their respective regression coefficients.

Although the relative novelty of the ZIBell model, numerous methodological variants have been proposed in its original definition. Recently, Ali and Pho (2024) proposed a change in the link function in p_i parameter regression, switching to the probit function and thus defining the Zero Inflated Probit Bell model (ZIPBell). On the parameter estimation method proposed for ZIBell, the conventional approach has been maximum likelihood, whose properties have been thoroughly examined by Ali et al. (2022). This approach has been maintained even when generalizing the model to the multivariate case (Lemonte, 2022), although there have been some advances in applying shrinkage methods to estimation (Seifollahi et al., 2025), obtaining results that reduce the mean square error of the estimated parameters. On another note, Amani et al. (2025) proposes a marginalized ZIBell model to calculate the effects of covariates on the population mean of counts. However, no method has been proposed that would allow a longitudinal structure to be incorporated into the ZIBell model, given that its definition clearly establishes it as a model that varies only at the individual level, which is too simple to reflect the temporal characteristics of microbiota data.

5.2 Joint models for longitudinal and survival data

Joint models that integrate longitudinal and time-to-event data are a powerful and increasingly relevant statistical tool in biomedical research. Their main strength is their ability to simultaneously model the evolution of repeated biomarkers and their association with relevant clinical events, which improves the accuracy of inferences and captures dynamic relationships between both processes. Although the fundamental idea behind these models is not new (see Tsiatis and Davidian, 2004, which compiles advances to date), with the advancement of personalized and data-driven medicine, they have taken a prevalent position in research. Despite the extensive range of studies on joint modeling, there are not many models that were developed with the specific intention of incorporating microbiota data and its distinctive characteristics. While certain models do provide distributions that could be adapted to microbial abundances, such as negative binomial and Poisson, these can be regarded as special cases of a Poisson-Gamma mixture (Joe and Zhu, 2005), which can better capture the challenging features of microbiota data, as mentioned in previous sections. Wang et al. (2021) proposed a specific model for microbiota count data using the zero-inflated negative binomial distribution and a semi-parametric accelerated failure time (AFT) model for time-to-event information. However, it should be noted that this estimation method comprises two distinct phases for the two different submodels. Consequently, from a methodological perspective, this model cannot be regarded as a joint estimation framework. As a starting point for proposing new joint schemes for longitudinal data and time-to-event information, we present another joint model that was

explicitly designed to incorporate microbiota data in its longitudinal part, the so called **JointMM** (Hu et al., 2022).

5.2.1 JointMM: bacterial compositional data and time-to-event model

The two submodels that define JointMM, longitudinal and survival, connected by a shared latent Gaussian process, are the following:

1. **Longitudinal submodel (ZIS-beta regression):** this component models the presence/absence and observed proportion (when not zero) of a given taxon, for each individual and point in time. Let Y_{it} be the relative abundance of a bacterial taxon in the individual i at time t, $1 \leq i \leq N$, $1 \leq t \leq T_i$. The model assumes that Y_{it} follows the distribution:

$$Y_{it} \sim \begin{cases} 0 & \text{with probability } 1 - p_{it}, \\ \text{ScaledBeta}(u_{it}, \phi, Q) & \text{with probability } p_{it} \end{cases} \tag{5.5}$$

with $\phi > 0$ and $0 < u_{it}, p_{it} < 1$. These two last components are characterized by

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = a_i + \boldsymbol{X}_{it}^{\top}\boldsymbol{\alpha}, \qquad \log\left(\frac{u_{it}}{1-u_{it}}\right) = b_i + \boldsymbol{Z}_{it}^{\top}\boldsymbol{\beta}, \tag{5.6}$$

where a_i and b_i are individual specific intercepts, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of regression coefficients and $\boldsymbol{X}_{it}$ and $\boldsymbol{Z}_{it}$ are covariates for each individual and time point. We further consider that each one of the random intercepts follows a normal distribution, independent from each other:

$$a_i \sim N(a, \sigma_1^2), \quad b_i \sim N(b, \sigma_2^2).$$

The Scaled Beta distribution allows proportions to be bounded within a range less than 1 (e.g., (0,Q]), better suited to microbiota data, which rarely reach absolute proportions. Its density function is defined by

$$f(y_{it}; u_{it}, \phi, Q) = \frac{1}{Q^{\phi-1}} \frac{\Gamma(\phi)}{\Gamma(u_{it}\phi)\Gamma((1-u_{it})\phi)} y_{it}^{u_{it}\phi-1}(Q - y_{it})^{(1-u_{it})\phi-1}.$$

Thus, this submodel is a particular case of the ZIBR model (Chen and Li, 2016), studied in depth in Chapter 2.

2. **Survival submodel (Cox model with random terms):** the time to event is modeled with an extended version of the Cox model, where the random effects of the longitudinal submodel are incorporated directly into the hazard function:

$$h_i(t) = h_0(t) \cdot \exp\left(\boldsymbol{\gamma}^{\top}\omega_i + \delta_1 a_i + \delta_2 b_i\right), \tag{5.7}$$

where $h_i(t)$ is the hazard function for individual i at time t, $h_0(t)$ is the baseline hazard function (e.g., Weibull), ω_i is the vector of clinical covariates for individual i, $\boldsymbol{\gamma}$ represents the coefficients associated with these covariates, a_i and b_i are the random effects of the longitudinal submodel (microbial presence and abundance, respectively), and δ_1 and δ_2 are parameters that quantify the association between the microbial trajectory and the clinical event.

Regarding the estimation of this model, it is based on approximating the likelihood function using Gaussian quadrature. In Chapter 2 we have shown that, in the case of the aforementioned ZIBR model, this method may be inaccurate in certain scenarios, suggesting the use of a stochastic variation of the EM algorithm. In addition, while the ZIBR model and its particular case, the ZIS-beta regression, are useful for studying microbiota expressed as compositional data, McMurdie and Holmes (2014) suggest that using the original count data with an appropriate model offers substantial advantages over data transformations, such as those used in the ZIBR model to express bacterial abundances as proportions.

These three models represent innovative alternatives for modeling that can be further explored along the same lines of research discussed in this paper. Although we can use the SAEM algorithm for estimation tasks, due to the structure of the models and according to other studies, Bayesian estimation is an interesting approach that is used in some similar models (Bhattacharjee et al., 2024; Rustand et al., 2024). Whichever method is chosen, there is still much work to be done in the analysis of microbiota data.

