Hindawi
Complexity
Volume 2022, Article ID 4907527, 16 pages
https://doi.org/10.1155/2022/4907527

Research Article

Bayesian Hierarchical Compositional Models for Analysing Longitudinal Abundance Data from Microbiome Studies

I. Creus Martí,[1,2] **A. Moya**,[1,3,4] **and F. J. Santonja**[2]

[1]*Instituto de Biología Integrativa de Sistemas (I2Sysbio), Universitat de València-CSIC, Valencia, Spain*
[2]*Departamento de Estadística e Investigación Operativa, Universitat de València, Valencia, Spain*
[3]*Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO), Valencia, Spain*
[4]*CIBER en Epidemiología y Salud Pública (CIBEResp), Madrid, Spain*

Correspondence should be addressed to F. J. Santonja; francisco.santonja@uv.es

Received 14 March 2022; Accepted 15 June 2022; Published 30 August 2022

Academic Editor: Chittaranjan Hens

Gut microbiome plays a significant role in defining the health status of subjects, and recent studies highlight the importance of using time series strategies to analyse microbiome dynamics. In this paper, we develop a Bayesian model for microbiota longitudinal data, based on Dirichlet distribution with time-varying parameters, that take into account the compositional paradigm and consider principal balances. The proposed model can be effective for predicting the future dynamics of a microbial community in the short term and for analysing the microbial interactions using the value of the estimated parameters. The usefulness of the proposed model is illustrated with six different datasets, and a comparison study with four alternative models is provided.

1. Introduction

The microbiome is defined in [1] as "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space." Recent studies suggest that gut microbiome plays a significant role in defining the health status of subjects. In fact, human gut microbiome influences brain function and behavior [2] and is related to several diseases such as cancer [3], HIV [4], Alzheimer [5], type 2 diabetes [6], cardiovascular disease [7], irritable bowel syndrome (IBS) [8], Crohn's disease [9], chronic fatigue syndrome (CFS) [10], obesity [11], and malnutrition [12], among others. What is more, in [13], the paramount interest of gut microbiome, pointing out the importance of microbes in regulating most if not all human bodily functions, has been underscored. As a result, monitoring human microbiome will probably lead us to improve human health, as has been mentioned in [14].

Given the importance of the microbiome, a concerted effort is being made to design models that enable us to extract information on bacterial dynamics. In [15], Martí et al. explain that temporal microbiota stability is linked to health status. In [16], Dinleyici et al. observed that a disruption in the stable state of the microbiome is related to a disease, but the stable state is reached again when the subject is cured. These facts highlight the importance of using time series strategies to study the microbiome. In that respect, longitudinal studies are an appropriate approach to tackle microbiome modeling, as they can help to gain a better understanding of microbiota regularities, bacterial interactions, and microbiome response to perturbations. Furthermore, this longitudinal approach allows us to predict future microbiome dynamics and analyse not only microbiome stability but also the time it needs to recover and reach a new stable state, if possible.

At present, there are studies where the microbiome is manipulated in humans in order to treat certain illnesses [17]. This manipulation is carried out by fecal microbiota transplantation, probiotic supplementation, or a combination of both strategies [17]. Nowadays, there are also nutrition-based approaches to treatment [18]. For instance, the work detailed in [19] points out that the administration of

Lactobacillus delbrueckii and antibiotics to treat bacterial vaginosis can provide a cure, [20, 21] indicate that fecal microbiota transplantation is a promising therapy for *C. difficile* infection, and [22] reports that fecal microbiota transplantation in HIV patients attenuates HIV-associated dysbiosis. Taking this research approach into account, modeling bacterial dynamics should be a research priority because these models will help us to design better interventions, as has been mentioned in [17].

One line of action in longitudinal analysis to model the abundance of each microbial entity at each time point is to use generalized Lotka–Volterra equations [23–26]. We must mention that these models have usually been used in a deterministic way; however, [27] takes stochastic processes into account, using multiplicative noise with normal distribution. These approaches consider pair-wise interactions, and consequently, they do not take into account the effect that a third microbe may have on an interacting pair of microbes. In [28], more limitations of using the Lotka–Volterra structure have been pointed out. In [29, 30], a nonparametric approach with an additive structure has been proposed. The advantage of these models is the absence of microbial relationship specifications, while the drawback is the fact that the additivity of relationships, allowed in this approach, may not be realistic for complex microbial communities. Other proposals consist of using state space models [31] or linear mixed models [32, 33]. In [34], and references therein, more proposal to analyse microbiota longitudinal data is detailed.

Recent investigations consider compositional data analysis [35] is also a valid approach to analyse microbiome high-throughput sequencing data, and therefore, microbiome datasets can be converted to relative abundance values, or normalized counts, prior to analysis (see, for instance, [36] or Chapter 3 in [34]). Fundamentally, this approach uses the log-ratio transformation techniques to convert microbiome data, thus removing the compositional constraints to make the standard techniques suitable for analysis. This transformation focuses on removing the compositional constraint: all microbial relative abundances sum to one. Microbiome data can be considered as a compositional dataset due to limitations of the techniques used to collect the data [36]. These limitations are based on the fact that datasets collected by 16S rRNA gene amplimers have an arbitrary total imposed by the instrument. This total is known as sequencing depth or library size. This implies that an increase in the abundance of one taxon results in a decrease of some of the other taxa because the sum of all of the raw abundances cannot exceed the sequencing depth. As the increase of a taxon produces a decrease in some of the other taxa, the correlation between the taxa is bogus. The work presented in [37] specified that the correlations between compositional components are spurious. This fact was described under various headings: the constant-sum problem, the closure problem, the negative bias problem, or the null correlation difficulty [38]. These characteristics of compositional data require that the analysis of this type of data must be conducted using appropriate methodology.

There are approaches that analyse microbiome taxa taking its compositional nature into account. In [39], Kynčlová et al. model the isometric log-ratio transformation of the data using Gaussian distribution and a VAR structure, while in [40] phylogenetic isometric log-ratio transformation, to allow off-the-shelf statistical tools to be optimally applied to microbiota datasets, is introduced. This approach takes into account phylogenies to construct the ILR transform and redefine data considering coordinates termed balances. Rivera-Pinto et al. in [41], take also into account balances. In this case, to identify microbial signatures and consider them as predictive covariates of a phenotype of interest, Joseph et al. describe microbial dynamics using generalized Lotka–Volterra (gLV) equations and compositional data analysis (CODA) [42]. In [43, 44], Bayesian CODA methods for microbiome time series which model the counting process and perform inference and apply log-ratio transformations to the model parameters are detailed.

Our proposal is based on modeling a normalized transformation of the observed counts. We assume that normalized transformation of the observed counts can be explained by an autoregressive structure considering Dirichlet distribution with time-varying parameters that take into account principal balances. Principal balances are a compositional tool that enables us to reduce the number of parameters of the model by selecting the balances that maximize the variance [45]. As balances with more variability will better explain the variability of the data, this formulation allows us to encapsulate information and decrease the number of parameters to be estimated. Our proposal is in line with [46], though this one is not used with microbiome datasets. We must mention that the Dirichlet distribution has been considered inappropriate for modeling compositional data due to its strong implied independence structure, which can be deduced from its definition by a set of independent, gamma-distributed, random variables, with equal scale parameter. This fact makes it inappropriate to consider this probability distribution for modeling compositional data [35]. However, it has also been found useful when used as a conditional distribution (see, for instance, [47, 48]). Our approach is presented as an alternative for predicting the future dynamics of a microbial community in the short term and for analysing the microbial interactions using the value of the estimated parameters.