BIBLIOGRAPHY

Abe, T. and Iwasaki, M. (2007). Evaluation of statistical methods for analysis of small-sample longitudinal clinical trials with dropouts. *Journal of the Japanese Society of Computational Statistics*, 20(1):1–18.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Ali, E., Diop, M. L., and Diop, A. (2022). Statistical inference in a zero-inflated Bell regression model. *Mathematical Methods of Statistics*, 31(3):91–104.

Ali, E. and Pho, K.-H. (2024). A novel model for count data: zero-inflated Probit Bell model with applications. *Communications in Statistics-Simulation and Computation*, pages 1–19.

Amani, K. M., Kouakou, K. J. G., and Hili, O. (2025). Marginalized zero-inflated Bell regression models for overdispersed count data. *Journal of Statistical Theory and Practice*, 19(2):17.

Amin, R., Dupuis, A., Aaron, S. D., and Ratjen, F. (2010). The effect of chronic infection with Aspergillus fumigatus on lung function and hospitalization in patients with cystic fibrosis. *Chest*, 137(1):171–176.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–10.

Arribas-Gil, A., Bertin, K., Meza, C., and Rivoirard, V. (2014). LASSO-type estimators for semiparametric nonlinear mixed-effects models estimation. *Statistics and Computing*, 24(3):443–460.

Baldelli, V., Scaldaferri, F., Putignani, L., and Del Chierico, F. (2021). The role of Enterobacteriaceae in gut microbiota dysbiosis in inflammatory bowel diseases. *Microorganisms*, 9(4):697.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68.

Batardière, B., Chiquet, J., Gindraud, F., and Mariadassou, M. (2024). Zero-inflation in the multivariate Poisson lognormal family. *arXiv preprint arXiv:2405.14711*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications*, volume 2. Wiley New York.

Bell, E. T. (1934a). Exponential numbers. *The American Mathematical Monthly*, 41(7):411–419.

Bell, E. T. (1934b). Exponential polynomials. *Annals of Mathematics*, 35(2):258–277.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1):289–300.

Bertrand, J., Comets, E., and Mentre, F. (2008). Comparison of model-based tests and selection strategies to detect genetic polymorphisms influencing pharmacokinetic parameters. *Journal of biopharmaceutical statistics*, 18(6):1084–1102.

Bhattacharjee, A., Rajbongshi, B. K., and Vishwakarma, G. K. (2024). jmBIG: enhancing dynamic risk prediction and personalized medicine through joint modeling of longitudinal and survival data in big routinely collected data. *BMC Medical Research Methodology*, 24(1):172.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Burger, D. A., Schall, R., Ferreira, J. T., and Chen, D.-G. (2020). A robust Bayesian mixed effects approach for zero inflated and highly skewed longitudinal count data emanating from the zero inflated discrete Weibull distribution. *Statistics in Medicine*, 39(9):1275–1291.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75(1):33–57.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*. Number 53 in Econometric Society Monographs. Cambridge University Press.

Cappellato, M., Baruzzo, G., and Di Camillo, B. (2022). Investigating differential abundance methods in microbiome data: A benchmark study. *PLoS Comput Biol*, 18(9):e1010467.

Castellares, F., Ferrari, S. L., and Lemonte, A. J. (2018). On the Bell distribution and its associated regression model for count data. *Applied Mathematical Modelling*, 56:172–185.

Celeux, G., Chauveau, D., and Diebolt, J. (1995). On Stochastic Versions of the EM Algorithm. Research report, INRIA.

Chan, A. W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P. C., Krumholz, H. M., Ghersi, D., and Van Der Worp, H. B. (2014). Increasing value and reducing waste: addressing inaccessible research. *The Lancet*, 383(9913):257–266.

Chen, E. Z. and Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617.

Chiquet, J., Mariadassou, M., and Robin, S. (2021). The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9.

Chiquet, J., Robin, S., and Mariadassou, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171. PMLR.

Comets, E., Karimi, B., Delattre, M., Ranke, J., Lavenu, A., Lavielle, M., Chanel, M., Guhl, M., Fayette, L., and Kaisaridi, S. (2021). Saemix user's guide, version 3.0. https://github.com/iame-researchCenter/saemix/blob/7638e1b09ccb01cdff173068e01c266e906f76eb/docsaem.pdf.

Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an R implementation of the SAEM algorithm. *Journal of Statistical Software*, 80(3):1–41.

Comets, E. and Mentré, F. (2001). Evaluation of tests based on individual versus population modeling to compare dissolution curves. *Journal of Biopharmaceutical Statistics*, 11(3):107–123.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational mathematics*, 5:329–359.

de la Cruz, R., Lavielle, M., Meza, C., and Núñez-Antón, V. (2024). A joint analysis proposal of nonlinear longitudinal and time-to-event right-, interval-censored data for modeling pregnancy miscarriage. *Computers in Biology and Medicine*, 182:109186.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, pages 94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dimitrakopoulos, C., Hindupur, S. K., Colombi, M., Liko, D., Ng, C. K., Piscuoglio, S., Behr, J., Moore, A. L., Singer, J., Ruscheweyh, H.-J., et al. (2021). Multi-omics data integration reveals novel drug targets in hepatocellular carcinoma. *BMC genomics*, 22:1–26.

Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics*, 15:1–8.

D'Agata, A. L., Wu, J., Welandawe, M. K. V., Dutra, S. V. O., Kane, B., and Groer, M. W. (2019). Effects of early life NICU stress on the developing gut microbiome. *Developmental Psychobiology*, 61(5):650–660.

Eggers, J. (2015). *On Statistical Methods for Zero-Inflated Models*. Thesis, Uppsala Universitet.

Elborn, J. S., Shale, D., and Britton, J. (1991). Cystic fibrosis: current survival and population estimates to the year 2000. *Thorax*, 46(12):881–885.

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550.

Florova, V., Romero, R., Tarca, A. L., Galaz, J., Motomura, K., Ahmad, M. M., Hsu, C.-D., Hsu, R., Tong, A., Ravel, J., et al. (2021). Vaginal host immune-microbiome interactions in a cohort of primarily African-American women who ultimately underwent spontaneous preterm birth or delivered at term. *Cytokine*, 137:155316.

Folkesson, A., Jelsbak, L., Yang, L., Johansen, H. K., Ciofu, O., Høiby, N., and Molin, S. (2012). Adaptation of Pseudomonas aeruginosa to the cystic fibrosis airway: an evolutionary perspective. *Nature Reviews Microbiology*, 10(12):841–851.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Fu, X., Li, Y., Meng, Y., Yuan, Q., Zhang, Z., Wen, H., Deng, Y., Norbäck, D., Hu, Q., Zhang, X., and Sun, Y. (2021). Derived habitats of indoor microbes are associated with asthma symptoms in chinese university dormitories. *Environ Res*, 194:110501.

Fu, X., Ou, Z., and Sun, Y. (2022). Indoor microbiome and allergic diseases: From theoretical advances to prevention strategies. *Eco Environ Health*, 1(3):133–146.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC.

Gange, S. J., Muñoz, A., Sáez, M., and Alonso, J. (1996). Use of the beta–binomial distribution to model the effect of policy changes on appropriateness of hospital stays. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 45(3):371–382.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

Gerber, G. K. (2015). Longitudinal microbiome data analysis. In *Metagenomics for microbiology*, pages 97–111. Elsevier.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*, 8.

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016). It's all relative: analyzing microbiome data as compositions. *Annals of epidemiology*, 26(5):322–329.

Han, Y., Baker, C., Vogtmann, E., Hua, X., Shi, J., and Liu, D. (2021). Modeling longitudinal microbiome compositional data: a two-part linear mixed model with shared random effects. *Statistics in Biosciences*, 13:243–266.

Handayani, D., Notodiputro, K. A., Sadik, K., and Kurnia, A. (2017). A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM). *AIP Conference Proceedings*, 1827(1):020033.

Heijerman, H. G., McKone, E. F., Downey, D. G., Van Braeckel, E., Rowe, S. M., Tullis, E., Mall, M. A., Welter, J. J., Ramsey, B. W., McKee, C. M., et al. (2019). Efficacy and safety of the elexacaftor plus tezacaftor plus ivacaftor combination regimen in people with cystic fibrosis homozygous for the F508del mutation: a double-blind, randomised, phase 3 trial. *The Lancet*, 394(10212):1940–1948.

Hossine, Z., Towers, I. N., and Kaehler, B. D. (2025). Network based differential abundance analysis: bridging community interactions and host microbiome dynamics. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 14(1):14.

Hu, J., Wang, C., Blaser, M. J., and Li, H. (2022). Joint modeling of zero-inflated longitudinal proportions and time-to-event data with application to a gut microbiome study. *Biometrics*, 78(4):1686–1698.

Hu, T., Gallins, P., and Zhou, Y.-H. (2018). A zero-inflated beta-binomial model for microbiome data analysis. *Stat*, 7(1):e185.

Huang, C., Gin, C., Fettweis, J., Foxman, B., Gelaye, B., MacIntyre, D. A., Subramaniam, A., Fraser, W., Tabatabaei, N., and Callahan, B. (2023). Meta-analysis reveals the vaginal microbiome is a better predictor of earlier than later preterm birth. *BMC biology*, 21(1):199.

Jiang, R., Zhan, X., and Wang, T. (2023). A flexible zero-inflated Poisson-Gamma model with application to microbiome sequence count data. *Journal of the American Statistical Association*, 118(542):792–804.

Joe, H. and Zhu, R. (2005). Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(2):219–229.

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature communications*, 10(1):5029.

Kloek, T. and van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica*, 46(1):1–19.

Kodikara, S., Ellul, S., and Lê Cao, K.-A. (2022). Statistical challenges in longitudinal microbiome data analysis. *Briefings in Bioinformatics*, 23(4):1–18.

Kroon, S. J., Ravel, J., and Huston, W. M. (2018). Cervicovaginal microbiota, women's health, and reproductive outcomes. *Fertility and sterility*, 110(3):327–336.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4):1020–1038.

La Cognata, V., Morello, G., and Cavallaro, S. (2021). Omics data and their integrative analysis to support stratified medicine in neurodegenerative diseases. *International Journal of Molecular Sciences*, 22(9):4820.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lemonte, A. J. (2022). Multivariate zero-inflated Bell distribution and its inference and applications. *Applied Mathematical Modelling*, 103:543–556.

Lemonte, A. J., Moreno-Arenas, G., and Castellares, F. (2020). Zero-inflated Bell regression models for count data. *Journal of Applied Statistics*.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C., et al. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell Host & Microbe*, 18(4):489–500.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687.

Liu, L., Shih, Y.-C. T., Strawderman, R. L., Zhang, D., Johnson, B. A., and Chai, H. (2019). Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review. *Statistical Science*, 34(2):253 – 279.

Liu, Q., Shepherd, B. E., Li, C., and Harrell Jr, F. E. (2017). Modeling continuous response variables using ordinal regression. *Statistics in medicine*, 36(27):4316–4335.

Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016). The healthy human microbiome. *Genome medicine*, 8:1–11.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15:1–21.

Luo, R. and Paul, S. (2018). Estimation for zero-inflated beta-binomial regression model with missing response data. *Statistics in medicine*, 37(26):3789–3813.

Maki, K. A., Kazmi, N., Barb, J. J., and Ames, N. (2021). The oral and gut bacterial microbiomes: Similarities, differences, and connections. *Biological research for nursing*, 23(1):7–20.

Martin, B. D., Witten, D., and Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The annals of applied statistics*, 14(1):94.

McCullagh, P. and Nelder, J. (1982). *Generalized linear models*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series Chapman and Hall, second edition.

McCulloch, C. E. and Searle, S. R. (2004). *Generalized, linear, and mixed models*. John Wiley & Sons.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4):e1003531.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meza, C., Jaffrézic, F., and Foulley, J.-L. (2007). REML estimation of variance parameters in nonlinear mixed effects models using the SAEM algorithm. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 49(6):876–888.

Meza, C., Osorio, F., and De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, 22(1):121–139.

Middleton, P. G., Mall, M. A., Dřevínek, P., Lands, L. C., McKone, E. F., Polineni, D., Ramsey, B. W., Taylor-Cousar, J. L., Tullis, E., Vermeulen, F., et al. (2019). Elexacaftor–tezacaftor–ivacaftor for cystic fibrosis with a single Phe508del allele. *New England Journal of Medicine*, 381(19):1809–1819.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19.