2. Some Basic Principles of CODa

The first steps towards developing a suitable methodology to analyse compositional data focused on projecting the sample space of compositional data (the simplex) to real space using log-ratio transformations [35]. The additive log-ratio transformation (alr) and centered log-ratio transformation (clr) were defined in [35] and are defined by the following expressions:

$$\begin{aligned}\operatorname{alr}(\mathbf{x}) &= \left(\log\left(\frac{x_1}{x_D}\right), \log\left(\frac{x_2}{x_D}\right), \ldots, \log\left(\frac{x_{D-1}}{x_D}\right)\right),\\ \operatorname{clr}(\mathbf{x}) &= \left(\log\left(\frac{x_1}{g(x)}\right), \log\left(\frac{x_2}{g(x)}\right), \ldots, \log\left(\frac{x_D}{g(x)}\right)\right),\end{aligned} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, y_D)$ denotes a D-composition and $g(\cdot)$ denotes the geometric mean. Note that when projecting the simplex to the real space, the clr transformation preserves the distances [49]; see equation (2), where $d_a(\mathbf{x}, \mathbf{y})$ denotes the Aitchison distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ and $d_e(\cdot, \cdot)$ denotes the Euclidean distance.

$$\begin{aligned}d_a^2(\mathbf{x}, \mathbf{y}) &= \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log\frac{x_i}{x_j} - \log\frac{y_i}{y_j}\right)^2\\ &= d_e^2(clr(\mathbf{x}), clr(\mathbf{y})).\end{aligned} \tag{2}$$

2.1. Balances. Balances are defined in [50]. Using the notation applied in [45], they are expressed as follows:

$$b = \sqrt{\frac{r \cdot s}{r + s}} \cdot \ln\left(\frac{g(\mathbf{x}_+)}{g(\mathbf{x}_-)}\right), \tag{3}$$

where $\mathbf{x}_+$ and $\mathbf{x}_-$ are two nonoverlapping groups of parts of a D-composition, r and s are the number of parts in $\mathbf{x}_+$ and $\mathbf{x}_-$, respectively, and $g(\cdot)$ denotes the geometric mean.

Balances are a compositional tool that makes it possible to study the relationship between the components of the numerator and the components of the denominator of the balance. Balances that are near zero indicate similarity between the two groups; however, the greater the absolute value of the balance, the lager is the dissimilarity between them.

Depending on which components we put in $\mathbf{x}_+$ and which we put in $\mathbf{x}_-$ we will have different balances. In this study, we want to choose the components of $\mathbf{x}_+$ and the components of $\mathbf{x}_-$ that maximize the variance of the balance because balances with more variability will better explain the variability of the microbiome data. Principal balances are the strategy selected to tackle this issue.

2.2. Principal Balances. Principal balances act as the Principal Component Analysis (PCA) for compositional data. In other words, they allow us to reduce the dimension with a minimum loss of information. In addition, the numerator and the denominator of the principal balances are selected so as to maximize the variance of the balance [45].

In order to obtain the principal balances, Ward's clustering method [51] is used (see details in [52]). We apply this method because it enhances the proportionality between groups of parts and reduces the computation time [52].

3. Model

3.1. Selecting Principal Balances. Let $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{Dt})$ be the relative abundances at time t, so that $y_{it} \in (0, 1)$ and $\sum_{i=1}^{D} y_{it} = 1$. Note that y_{it} denotes the relative abundance of taxon i at time point t. Let T be the number of time points available in the data. In order to implement Ward's clustering method, we denote as $\mathbf{S}_i$ the vector that contains the clr transformation of the relative abundance of species i at all time points available in the data. As a result,

$$\mathbf{S}_i = \left(\log\left(\frac{y_{i1}}{g(y_1)}\right), \log\left(\frac{y_{i2}}{g(y_2)}\right), \ldots, \log\left(\frac{y_{iT}}{g(y_T)}\right)\right) \text{with } i = 1, 2, \ldots, D. \tag{4}$$

First of all, we calculate the matrix $M = (m_{ij})_{(D-1)x(D-1)}$ where $m_{ij} = d_e(\mathbf{S}_i, \mathbf{S}_j)$. This matrix is proportional to the variation matrix due to the following equation defined by [52].

$$\operatorname{Var}\left[\log\left(\frac{\mathbf{z}_i}{\mathbf{z}_j}\right)\right] = \frac{1}{T}d_a^2\left(\mathbf{z}_i, \mathbf{z}_j\right) = \frac{1}{T}d_e^2\left(\mathbf{S}_i, \mathbf{S}_j\right). \tag{5}$$

In this equation, $\mathbf{z}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ denotes the vector that contains the relative abundance of species i at all time points available in the data. Before going into further detail, it is important to point out that Ward's method distance between two clusters B and C is defined by equation (6), where $\overline{C}$ and $\overline{B}$ denote the centroids (i.e., barycenters) of C and B, respectively.

$$d(C, B) = \frac{2|C\|B|}{|C| + |B|}d_e(\overline{C}, \overline{B}). \tag{6}$$

Second, we apply Ward's method, because according to [52], the variance of a balance is proportional to Ward's method distance between the numerator and the denominator of the balance. Ward's method is hierarchical, and it detects the smallest entry in the matrix M (which functions as the variation matrix), and the corresponding parts are merged to form a group. The matrix M is recalculated at each iteration using equation (6) and, the final stage of the process, the last two remaining groups are fused into one, which gives the balance with the largest variance. We have implemented this process using the *hclust* function included in the *fastcluster* package in *R* [53].

During the hierarchical process, we generate a dendrogram that plots the groups created in each iteration of the method. As a result, the groups merged in the dendrogram give us the numerator and the denominator of the principal balances. In conclusion, through this process, we have applied the new approach to Ward's clustering method in

order to obtain the principal balances, which are denoted as $P\text{Bal}_1, P\text{Bal}_2, \ldots, P\text{Bal}_{D-1}$.

Finally, we select the principal balances for which the sum of the percentage of variance is higher than 80%. We choose this percentage because we consider that taking account of the balances that describe at least 80% of the variability is enough for our model to capture the variability of the data [54]. We denote the selected principal balances as $SP\text{Bal}_1, SP\text{Bal}_2 \ldots, SP\text{Bal}_M$, where M denotes the total amount of SPBal. We must mention that a detailed example of the process described in this section to obtain the selected principal balances is provided in Section 5.

3.2. Proposed Model (PM). We assume that $\mathbf{y}_t$ follows a Dirichlet distribution with nonnegative parameters $\alpha_t = (\alpha_{1t}, \alpha_{2t} \ldots, \alpha_{Dt})$. Reference [46] presents the ability of the Dirichlet distribution with an autoregressive structure that uses the alr or the clr transformation as a link function and imposes a parameter that is equal to the sum of the Dirichlet parameters. Our proposal is in line with this because it uses the Dirichlet distribution and an autoregressive structure; however, we use a Bayesian framework with the logarithm as a link function, we introduce the principal balances in the autoregressive structure, and we do not need to estimate an extra parameter that is equal to the sum of the Dirichlet parameters.

We model each transformed parameter $\log(\alpha_{it})$ as a linear function of the selected principal balances:

$$\log(\alpha_{it}) = a_{i0} + \sum_{j=0}^{M} a_{ij} \cdot \text{SPBal}_{jt-1}. \tag{7}$$

Note that $\text{SPBal}_{j,t-1}$ are calculated using the values of $\widehat{y}_{i,t-1}$, and the parameters of the model are a_{ij}. The Gaussian distribution with zero mean and standard deviation σ_{ij} is considered as a prior distribution of a_{ij}. As [55] suggests, σ_{ij} is assigned the uniform distribution in the interval $[0, 5]$. Note that $\alpha_{it} = \exp(a_{i0} + \sum_{j=0}^{M} a_{ij} \cdot \text{SPBal}_{jt-1})$ and then $\text{E}(y_{it})$ can be estimated by $\text{E}(y_{it}) = \alpha_{it} / \sum_{i=1}^{D} \alpha_{it}$

As we have mentioned before, the numerator and the denominator of the principal balances are selected so as to maximize the variance of the balance. As balances with more variability will better explain the variability of the data, principal balances enable us to encapsulate information and decrease the number of parameters to be estimated. Balances provide information about the similarity between the components of the numerator and the elements of the denominator: balances that are approximately zero indicate that the two groups are similar; however, the larger the absolute value of the balances, the greater the dissimilarity between them.

4. Datasets

We demonstrated the ability of our model by using both publicly available and simulated datasets. We studied the gut microbiome time series data of two healthy subjects, one female and one male. In both subjects, the samples were collected daily. The data of the female subject were extracted from [56] and were analysed on different taxonomic levels. We used Subject C in [57] as the male subject, taking into account that this subject is obese class I according to his body mass index.

We simulated three databases with fifteen, twenty, and sixty microbial taxa, respectively. Following the scheme given by [58], we generated the interaction matrix using the algorithm proposed by [59], and we generated the initial abundances using the Poisson distribution. With these two tools, we simulated the data using the generalized Lotka–Volterra structure. We carried out the simulation using the *R* package *seqtime* [58]. Focusing on technical details, to generate the interaction matrix we set the clique size at 10, the diagonal values at −1, the interaction connectance at 0.01, the positive edge percentage at 64%, and the maximal absolute off-diagonal interaction strength at 1.

Microbial datasets are zero-inflated [60]; nevertheless, zeros in a compositional data set are below detection limit values [61]. For this reason, in order to analyse the publicly available datasets, we selected the bacterial taxa that did not have zeros at any time point, and we gathered the information of the nonselected taxa in the *Other* collection.

5. Results

First of all, we will explain in detail how we applied Ward's clustering method to obtain the principal balances and present a comparative study on selecting the principal balances. Second, we analyse the information that estimated parameters provide on bacterial dynamics. Third, we offer a study on the presence of zeros at credible intervals. After this, we examine the graphic results and test the results when modeling datasets with a large number of microbial taxa. Finally, we compare the proposed model with other approaches.

5.1. Applying Ward's Clustering Method and Selecting the Principal Balances. In order to introduce the results in an accessible way, we first present the outcomes obtained with the publicly available datasets. Moreover, we start by analysing the female dataset at family level, because it has aspects in common with the other two datasets and, consequently, it is easily comparable.

By applying Ward's clustering method as it is described in Section 3.1, we obtain the dendrogram shown in Figure 1. As it is presented by [62], the horizontal bars of the dendrogram describe the groups of parts defined at each iteration of the process; for this reason, the groups merged in the dendrogram give us the numerator and the denominator of the principal balances. Moreover, the sum of all vertical bars represents the total variance of the sample; therefore, a long vertical bar implies a balance that explains a good deal of the cc variance and a short vertical bar means that balance has a little variability in the sample. In Figure 1, we can also see the principal balances described by the dendogram and their percentage of variance (Table1).

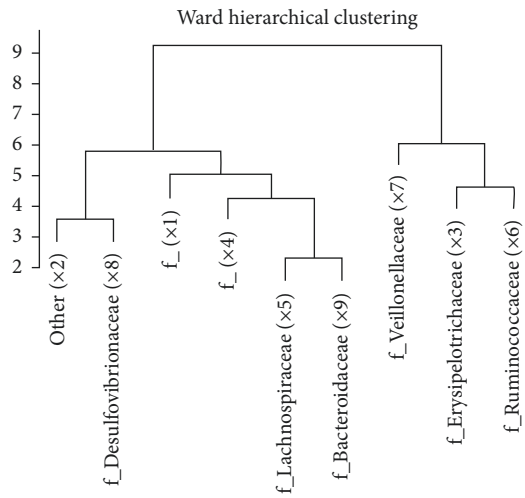


Figure 1: Female dataset at family taxonomic level. Dendogram obtained using Ward's method for obtaining the principal balances. The table shows the expression of the principal balances and their percentage of variance.

As explained in Section 3.1, we select the principal balances that accounted for at least 80% of the variability. Consequently, we pick PBal_8, PBal_2, PBal_7, PBal_5, and PBal_1 because they explain a cumulative variance of 84.90% (see Table 2). Note that we have written the principal balances in descending order by percentage of variance and we denote them as $\text{SPBal}_1, \text{SPBal}_2 \ldots \text{SPBal}_5$, respectively. For comparative purposes, we have estimated the proposed model using different selected principal balances to compare the results when extracting different percentages of variance of the data. In Table 2, we can observe that the formulation with $\text{SPBal}_1, \text{SPBal}_2 \ldots \text{SPBal}_5$ is the one with smallest value of 2002 DIC, 2004 DIC, and RMSD when fitting; consequently, this formulation is the most suitable option. In fact, it is underscored in Section 5.2 that SPBal_5 is particularly interesting for explaining the dynamics of the microbial community.

5.2. Estimated Parameters. Once we have selected the principal balances, we can estimate the proposed model. We are going to do so using Bayesian analysis because it allows us to simulate the posterior distribution of the model parameters using Markov Chain Monte Carlo (MCMC), as implemented in the free statistical software JAGS [65]. We fixed a burn-in period of 10000 iterations and an adaptation period of 1000 iterations. We kept one posterior sample in 100 iterations until a set of 20000 iterations was obtained, except for the first and second simulation datasets where just 2000 iterations were needed. We implemented this process using the *runjags* package in *R* [66]. Note that in Table S1 (see supplementary material), we can find the potential scale reduction factor defined by [67] and obtained using the *gelman.diag* function of the Coda package [68]. Table S1 shows that the potential scale reduction factor belongs to the interval $[1, 1.1]$ and is smaller than its upper 95% confidence limit.

In Table 3 we can find the estimated parameters, denoted in Section 3.2 as a_{ij}. Note that in each equation in the model (see equation (7)), we need to estimate six parameters, the intercept, and the five parameters associated with the selected principal balances. The estimated value of a_{i0} represents the abundance of the family i in the community; in fact, tiny intercepts are related to bacterial families of low abundance, whereas large intercepts indicate taxa of high abundance. In addition, the absolute value of the estimated a_{ij} with $j \neq 0$ represents the weight that the balance SPBal_j has in describing the relative abundance of the family i. As a result, we can say that the absolute value of a_{ij} with $j \neq 0$ represents the influence of the balance SPBal_j on the relative abundance of the family i.

Analysing the estimated parameters allows us to extract some information about the bacterial dynamics. Focusing on intercepts described in Table 3, we observe that the families *o_Clostridiales; f_*, *f_Desulfovibrionaceae*, and *f_Erysipelotrichaceae* have a low intercept, while the intercepts of *f_Bacteroidaceae* and *Other* are meaningful. A negative intercept indicates families of low abundance, whereas large positive intercepts indicate abundant ones. We define the percentage of sway that each family has on the bacterial community as $|a_{i0}| / \sum_{i=1}^{D} |a_{i0}|$. This percentage of sway provides information about the abundance of a family taking into account the other families; consequently, it enables us to compare the abundance of the different families. When the intercept is positive, a high percentage indicates high abundance and vice versa. However, when the intercept is negative, high percentage is related to low abundance and vice versa. In Table 3, the percentage of sway is written in bold, and we can observe that *f_Bacteroidaceae* and *Other*, which have positive intercepts and a percentage of sway of 33.91% and 17.50%, respectively, are the two most abundant families. In addition, *o_Clostridiales; f_* has a negative intercept and a percentage of 13.75%; it is therefore the least abundant family. The percentage of sway can help to gain a better understanding of the different dynamics of microbes because low percentage indicates small families, and high percentage is related to the extreme families: it indicates the most abundant if the intercept is positive or the least abundant if the percentage is negative.