Mirsepasi-Lauridsen, H. C., Vallance, B. A., Krogfelt, K. A., and Petersen, A. M. (2019). Escherichia coli pathobionts associated with inflammatory bowel disease. *Clinical Microbiology Reviews*, 32(2):e00060–18.

Molenberghs, G. and Verbeke, G. (2006). Longitudinal data analysis. *Wiley StatsRef: Statistics Reference Online*, pages 1–28.

Myers, W. R. (2000). Handling missing data in clinical trials: an overview. *Drug information journal: DIJ/Drug Information Association*, 34:525–533.

Márquez, M., Meza, C., Lee, D.-J., and De la Cruz, R. (2023). Classification of longitudinal profiles using semi-parametric nonlinear mixed models with P-splines and the SAEM algorithm. *Statistics in Medicine*, 42(27):4952–4971.

Najera-Zuloaga, J., Lee, D.-J., and Arostegui, I. (2019). A beta-binomial mixed-effects model approach for analysing longitudinal discrete and bounded outcomes. *Biometrical Journal*, 61(3):600–615.

Nelder, J. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 54(1):273–284.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384.

Nugent, R. P., Krohn, M. A., and Hillier, S. L. (1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of Gram stain interpretation. *Journal of clinical microbiology*, 29(2):297–301.

Ospina, R. and Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623.

Pais, D., Brás, S., and Sebastião, R. (2024). Overcoming the small dataset challenge in healthcare. In *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, pages 497–502. IEEE.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.

Pollock, J., Glendinning, L., Wisedchanwet, T., and Watson, M. (2018). The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. *Applied and environmental microbiology*, 84(7):e02627–17.

Powney, M., Williamson, P., Kirkham, J., and Kolamunnage-Dona, R. (2014). A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*, 15(1):1–11.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833–844.

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., et al. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences*, 108(supplement_1):4680–4687.

Razzaghi, M. (2022). An alternative to the beta-binomial distribution with application in developmental toxicology. *Journal of the Iranian Statistical Society*, 20(1):333–345.

Rigby, R. A. and Stasinopoulos, D. (1996). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6:57–65.

Rigby, R. A. and Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, 54:507–554.

Rodrigues-Motta, M., Gianola, D., and Heringstad, B. (2010). A mixed effects model for overdispersed zero inflated Poisson data with an application in animal breeding. *Journal of Data Science*, 8(3):379–396.

Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Nikita, L., Galuppi, M., Lamont, R. F., Chaemsaithong, P., Miranda, J., et al. (2014). The composition and stability of the vaginal microbiota of normal pregnant women is different from that of non-pregnant women. *Microbiome*, 2(1):1–19.

Romero, R., Theis, K. R., Gomez-Lopez, N., Winters, A. D., Panzer, J. J., Lin, H., Galaz, J., Greenberg, J. M., Shaffer, Z., Kracht, D. J., et al. (2023). The vaginal microbiota of pregnant women varies with gestational age, maternal age, and parity. *Microbiology spectrum*, 11(4):e03429–22.

Rowe, S. M., Miller, S., and Sorscher, E. J. (2005). Mechanisms of disease: cystic fibrosis. *New England Journal of Medicine*, 352(19):1992–2001.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rustand, D., Van Niekerk, J., Krainski, E. T., Rue, H., and Proust-Lima, C. (2024). Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested Laplace approximations. *Biostatistics*, 25(2):429–448.

Salaun-Ferron, C., Oufkir, T., Ouedraogo, N. P., and Avalos, M. (2023). Association between indoor environmental microbiota of living spaces and chronic asthma and respiratory allergies in Europe: A systematic review. WoM 2023 - 4th International World of Microbiome Conference.

Samson, A., Lavielle, M., and Mentré, F. (2007). The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed effects model. *Statistics in Medicine*, 26(27):4860–4875.

Savic, R. M., Mentré, F., and Lavielle, M. (2011). Implementation and evaluation of the SAEM algorithm for longitudinal ordered categorical data with an illustration in pharmacokinetics–pharmacodynamics. *The AAPS Journal*, 13(1):44–53.

Schuckers, M. E. (2003). Using the beta-binomial distribution to assess performance of a biometric identification device. *International Journal of Image and Graphics*, 3(03):523–529.

Seber, G. and Wild, C. (2005). *Nonlinear Regression*. Wiley Series in Probability and Statistics. Wiley.

Seifollahi, S., Bevrani, H., and Algamal, Z. Y. (2025). Shrinkage estimators in zero-inflated Bell regression model with application. *Journal of Statistical Theory and Practice*, 19(1):1–20.

Serrano, M. G., Parikh, H. I., Brooks, J. P., Edwards, D. J., Arodz, T. J., Edupuganti, L., Huang, B., Girerd, P. H., Bokhari, Y. A., Bradley, S. P., et al. (2019). Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. *Nature medicine*, 25(6):1001–1011.

Severgnini, M., Morselli, S., Camboni, T., Ceccarani, C., Laghi, L., Zagonari, S., Patuelli, G., Pedna, M. F., Sambri, V., Foschi, C., et al. (2022). A deep look at the vaginal environment during pregnancy and puerperium. *Frontiers in Cellular and Infection Microbiology*, 12:838405.

Shi, P., Zhang, A., and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):257–261.

Sokal, R. R. and Rohlf, F. J. (1995). *Biometry: the principles and practice of statistics in biological research*. W.H. Freeman, 3 edition.

Sorbara, M. T., Littmann, E. R., Fontana, E., Moody, T. U., Kohout, C. E., Gjonbalaj, M., Eaton, V., Seok, R., Leiner, I. M., and Pamer, E. G. (2020). Functional and genomic variation between human-derived isolates of Lachnospiraceae reveals inter-and intra-species diversity. *Cell host & microbe*, 28(1):134–146.

Srinivasan, S., Liu, C., Mitchell, C. M., Fiedler, T. L., Thomas, K. K., Agnew, K. J., Marrazzo, J. M., and Fredricks, D. N. (2010). Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. *PloS one*, 5(4):e10197.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*. CRC Press, Taylor & Francis Group.

Thukral, A. K. (2017). A review on measurement of alpha-diversity in biology. *Agricultural Research Journal*, 54(1):1–10.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834.

Vandenborght, L.-E., Enaud, R., Urien, C., Coron, N., Girodet, P.-O., Ferreira, S., Berger, P., and Delhaes, L. (2021). Type 2-high asthma is associated with a specific indoor mycobiome and microbiome. *J Allergy Clin Immunol*, 147(4):1296–1305.e6.

Verbeke, G., Molenberghs, G., and Rizopoulos, D. (2010). Random effects models for longitudinal data. In *Longitudinal research with latent variables*, pages 37–96. Springer.

Veščičík, P., Musilová, K., Stráník, J., Štěpán, M., Kacerovskỳ, M., et al. (2020). Lactobacillus crispatus dominant vaginal microbiota in pregnancy. *Ceska gynekologie*, 85(1):67–70.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17:395–416.

Vonesh, E. and Chinchilli, V. M. (1996). *Linear and nonlinear models for the analysis of repeated measurements*. CRC press.

Walther-António, M. R., Jeraldo, P., Berg Miller, M. E., Yeoman, C. J., Nelson, K. E., Wilson, B. A., White, B. A., Chia, N., and Creedon, D. J. (2014). Pregnancy's stronghold on the vaginal microbiome. *PloS ONE*, 9(6):e98514.

Wang, C., Fan, A., Li, H., Yan, Y., Qi, W., Wang, Y., Han, C., and Xue, F. (2020). Vaginal bacterial profiles of aerobic vaginitis: a case–control study. *Diagnostic microbiology and infectious disease*, 96(4):114981.

Wang, J., Reyes-Gibby, C. C., and Shete, S. (2021). An approach to analyze longitudinal zero-inflated microbiome count data using two-stage mixed effects models. *Statistics in Biosciences*, 13:267–290.

Warton, D. I. and Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, 92(1):3–10.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Williams, D. (1975). 394: The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, pages 949–952.

Wooldridge, J. M. (2003). Introductory econometrics. a modern approach. *Mason, OH: South-Western College Publishing*.

Wu, H., Zhang, Y., and Long, J. D. (2017). Longitudinal beta-binomial modeling using GEE for overdispersed binomial data. *Statistics in medicine*, 36(6):1029–1040.

Yang, L. and Chen, J. (2022). A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome*, 10(1):130.

Yang, R., Li, X., Ying, Z., Zhao, Z., Wang, Y., Wang, Q., Shen, B., and Peng, W. (2023). Prematurely delivering mothers show reductions of Lachnospiraceae in their gut microbiomes. *BMC microbiology*, 23(1):169.

Yi, N. (2024). *NBZIMM: Negative Binomial and Zero-Inflated Mixed Models*. R package version 1.0, commit 5e454376e0a84f62cbb61d32bd1713b19bb9b71a.

Zhang, X., Guo, B., and Yi, N. (2020). Zero-inflated gaussian mixed models for analyzing longitudinal microbiome data. *PloS ONE*, 15(11):e0242073.

Zhang, X., Pei, Y.-F., Zhang, L., Guo, B., Pendegraft, A. H., Zhuang, W., and Yi, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Frontiers in Microbiology*, 9:1683.

Zhang, X. and Yi, N. (2020). Fast zero-inflated negative binomial mixed modeling approach for analyzing longitudinal metagenomics data. *Bioinformatics*, 36(8):2345–2351.

Zhang Chen, E. (2023). *ZIBR: A Zero-Inflated Beta Random Effect Model*. R package version 1.0.2.

Zhao, J., Schloss, P. D., Kalikin, L. M., Carmody, L. A., Foster, B. K., Petrosino, J. F., Cavalcoli, J. D., VanDevanter, D. R., Murray, S., Li, J. Z., et al. (2012). Decade-long bacterial community dynamics in cystic fibrosis airways. *Proceedings of the National Academy of Sciences*, 109(15):5809–5814.

Zhu, B., Tao, Z., Edupuganti, L., Serrano, M. G., and Buck, G. A. (2022). Roles of the microbiota of the female reproductive tract in gynecological and reproductive health. *Microbiology and Molecular Biology Reviews*, 86(4):e00181–21.

Zhu, H.-T. and Lee, S.-Y. (2002). Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte-Carlo method. *Statistics and Computing*, 12(2):175–183.

Appendices

APPENDIX A

ADDITIONAL TABLES AND FIGURES FOR CHAPTER 2

Table A.1 shows the adjusted p-values for the variables considered in the model for each bacterial taxon. Figure A.1 shows the convergence behavior of the SAEM algorithm for the ZIBR model in the mentioned taxon, evidencing one of its great advantages, which is not needing a large number of iterations when entering the phase of descent to the final values. Tables A.2 and A.3 show the estimated coefficients for the ZIBR model in each of the bacterial taxa considered in Section 2.4.2, according to Model 1 and Model 2, respectively.

Table A.1: *P-values obtained by the Likelihood Ratio Test based on the SAEM algorithm for the bacterial taxa of the IBD patients data. The p-values were corrected using the Benjamini-Hochberg process to decrease the false discovery rate.*

Species	**Baseline**	**Time**	**Treat**
Bacteroides	0.0000	0.1172	0.2133
Ruminococcus	0.0008	0.1203	0.0033
Faecalibacterium	0.0000	0.2645	0.0009
Bifidobacterium	0.0000	0.1621	0.0008
Escherichia	0.0000	0.0293	0.0308
Clostridium	0.0003	0.2905	0.0934
Dialister	0.0008	0.2227	0.0045
Eubacterium	0.0049	0.0106	0.0114
Roseburia	0.0000	0.1587	0.0708
Streptococcus	0.0170	0.1706	0.0000
Dorea	0.0000	0.3773	0.0605
Parabacteroides	0.0000	0.0989	0.2028
Lactobacillus	0.0000	0.2208	0.0020
Veillonella	0.0000	0.7333	0.0053
Haemophilus	0.0000	0.1961	0.0000
Alistipes	0.0000	0.4010	0.0000
Collinsella	0.0000	0.1228	0.0078
Coprobacillus	0.0000	0.5255	0.3298