As we have mentioned before, the absolute value of the estimated parameters a_{ij} with $j \neq 0$ represents the weight that the balance has in describing the relative abundance of the family, and they therefore provide information about the influence that each balance has on the family. Centering our attention on the absolute value of the estimated a_{ij}, we can observe that SPBal_5 is the one that most influences *f_Ruminococcaceae*, *f_Lachnospiraceae*, *f_Veillonellaceae*, and *f_Bacteroidaceae*. Consequently, the families that belong to the SPBal_5 are particularly interesting to explain the dynamics of these families. Note that $\text{SPBal}_{5,t-1} = \sqrt{1 \cdot 1/1 + 1} \cdot \ln(\mathbf{y}_{3,t-1}/\mathbf{y}_{6,t-1})$; therefore, this balance explains the relationship between the components y_3 and y_6 (which are *f_Erysipelotrichaceae* and *f_Ruminococcaceae* respectively) at time point $t-1$. A balance value close to zero can be interpreted as a high similarity between the numerator elements and the denominator

Table 1: The expression of the principal balances and their percentage of variance.

PBal	Expression	% var
$PBal_1$	$\sqrt{1 \cdot 1/1 + 1} \cdot \ln(\mathbf{x}_3/\mathbf{x}_6)$	9.00
$PBal_2$	$\sqrt{2 \cdot 1/2 + 1} \cdot \ln((\mathbf{x}_3 \cdot \mathbf{x}_6)^{1/2}/\mathbf{x}_7)$	15.38
$PBal_3$	$\sqrt{1 \cdot 1/1 + 1} \cdot \ln(\mathbf{x}_5/\mathbf{x}_9)$	2.15
$PBal_4$	$\sqrt{2 \cdot 1/2 + 1} \cdot \ln((\mathbf{x}_5 \cdot \mathbf{x}_9)^{1/2}/\mathbf{x}_4)$	7.62
$PBal_5$	$\sqrt{3 \cdot 1/3 + 1} \cdot \ln((\mathbf{x}_5 \cdot \mathbf{x}_9 \cdot \mathbf{x}_4)^{1/3}/\mathbf{x}_1)$	10.62
$PBal_6$	$\sqrt{1 \cdot 1/1 + 1} \cdot \ln(\mathbf{x}_2/\mathbf{x}_8)$	5.31
$PBal_7$	$\sqrt{4 \cdot 2/4 + 2} \cdot \ln((\mathbf{x}_5 \cdot \mathbf{x}_9 \cdot \mathbf{x}_4 \cdot \mathbf{x}_1)^{1/4}/(\mathbf{x}_2 \cdot \mathbf{x}_8)^{1/2})$	14.08
$PBal_8$	$\sqrt{6 \cdot 3/6 + 3} \cdot \ln((\mathbf{x}_5 \cdot \mathbf{x}_9 \cdot \mathbf{x}_4 \cdot \mathbf{x}_1 \cdot \mathbf{x}_2 \cdot \mathbf{x}_8)^{1/6}/(\mathbf{x}_7 \cdot \mathbf{x}_3 \cdot \mathbf{x}_6)^{1/3})$	35.81

Table 2: Female dataset at family taxonomic level. Results obtained when estimating the proposed model using different selected principal balances. Additionally, the table shows the deviance information criterion (DIC) defined in [63], the DIC defined in [64], the root mean square deviation (RMSD) of the results obtained fitting the model, and the RMSD of the results obtained using the model for prediction.

Proposed model different SPBal	% var	$\overline{D}$	pD (2002)	DIC (2002)	pD (2004)	DIC (2004)	RMSD (fitting)	RMSD (predicting)
$SPBal_1$	35.81	−1820.76	16.93	−1803.83	17.60	−1803.15	0.0299	0.0408
$SPBal_1$, $SPBal_2$	51.19	−1827.11	24.16	−1802.94	31.55	−1795.56	0.0292	0.0409
$SPBal_1$, $SPBal_2$, $SPBal_3$	65.27	−1826.61	31.08	−1795.52	40.64	−1785.96	0.0288	0.0442
$SPBal_1, SPBal_2, \ldots, SPBal_4$	75.89	−1832.65	37.68	−1794.96	45.90	−1786.74	0.0282	0.0452
$SPBal_1, SPBal_2, \ldots, SPBal_5$	84.90	−1872.23	45.48	−1826.75	64.49	−1807.74	0.0264	0.0498

Table 3: Estimated parameters (a_{ij}) and credible interval obtained by applying the proposed model to the female dataset at family taxonomic level. The table also contains the percentage of influence that each $SPBal_j$ has on each family and the percentage of sway that each family has on the bacterial community is represented in bold.

	Family taxonomic level					
a_{ij}	Intercept	$SPBal_1$	$SPBal_2$	$SPBal_3$	$SPBal_4$	$SPBal_5$
*o_bacteroidales; f_*	−0.0366 (−1.06, 0.96) **0.58%**	0.4255 (0.13, 0.69) 33.25%	0.1666 (−0.11, 0.46) 13.02%	0.0145 (−0.33,0.37) 1.13%	−0.3427 (−0.76, 0.04) 26.78%	−0.3303 (−0.69, 0.009) 25.81%
Other	1.1076 (0.25, 1.95) **17.50%**	0.7105 (0.44, 0.96) 57.00%	−0.0753 (−0.31, 0.13) 6.04%	0.0913 (−0.20, 0.39) 7.32%	0.2463 (−0.04, 0.53) 19.76%	−0.1231 (−0.47, 0.18) 9.88%
f_erysipelotrichaceae	−0.5757 (−2.05, 0.61) **9.10%**	−0.0497 (−0.40, 0.25) 26.03%	0.1025 (−0.32, 0.54) 53.69%	−0.0191 (−0.50, 0.47) 10.01%	0.0141 (−0.53, 0.56) 7.39%	0.0055 (−0.50, 0.46) 2.88%
*o_clostridiales; f_*	−0.8705 (−2.44, 0.38) **13.75%**	0.1802 (−0.15, 0.54) 67.62%	−0.0346 (−0.46, 0.35) 12.98%	0.0038 (−0.45, 0.48) 1.43%	−0.0281 (−0.56, 0.50) 10.54%	0.0198 (−0.50, 0.50) 7.43%
f_lachnospiraceae	0.4847 (−0.33, 1.47) **7,66%**	0.3356 (0.04, 0.61) 36,26%	−0.0199 (−0.27, 0.21) 2,15%	0.0123 (−0.29, 0.31) 1,33%	−0.0225 (−0.36, 0.30) 2,43%	−0.5353 (−0.88, −0.19) 57,83%
f_ruminococcaceae	0.2661 (−0.71, 1.44) **4.20%**	−0.2023 (−0.52, 0.06) 15.42%	0.1230 (−0.19, 0.44) 9.37%	−0.2109 (−0.56, 0.10) 16.07%	0.0594 (−0.30, 0.43) 4.53%	−0.7166 (−1.11, −0.30) 54.61%
f_veillonellaceae	−0.2057 (−1.26, 0.75) **3,25%**	0.2451 (−0.05, 0.54) 30,49%	−0.1049 (−0.45, 0.16) 13,05%	−0.0496 (−0.42, 0.30) 6,17%	−0.0238 (−0.41, 0.34) 2,96%	−0.3804 (−0.76, −0.02) 47,33%
f_desulfovibrionaceae	−0.6360 (−2.03, 0.45) **10,05%**	0.2993 (−0.03, 0.63) 51,10%	0.0363 (−0.34, 0.40) 6,20%	0.0866 (−0.32, 0.55) 14,79%	−0.1100 (−0.62, 0.36) 18,78%	−0.0535 (−0.52, 0.35) 9,13%
f_bacteroidaceae	2.1460 (1.30, 3.04) **33,91%**	0.2476 (−0.008, 0.49) 24,09	−0.0018 (−0.25, 0.22) 0,18	−0.1890 (−0.47, 0.08) 18,39	−0.0426 (−0.33, 0.25) 4,14	−0.5470 (−0.87, −0.24) 53,21

components; however, the larger the absolute value of the balance, the greater the dissimilarity between them. Furthermore, in Table 3, we can also observe that the coefficient of SPBal_1 has a significant absolute value in all families. As a result, SPBal_1, which is the selected principal balance with the highest percentage of variance, is also the one that has a significant influence on most of the bacterial families.