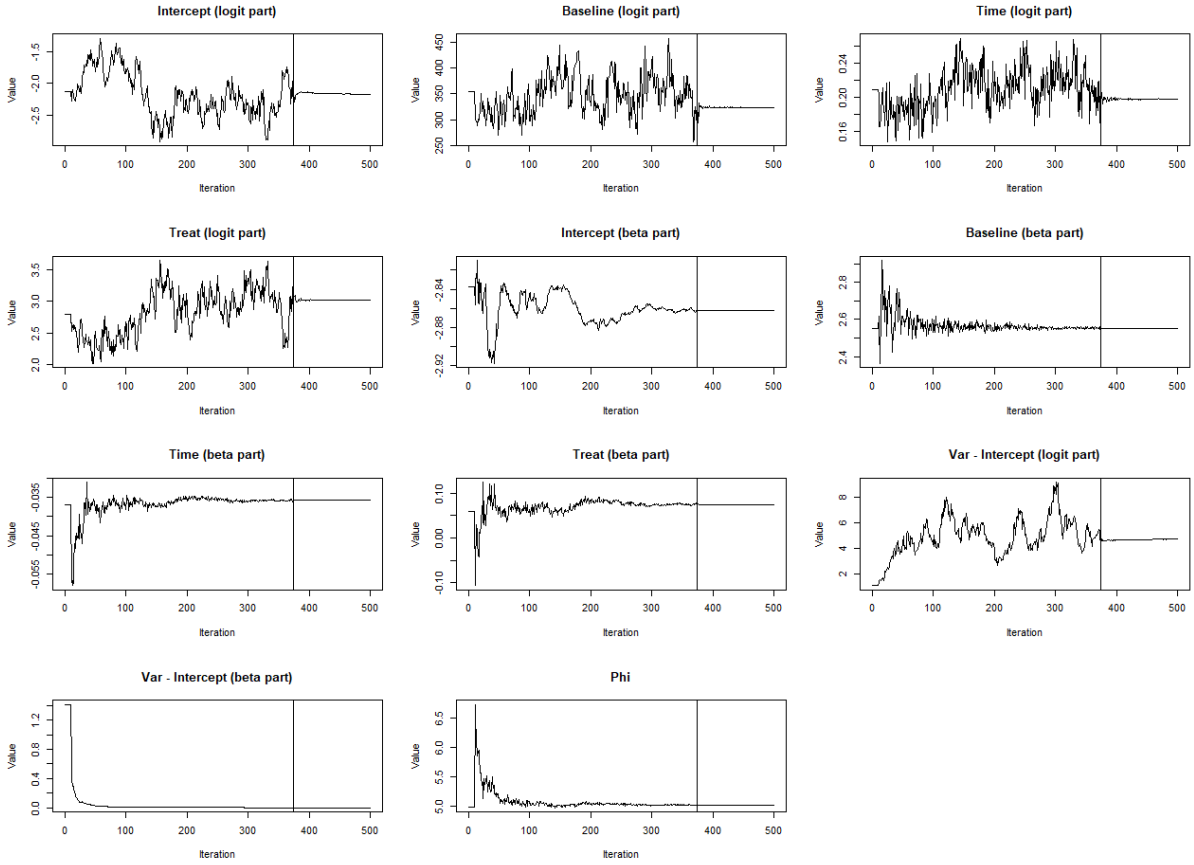


Figure A.1: *Convergence of the ML estimates of the parameters of the ZIBR model for the Escherichia genus calculated by the SAEM algorithm. The SAEM routine was implemented with 5 Markov chains and 500 iterations.*

Table A.2: *ML estimates calculated by the SAEM algorithm for the parameters of Model 1 on the vaginal microbiome data*