Families with low abundances will have low parameter values; as a result, in order to compare the influence of one SPBal_{j_0} in different families, it is interesting to study the percentage of influence $a_{ij}/\sum_{j=1}^{M} aij$. Observing Table 3, it seems that SPBal_1 does not have an special influence on *f_Erysipelotrichaceae* because $|a_{ij}| < 0.1$, while in the rest of the families $|a_{ij}| > 0.1$, however, if we observe the value of all $|a_{ij}|$ at *f_Erysipelotrichaceae*, we can observe that all the values are small and, in comparison, SPBal_1 has a significant influence. This is the reason why it is interesting to study the percentage of influence present in Table 3. In *f_Erysipelotrichaceae*, the SPBal_1 has a 26.03% percentage of influence although the selected principal balance that most affects this family is SPBal_2 with an influence of 53.69%. It would be a mistake to think that SPBal_1 has more influence on *f_Ruminococcaceae* than on *f_Erysipelotrichaceae* because the absolute value of the estimated parameter in *f_Ruminococcaceae* is greater than the value in *f_Erysipelotrichaceae*; in fact, the percentage of influence of SPBal_1 on *f_Ruminococcaceae* is 15.42%, which is smaller than the percentage of influence of the selected principal balance on *f_Erysipelotrichaceae*.

In Table S2 (see supplementary material), we present the estimated parameters obtained with the female dataset at genus level as it is the publicly available dataset with most parameters. Focusing on intercepts, we observe that *g_Bacteroides* and *Other* are the two genera that most influence the bacterial community, and *g_Dorea* is the genus with the smallest intercept. Studying the percentage of influence that each selected principal balance has on each genus, we can observe in Table S2 that SPBal_5 is the one that most influences eight genera, and SPBal_4 is the one that most influences six genera. This reflects the importance of SPBal_5 and SPBal_4, which are the balances with the least percentage of variance. Consequently, taking into account the principal balances that accumulate at least 80% of the variability is paramount in order to collect enough information. In addition, in Table S3, we can observe the estimated parameters obtained by analysing the male dataset, where both the selected principal balances with the lowest percentage of variance are also essential for describing the microbiome dynamics of the male subject.

5.3. Credible Intervals. The credible intervals present in Tables 3, S2, and S3 (see supplementary material) are noteworthy because most of them contain zero. In order to study this fact, we analyse the percentage of zeros included at the credible intervals and the number of parameters present in each equation of the model for all databases. We scrutinise these two aspects with the proposed model and the four approaches used for comparison; see details in Section 5.5.

Table 4 collects this analysis and points out two overall trends: (a) as the number of microbial entities present in the data grows, the number of parameters that contain zero in his credible interval rises; (b) the model with the fewest parameters in each equation of the model is the one with the fewest parameters containing the zero in its credible interval.

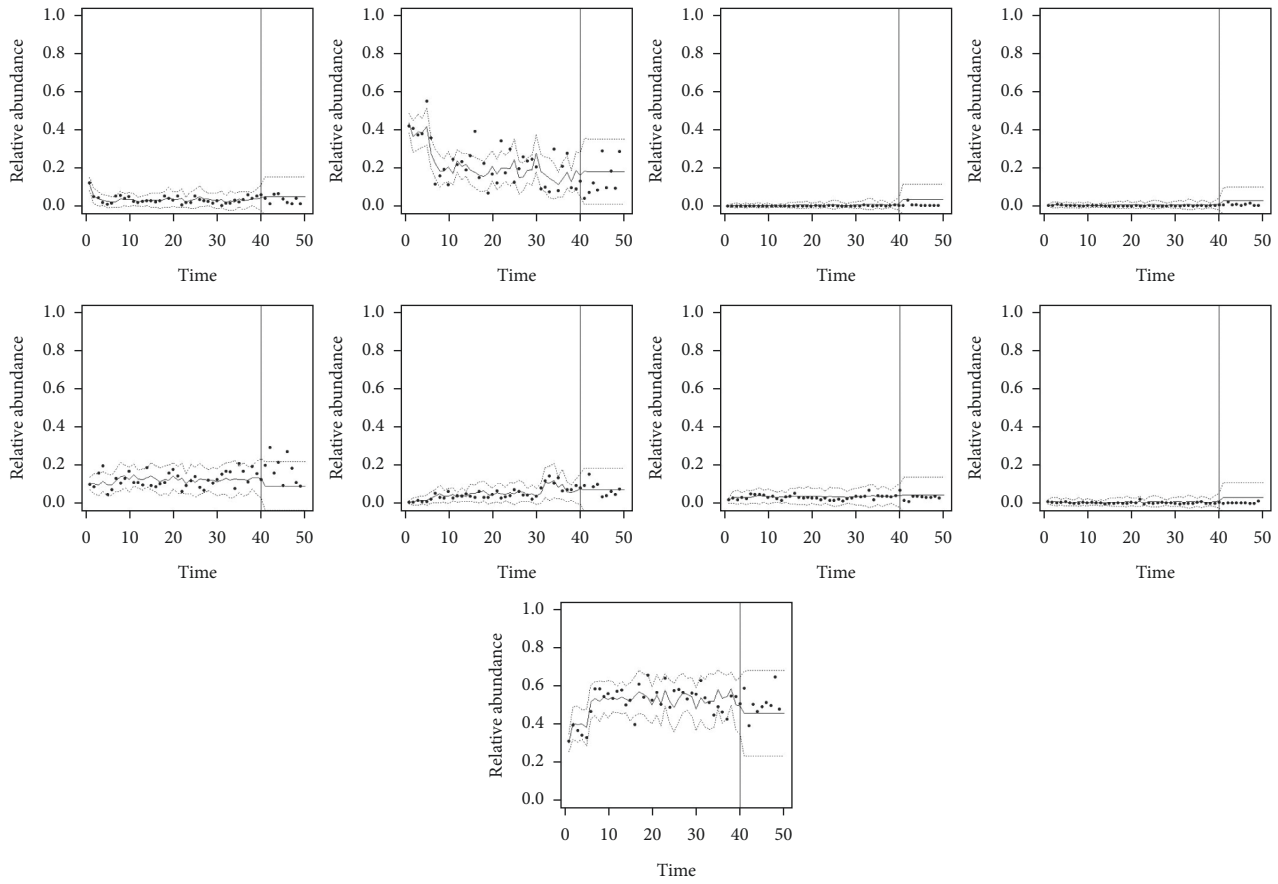
In all models, the expected value of each microbial entity depends on its corresponding autoregressive expression. In Table 4, we have denoted the number of parameters in each autoregressive expression as *Param/equation*. As the autoregressive expressions have the same number of parameters for all bacteria, the estimate parameters of the less abundant microbial entities must necessarily be small. In conclusion, small families will inevitably have small parameters, and as a result of working with small parameters, the credible intervals contain zero. As we are working with compositional data, the more microbial entities we have, the smaller these microbial entities are. As a result, the number of small parameters increases in datasets with a lot of microbial entities. To sum up, a zero present in a credible interval does not mean that the parameter is irrelevant; the zeros appear due to working with compositional data.

5.4. Graphic Results and Modeling Datasets with a Large Number of Microbial Entities. Graphic results show the ability of our model to both explain and predict the dynamics of microbiome communities. In order to predict, we estimate the model, and we denote as $a_{ij}^{(k)}$ the value of the parameter a_{ij} at the iteration k of the Markov Chain, with $k = 1, \ldots, J$. In order to obtain predictions using the model, we apply the following reasoning:

$$\widehat{y}_{it}^{(k)} = \frac{\widehat{\alpha_{it}}^{(k)}}{\sum_{i=1}^{D} \widehat{\alpha_{it}}^{(k)}} \text{ where } \widehat{\alpha_{it}}^{(k)} = \exp\left(a_{i0}^{(k)} + \sum_{j=0}^{M} a_{ij}^{(k)} \cdot \text{SPBal}_{jt-1}\right). \tag{8}$$

Finally, $\widehat{y}_{it} = \text{mean}(\widehat{y}_{it}^{(1)}, \ldots, \widehat{y}_{it}^{(K)})$ denotes the predictions obtained with the model. In Figures 2 and 3 we can observe how many time points we have used to estimate the model and to make predictions for each dataset.

First, we analyse the graphic results obtained with the female dataset at family taxonomic level because it is straightforwardly comparable with the male dataset and the genus study, as mentioned in Section 5.1. Figure 2(a) indicates that the two families with the highest intercept values



(a)

Figure 2: Continued.

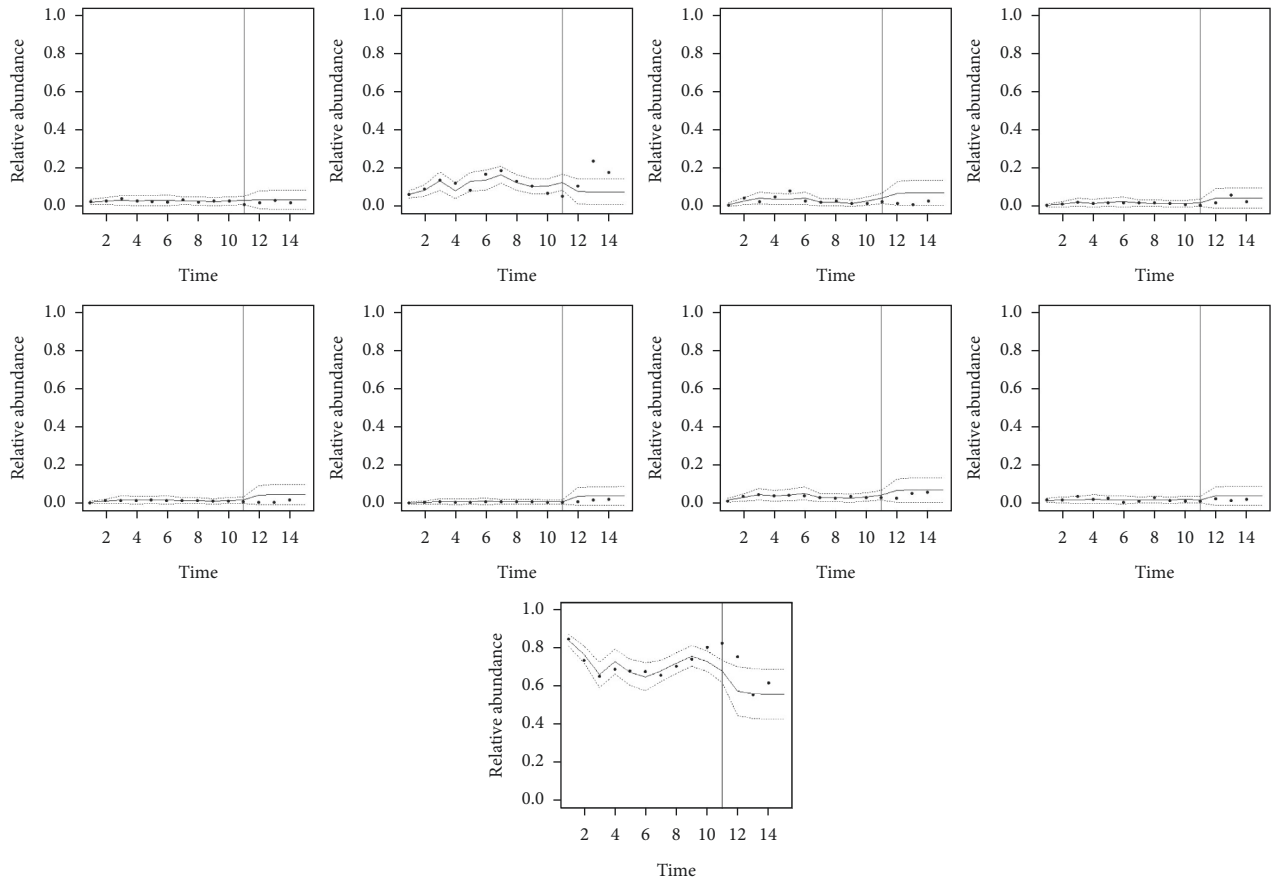


Figure 2: Model estimation and prediction at family taxonomic level. Black points represent the data, the black line represents the expected value obtained with the proposed model, and dashed lines represent the value of $E[y_{it}] \pm 2\sqrt{(\mathrm{Var}(y_{it}))}$. The vertical black line represents when the estimation of the model finishes and the prediction starts. (a) Female: the families of bacteria are, from left to right, *o_bacteroidales; f_*, *other*, *f_erysipelotrichaceae*, *o_clostridiales;f_*, *f_lachnospiraceae*, *f_ruminococcaceae*, *f_veillonellaceae*, *f_desulfovibrionaceae*, and *f_bacteroidaceae*. (b) Male: the families of bacteria are, from left to right, *f_alcaligenaceae*, *f_bacteroidaceae*, *f_lachnospiraceae*, *f_porphyromonadaceae*, *other*, *f_rikenellaceae*, *f_ruminococcaceae*, *f_veillonellaceae*, and *f_prevotellaceae*.

(*f_Bacteroidaceae* and *Other*, mentioned in Section 5.2) are also the families with the most relative abundance and variability. What is more, *o_Clostridiales*, *f_Desulfovibrionaceae*, and *f_Erysipelotrichaceae*, which are the families with the lowest intercepts, are the families with the lowest abundance. In order to measure the goodness of fit of the estimating and predicting part, we can observe the RMSD presented in Table 5.

In Figure 2 and Table 5, we can compare the results of the female and male datasets. Analysing Figure 2, we conclude that the families of the male have less variability than the families of the female. This leads to a decrease in the RMSD value obtained with the male dataset not only in the fitting part but also in the predicting part. Note that this drop is illustrated at Table 5.

Hitherto, we have analysed the datasets at family taxonomic level. In Figure 2, we can see that both female and male datasets have $D = 9$ families. When we increase the number D, we increase the number of equation (7) that our model has (see Section 3.2). In order to test our model with a higher value of D, we analyse the female dataset at genus taxonomic level. In this case, $D = 18$ and the dimension of the parameter estimation matrix a_{ij} is 18x6, as we can observe at Table S2. We must mention that the dimensions of a_{ij} with the female and male dataset at family taxonomic level are 9x6 and 9x5, respectively (see Tables 5 and S3).

In Figure 3 and Table 5, we can observe the results of the estimation and prediction when analysing the data at the genus taxonomic level. Figure 3 shows that *g_Bacteroides*, *g_Akkermansia*, and *Other* are the genera with the most variability. Indeed, we can observe in Table S2 that *g_Bacteroides* and *Other* are the two genera with the highest intercept values, and *g_Akkermansia* is especially influenced by SPBal_1 and SPBal_4.