	Logistic part			Beta part		
Taxa	**Time**	**Pregnancy**	**Age**[1]	**Time**	**Pregnancy**	**Age**[1]
More present in non-pregnant						
Actinomycetales	0.1643	**-2.9565**	-4.3481	0.2711	**-0.5074**	0.1593
Peptoniphilus	-0.5538	**-3.3886**	-4.9629	0.3549	**-0.5902**	0.4930
Finegoldia magna	0.2397	**-3.0966**	-4.3831	-0.2211	**-0.4315**	-0.0565
Anaerococcus	0.3107	**-3.1733**	-4.4771	-0.0889	**-0.4276**	0.1836
Clostridiales Family XI Incertae Sedis	0.6038	**-5.0301**	-5.4008	-0.2489	**-0.1186**	0.4467
Prevotella	0.0793	**-2.7107**	-4.7978	0.0155	**-0.5715**	0.0004
Anaerococcus vaginalis	0.8104	**-4.6513**	-5.3721	-0.2472	**-0.3762**	0.2780
Prevotella genogroup 2	0.5090	**-1.6274**	-4.0081	0.7375	**-1.2336**	-0.3287
Dialister	2.2496	**-2.9795**	-5.6544	0.4761	**-0.2715**	0.0820
Atopobium	2.0765	**-5.5589**	-5.2281	-0.2565	**-0.5111**	-0.0844
Gardnerella vaginalis	0.6939	**-2.9834**	-4.2448	0.4555	**-0.9249**	-0.4500
Leptotrichia amnionii	0.5094	-2.7212	-3.0925	-0.2628	-0.4553	-0.3020
Bacteroides	0.6935	**-3.2685**	-5.5691	-0.9732	**0.1607**	0.0274
Clostridiales	0.4412	-1.1113	-4.1706	0.6240	-0.8599	0.3289
Streptococcus anginosus	1.1914	**-2.7476**	-3.2483	-0.4074	**-0.3116**	-0.3012
Staphylococcus	-1.2574	-1.8653	-2.9380	-0.4051	-0.2712	-0.5260
Bacteria	-0.8519	-0.0419	-5.5588	-0.2858	-0.2568	0.2760
Peptoniphilus lacrimalis	0.0722	-2.1754	-4.4349	-0.0621	-0.5729	0.1260
Corynebacterium accolens	0.2755	**-3.5430**	-6.4496	-0.1933	**-0.1842**	0.1541
Parvimonas micra	1.7007	-1.9734	-4.0355	-0.5192	-0.1024	0.5413
Porphyromonas	-0.4227	-2.2421	-5.6416	0.4346	-1.0118	-0.3632
Prevotella bivia	0.6292	**-2.7653**	-4.0522	0.5408	**-0.7516**	-0.4583
Bifidobacteriaceae	1.2630	-2.0531	-4.9472	0.2806	-0.1624	-0.1074
Peptoniphilus asaccharolyticus	-0.1290	**-3.6205**	-6.3497	-0.6669	**-0.2086**	-0.0705
Peptoniphilus harei	0.3274	**-2.4012**	-6.4962	-0.4734	**-0.2088**	0.2916
Actinomyces	1.7491	**-2.8264**	-6.0772	0.2213	**-0.5420**	-0.0433
Sneathia	0.7347	-4.7077	-4.9721	-0.7067	-0.3644	-0.7778
Gemella	0.6244	-3.1960	-4.3473	0.0634	-0.3158	0.1018
Coriobacteriaceae	0.7464	**-3.6256**	-5.1390	-0.6120	**-0.2043**	-0.2116
Veillonellaceae	1.3996	**-3.8063**	-6.2298	-0.2779	**0.1731**	-0.2848
Eggerthella	3.9614	**-3.8869**	-4.4172	0.2629	**-1.0546**	-1.0317
Lachnospiraceae	0.4803	-3.3161	-4.0073	-1.3028	-0.0050	-0.7322
Bacteroidales	0.8925	**-2.6188**	-5.7584	-0.5879	**0.1703**	-1.1204
Peptostreptococcus	-1.5022	-2.0752	-3.6386	-0.7837	-0.3178	-0.3577
Aerococcus christensenii	0.8137	**-1.7496**	-4.0424	0.6430	**-0.6755**	-0.7022
Dialister sp type 3	-0.4974	**-2.8009**	-5.9637	-0.5940	**-0.4529**	-0.5076
Mobiluncus curtisii	0.6626	-4.2598	-3.8239	-0.8479	-0.1577	-0.7071
Lactobacillales	-0.0490	**-3.7421**	-6.4767	-0.6701	**-0.1307**	0.4199
Actinomycetaceae	-0.2150	0.0529	-4.8075	0.3671	-0.5665	-0.6179
Anaerococcus tetradius	-1.8136	-0.3787	-4.3929	0.0235	-0.5861	-0.2021
Prevotella genogroup 3	-3.1242	1.4811	-4.4904	-0.1353	0.7275	0.4336
Firmicutes	0.2795	-1.6765	-5.7859	-1.7511	1.1085	-0.4444
Atopobium vaginae	2.6754	**-2.3296**	-1.8858	1.0732	**-1.7974**	-0.7179
Streptococcus	1.4210	-2.3253	-4.1598	-0.6664	-0.1194	-0.6794
Aerococcus	0.6551	**-1.6029**	-5.4446	0.7116	**-1.3571**	-3.0284
BVAB1	-2.0914	2.6161	-3.9562	-0.3836	-0.2357	0.1702
Ureaplasma	0.6215	-1.5126	-5.5155	-1.0830	0.0987	-0.2251
More present in pregnant						
Lactobacillus jensenii	1.1874	2.6978	-2.9872	0.4672	0.1064	-0.6276
Lactobacillus crispatus	1.2249	**2.4304**	-2.1262	0.0366	**1.2101**	1.0443
Lactobacillus vaginalis	-1.2876	3.1935	-5.5519	0.3050	-0.2523	-1.1646
Lactobacillus gasseri	0.1602	0.3210	-2.8409	-0.8749	0.3214	-0.1971
Lactobacillus	2.8199	0.5166	-2.7068	-0.3749	-0.5104	-1.0172
Lactobacillus iners	3.0629	-0.2200	-0.4422	-0.2367	0.4068	-0.8884
Prevotella genogroup 1	-1.4204	3.3757	-2.3026	-1.4471	-0.7525	-2.6299
Megasphaera sp type 1	0.4711	1.2356	-3.7876	0.3543	-0.5830	-0.2166
Sneathia sanguinegens	-0.3651	**2.9115**	-3.0625	-0.1183	**-1.3089**	-1.6596
Proteobacteria	-0.3355	0.8134	-6.0708	0.2224	-0.2306	0.0222

Note: Bold coefficients represent a statistically significant variable for the corresponding taxa according to LRT ($\alpha = 0.05$).

[1] Variable scaled to $[0, 1]$.

Table A.3: *ML estimates calculated by the SAEM algorithm for the parameters of Model 2 adjusted on the vaginal microbiome data*