Using simulated data, we have also managed to test our model when $D = 20$ and $D = 60$. In Table 4, we can see the results obtained, and we can conclude that the model achieves modeling and predicting in short-term microbiome dynamics with higher values of D. In fact, Table 5 shows that the RMSD of the fitting part of the data from the third simulation (with $D = 60$) is smaller than the RMSD of the data from the second simulation (with $D = 20$).

Table 4: Percentage of credible intervals that do not contain zero and the number of parameters in each equation of the model for all databases. *S*1, *S*2, and *S*3 denote the first, second, and third simulations. Note that the AR(1)-clr model does not achieve convergence in the F. genus dataset.

Model	Information	F. families $D = 9$	F. genus $D = 18$	Male $D = 9$	*S*1 $D = 15$	*S*2 $D = 20$	*S*3 $D = 60$
AR (1)-alr	% zeros	**12.5%**	**6.86%**	**0%**	**3.80%**	**1.57%**	**0%**
	Param/equation	9	18	9	15	20	60
alr-alr-Bal	% zeros	**41.66%**	**39.21%**	**4.16%**	**11.90%**	**12.28%**	**0%**
	Param/equation	3	3	3	3	3	3
PM	% zeros	**9.54%**	**8.33%**	**4.44%**	**7.69%**	**0.1%**	**0%**
	Param/equation	6	6	5	13	13	15
AR (1)-clr	% zeros	**2.22%**		**0%**	**4.16%**	**4.76%**	**0%**
	Param/equation	10	19	10	16	21	61
alr-Bal	% zeros	**43.75%**	**35.29%**	**31.25%**	**14.28%**	**15.78%**	**0%**
	Param/equation	2	2	2	2	2	2

Table 5: Deviance information criterion (DIC) defined in [63] and the DIC defined in [64]. The root mean square deviation (RMSD) of the results obtained fitting the model and the RMSD of the results obtained using the model for prediction.

Data	Model	$\overline{D}$	pD (2002)	DIC (2002)	pD (2004)	DIC (2004)	RMSD (fitting)	RMSD (predicting)
Female families ($D = 9$)	PM	$-1.87 \cdot 10^3$	$4.54 \cdot 10^1$	$-1.82 \cdot 10^3$	$6.44 \cdot 10^1$	$-1.80 \cdot 10^3$	0.0264	0.0498
	AR(1)-alr	$2.98 \cdot 10^4$	$2.09 \cdot 10^4$	$5.07 \cdot 10^4$	$9.36 \cdot 10^8$	$9.37 \cdot 10^8$	0.0245	0.0398
	AR(1)-clr	$3.63 \cdot 10^4$	$2.34 \cdot 10^4$	$5.98 \cdot 10^4$	$1.03 \cdot 10^9$	$1.03 \cdot 10^9$	0.0245	0.0395
	alr-Bal	$5.81 \cdot 10^4$	$4.48 \cdot 10^4$	$1.02 \cdot 10^5$	$3.55 \cdot 10^9$	$3.55 \cdot 10^9$	0.0346	0.0414
	alr-alr-Bal	$4.95 \cdot 10^4$	$3.25 \cdot 10^4$	$8.20 \cdot 10^4$	$2.15 \cdot 10^9$	$2.15 \cdot 10^9$	0.0315	0.0399
Female genera ($D = 18$)	PM	$-5.25 \cdot 10^3$	$8.96 \cdot 10^1$	$-5.16 \cdot 10^3$	$9.57 \cdot 10^1$	$-5.15 \cdot 10^3$	0.0174	0.0280
	AR(1)-alr	$4.39 \cdot 10^5$	$2.80 \cdot 10^5$	$7.19 \cdot 10^5$	$3.37 \cdot 10^{10}$	$3.37 \cdot 10^{10}$	0.0172	0.0368
	alr-Bal	$8.31 \cdot 10^5$	$2.20 \cdot 10^5$	$1.05 \cdot 10^6$	$7.49 \cdot 10^{10}$	$7.49 \cdot 10^{10}$	0.0271	0.0216
	alr-alr-Bal	$4.38 \cdot 10^5$	$1.41 \cdot 10^5$	$5.79 \cdot 10^5$	$2.49 \cdot 10^{10}$	$2.49 \cdot 10^{10}$	0.0208	0.0207
Male ($D = 9$)	PM	$-5.19 \cdot 10^2$	$3.34 \cdot 10^1$	$-4.86 \cdot 10^2$	$5.77 \cdot 10^1$	$-4.61 \cdot 10^2$	0.0111	0.0458
	AR(1)-alr	$2.96 \cdot 10^3$	$3.42 \cdot 10^3$	$6.38 \cdot 10^3$	$2.72 \cdot 10^7$	$2.72 \cdot 10^7$	0.0098	0.0278
	AR(1)-clr	$6.09 \cdot 10^3$	$6.49 \cdot 10^3$	$1.25 \cdot 10^4$	$6.67 \cdot 10^7$	$6.67 \cdot 10^7$	0.0089	0.0289
	alr-Bal	$1.22 \cdot 10^3$	$1.37 \cdot 10^3$	$2.60 \cdot 10^3$	$6.60 \cdot 10^6$	$6.60 \cdot 10^6$	0.0144	0.0309
	alr-alr-Bal	$1.56 \cdot 10^3$	$1.70 \cdot 10^3$	$3.26 \cdot 10^3$	$8.43 \cdot 10^6$	$8.43 \cdot 10^6$	0.0142	0.0306
Simulated data ($D = 15$)	PM	$-1.40 \cdot 10^3$	$1.23 \cdot 10^2$	$-1.28 \cdot 10^3$	$1.93 \cdot 10^2$	$-1.21 \cdot 10^3$	0.0283	0.0401
	AR(1)-alr	$-8.59 \cdot 10^2$	$7.81 \cdot 10^2$	$-7.74 \cdot 10^1$	$5.40 \cdot 10^4$	$5.32 \cdot 10^4$	0.0236	0.0398
	AR(1)-clr	$-3.67 \cdot 10^0$	$1.24 \cdot 10^3$	$8.77 \cdot 10^0$	$1.02 \cdot 10^6$	$1.02 \cdot 10^6$	0.0239	0.0413
	alr-Bal	$-1.22 \cdot 10^3$	$6.94 \cdot 10^1$	$-1.15 \cdot 10^3$	$2.19 \cdot 10^3$	$9.72 \cdot 10^2$	0.0339	0.0382
	alr-alr-Bal	$-1.20 \cdot 10^3$	$1.07 \cdot 10^2$	$-1.10 \cdot 10^3$	$3.13 \cdot 10^3$	$1.92 \cdot 10^3$	0.0329	0.0385
Simulated data ($D = 20$)	PM	$-2.20 \cdot 10^3$	$2.16 \cdot 10^2$	$-1.98 \cdot 10^3$	$3.74 \cdot 10^2$	$-1.83 \cdot 10^3$	0.0190	0.0310
	AR(1)-alr	$-1.11 \cdot 10^3$	$1.63 \cdot 10^3$	$5.13 \cdot 10^2$	$2.02 \cdot 10^5$	$2.01 \cdot 10^5$	0.0144	0.0327
	AR(1)-clr	$6.63 \cdot 10^0$	$3.38 \cdot 10^3$	$4.05 \cdot 10^3$	$1.55 \cdot 10^7$	$1.55 \cdot 10^7$	0.0147	0.0331
	alr-Bal	$-2.02 \cdot 10^3$	$7.67 \cdot 10^1$	$-1.94 \cdot 10^3$	$1.22 \cdot 10^3$	$-7.96 \cdot 10^2$	0.0257	0.0308
	alr-alr-Bal	$-2.00 \cdot 10^3$	$1.10 \cdot 10^2$	$-1.89 \cdot 10^3$	$1.56 \cdot 10^3$	$-4.43 \cdot 10^2$	0.0252	0.0309
Simulated data ($D = 60$)	PM	$-3.50 \cdot 10^3$	$3.51 \cdot 10^2$	$-3.15 \cdot 10^3$	$1.60 \cdot 10^4$	$1.25 \cdot 10^4$	0.0001	0.0143
	AR(1)-alr	$5.12 \cdot 10^5$	$5.16 \cdot 10^5$	$1.02 \cdot 10^6$	$1.92 \cdot 10^{11}$	$1.92 \cdot 10^{11}$	0.0009	0.0142
	AR(1)-clr	$4.91 \cdot 10^5$	$4.95 \cdot 10^5$	$9.86 \cdot 10^5$	$1.89 \cdot 10^{11}$	$1.89 \cdot 10^{11}$	0.0009	0.0124
	alr-Bal	$-2.24 \cdot 10^3$	$4.86 \cdot 10^2$	$-1.75 \cdot 10^3$	$1.49 \cdot 10^5$	$1.47 \cdot 10^5$	0.0052	0.0060
	alr-alr-Bal	$-1.89 \cdot 10^3$	$8.64 \cdot 10^2$	$-1.03 \cdot 10^3$	$7.79 \cdot 10^5$	$7.77 \cdot 10^5$	0.0050	0.0061