	Logistic part				Beta part			
Taxa	**Time**	**Pregnancy**	**Age**[1]	**Interaction**	**Time**	**Pregnancy**	**Age**[1]	**Interaction**
More present in non-pregnant								
Actinomycetales	0.3948	-2.2550	1.2636	-4.5273	0.2574	-0.2449	0.4723	-0.1987
Peptoniphilus	0.3700	0.2810	0.9554	**-5.1428**	0.4979	-0.0130	0.7428	**-0.8741**
Finegoldia magna	0.9772	-0.3225	-0.6225	**-4.4261**	-0.1683	-0.4308	-0.0207	**0.0193**
Anaerococcus	0.8629	-0.6215	1.8259	-4.6912	0.0549	0.0580	0.4514	-0.5976
Clostridiales Family XI Incertae Sedis	0.7207	-10.1685	1.4287	-5.5151	-0.1724	-0.0642	0.5863	0.0701
Prevotella	0.4429	-1.3821	1.6417	-4.8826	0.1590	-0.1239	0.1221	-0.8141
Anaerococcus vaginalis	0.7951	-4.0813	-0.0944	-5.4910	-0.1770	-0.1281	0.4443	-0.1729
Prevotella genogroup 2	0.9482	-0.5006	2.2294	-4.4230	0.9075	-0.3012	0.3405	-1.2531
Dialister	2.4420	-2.6952	0.3451	-5.7329	0.6666	0.2014	0.1664	-0.9042
Atopobium	3.3311	-3.1559	-2.0620	**-5.4275**	-0.1449	0.0914	0.2228	**-0.7851**
Gardnerella vaginalis	0.6400	-2.7894	-0.8232	-4.3367	0.5297	-0.7076	-0.2064	-0.4432
Leptotrichia amnionii	1.4334	-0.6080	-0.8310	-3.2125	-0.2064	-0.0218	-0.0096	-0.5715
Bacteroides	0.7602	-5.7682	2.1606	-5.7675	-0.8673	-7.3757	0.2836	9.7135
Clostridiales	0.6560	-0.9829	1.4696	**-4.3330**	0.9404	1.0325	0.4875	**-2.7669**
Streptococcus anginosus	2.0122	-0.3166	0.9938	-3.3596	-0.3247	-0.3774	-0.2174	0.1301
Staphylococcus	-1.5362	-3.6995	-0.6303	-2.9711	-0.4023	-0.5308	-0.4955	0.3844
Bacteria	-0.6773	0.2644	2.3964	-5.5534	-0.4084	-0.5027	0.2986	0.4826
Peptoniphilus lacrimalis	-0.0278	-2.0074	3.6833	-4.6556	-0.0100	1.4279	0.4354	-2.5014
Corynebacterium accolens	0.5597	-0.7272	0.1489	-6.5660	-0.1154	-0.5092	0.2607	0.6831
Parvimonas micra	1.4761	-2.9152	2.2397	-4.3283	-0.5811	0.0751	1.1713	0.0643
Porphyromonas	-0.1872	1.5996	2.2892	-5.8198	0.4243	-0.6574	-0.1062	-0.2790
Prevotella bivia	1.6976	0.5426	-1.3359	**-4.1394**	0.6937	-0.3713	-0.3830	**-0.7431**
Bifidobacteriaceae	1.7009	-1.9078	-0.6436	-5.0737	0.2531	-0.1764	0.1672	0.0999
Peptoniphilus asaccharolyticus	-0.0138	-5.6375	1.3374	-6.6523	-0.5590	-0.1947	0.3948	0.0481
Peptoniphilus harei	0.7404	1.5115	1.1492	-6.6702	-0.3795	-0.0607	0.5226	-0.0205
Actinomyces	1.3715	**-7.1737**	0.7501	-6.2155	0.3464	**-0.4793**	0.1216	-0.0661
Sneathia	1.1231	0.1323	0.1920	-5.1472	-0.6467	-0.5812	-0.4221	0.4894
Gemella	1.4290	0.5889	-0.2813	**-4.4960**	0.3385	0.2498	0.1827	**-1.1261**
Coriobacteriaceae	1.0181	-0.9975	1.5135	-5.2714	-0.5296	-0.0980	-0.0432	0.0134
Veillonellaceae	2.0141	3.2373	1.9231	**-6.2525**	-0.3010	4.9707	-0.2761	**-10.8236**
Eggerthella	7.0001	1.3307	0.1952	**-4.6386**	0.7614	0.8781	-0.8357	**-3.1145**
Lachnospiraceae	0.4188	-6.9168	1.3680	-4.1897	-1.1530	-0.0616	-0.5525	0.0708
Bacteroidales	0.7965	-4.1245	2.7556	-6.2071	-0.3769	3.4525	-0.4391	-4.0362
Peptostreptococcus	-0.7720	0.7950	0.5703	-3.7720	-0.6830	-0.4789	-0.2520	0.3792
Aerococcus christensenii	1.1763	-1.3857	-1.3277	-4.2673	0.7233	-0.5886	-0.3450	0.0167
Dialister sp type 3	-0.2289	-1.6974	0.2424	-6.1968	-0.4251	-0.2131	-0.2086	-0.2637
Mobiluncus curtisii	0.3987	-4.1474	1.6940	-4.1145	-1.0627	0.1107	-0.1854	0.2001
Lactobacillales	-0.0543	-4.3852	0.8537	-6.4998	-0.5817	-0.1442	0.4201	0.0153
Actinomycetaceae	0.2268	1.2868	2.9645	-5.0411	0.3207	-1.4314	-0.2794	1.5222
Anaerococcus tetradius	-1.2706	1.1320	1.5860	-4.6129	0.1162	-0.0854	0.0456	-0.7397
Prevotella genogroup 3	-1.7386	**5.3548**	3.9245	-4.4041	-0.1681	**1.4938**	0.3868	-1.8901
Firmicutes	0.1821	-3.8383	3.1218	-5.9893	-1.6163	-3.2600	-0.1226	5.5245
Atopobium vaginae	3.3931	-1.8104	-0.2133	-2.2539	1.5484	-0.1086	-0.2015	-2.3932
Streptococcus	2.2390	-0.7838	-2.3284	-4.1217	-1.0025	-1.1272	-0.6405	1.8209
Aerococcus	1.0920	**-0.1265**	0.3653	**-6.4798**	0.4889	**-1.1422**	-0.9858	**1.0428**
BVAB1	-2.4305	1.5174	0.0303	-4.0814	-0.2639	0.2136	0.2991	-0.5858
Ureaplasma	0.4596	-1.5496	-2.9635	-5.5847	-1.0195	-0.0987	-0.0569	0.2484
More present in pregnant								
Lactobacillus jensenii	-0.4242	**1.1176**	-1.0204	**-3.1727**	-0.7723	**-1.3386**	-0.1073	**3.3769**
Lactobacillus crispatus	1.5798	**2.9938**	1.7026	-1.9845	0.0106	**0.7945**	0.8268	0.4186
Lactobacillus vaginalis	-3.5236	-2.9340	-5.3811	**-6.1216**	-1.3030	-0.7328	0.3448	**3.0030**
Lactobacillus gasseri	-1.2666	-1.9058	-1.5392	**-2.8785**	-0.9620	0.0504	-0.0161	**0.2846**
Lactobacillus	2.6612	**1.1124**	-2.7797	-2.7404	-0.6025	**-1.0576**	-0.8576	1.0133
Lactobacillus iners	2.9543	-0.1484	-2.8471	-0.4604	-0.2748	0.3779	-0.8028	0.0509
Prevotella genogroup 1	1.4486	**5.5025**	1.2280	**-2.9843**	-0.7541	**0.5118**	-1.8807	**-1.8349**
Megasphaera sp type 1	3.2792	**3.7197**	0.4440	**-3.9302**	0.4912	**-0.6149**	-0.1020	**0.0530**
Sneathia sanguinegens	0.9750	**4.0012**	3.2933	-3.1000	-0.0251	**-1.1921**	-1.6335	-0.2152
Proteobacteria	-1.3540	-0.9926	0.9526	**-6.1615**	0.2327	-0.6051	0.1561	**0.5710**

Note: Bold coefficients represent a statistically significant variable for the corresponding taxa according to LRT ($\alpha = 0.05$).
[1] Variable scaled to $[0, 1]$.