5.5. Comparison. For comparative purposes, we also show the results obtained with alternative models that differ from the proposed model due to the autoregressive component and the reparameterization of the Dirichlet parameters.

Reference [46] highlights the ability of the Dirichlet distribution with an autoregressive structure that uses the alr or the clr transformation as a link function and imposes a parameter that is equal to the sum of the Dirichlet parameters. On the one hand, we proposed the AR(1)-alr and AR(1)-clr models that take this construction and use it in a Bayesian framework while writing in the autoregressive structured an AR expression with the alr or clr

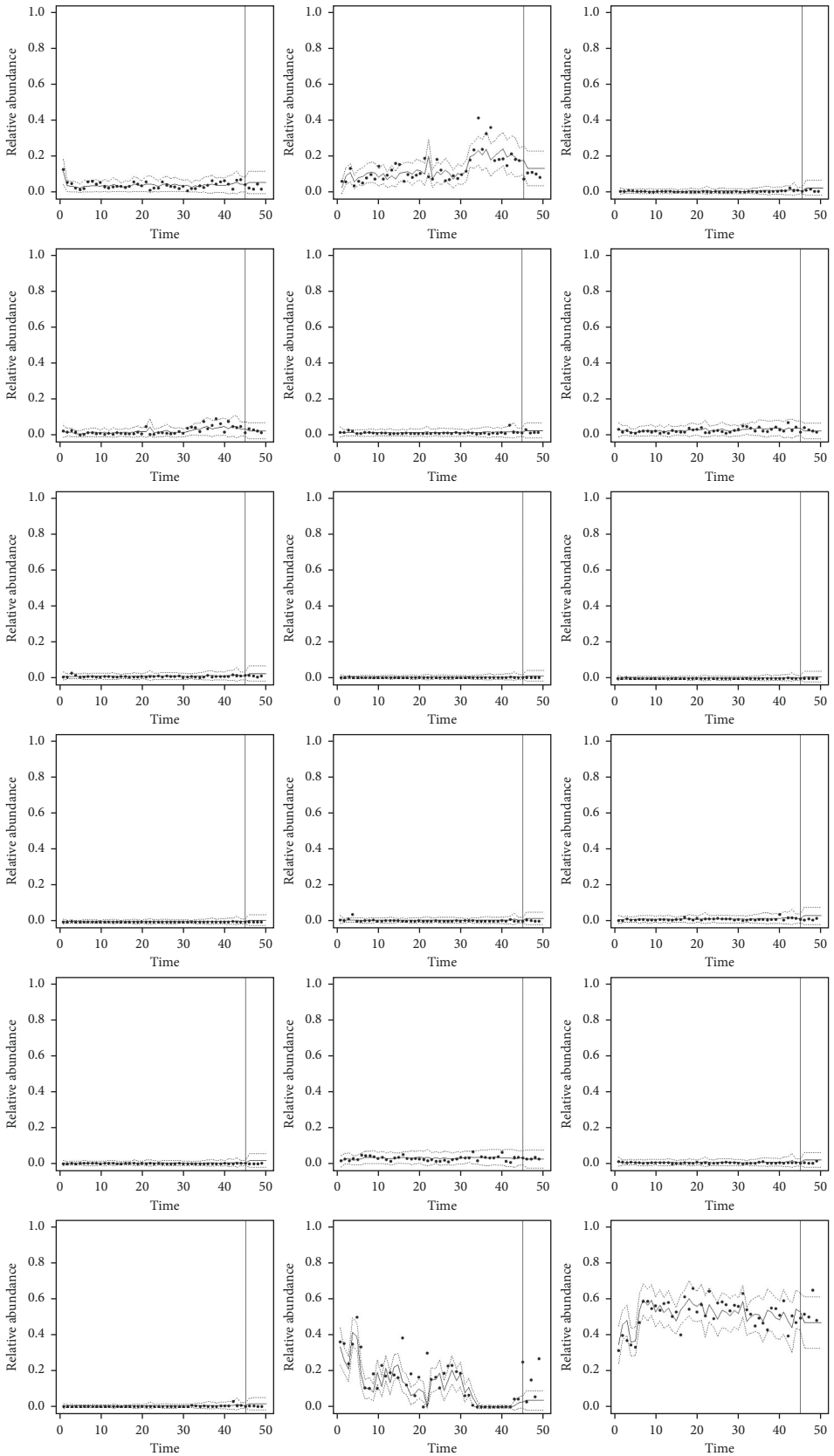


Figure 3: Model estimation and prediction of the female dataset at genus taxonomic level. Black points represent the data, the black line represents the expected value obtained with the proposed model, and dashed lines represent the value of $E[y_{it}] \pm 2\sqrt{(\mathrm{Var}(y_{it}))}$. The vertical black line represents when the estimation of the model finishes and the prediction starts. The genera are, from left to right, *o_bacteroidales;f_; g_*, *other, o_clostridiales;f_;g_*, *f_lachnospiraceae;g_*, *g_blautia*, *f_lachnospiraceae;g_clostridium*, *g_coprococcus*, *g_dorea, g_eubacterium*, *g_roseburia, g_ruminococcus*, *f_ruminococcaceae;g_*, *f_ruminococcaceae;g_clostridium*, *g_phascolarctobacterium*, *g_desulfovibrio*, *f_erysipelotrichaceae;g_clostridium*, *g_akkermansia,* and *g_bacteroides.*

transformation, respectively. On the other hand, [41] emphasizes the importance of balances to identify microbial signatures; for this reason, we present alr-Bal and alr-alr-Bal models, which are two alternative that maintain the general construction of [46] but introduce balances in the autoregressive structure and use it in a Bayesian framework. The following alternatives have been implemented: a modification of the alr-alr-Bal model using a balance whose numerator does not include the microbial entity i, the alr-alr-Bal model but using the clr transformation to accompany the a_{i1} parameter and a balance that compares all microbial entities except i with all the microbial entities, and an alteration of this last model applying the clr to transform the Dirichlet parameters. However, these alternatives did not lead us to obtain relevant fitting results, and consequently, they are not included in this article.

5.5.1. Comparison with AR(1)-alr Model. In this case, in order to link $\alpha_t = (\alpha_{1t}, \alpha_{2t} \ldots, \alpha_{Dt})$ with $\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \mathbf{y}_{t-3}, \ldots, \mathbf{y}_1$, we propose:

$$\log\left(\frac{\alpha_{it}}{\alpha_{Dt}}\right) := \mu_{it}, \tag{9}$$

with

$$\mu_{it} = a_{i0} + \sum_{j=1}^{D-1} a_{ij} \cdot \log\left(\frac{y_{jt-1}}{y_{Dt-1}}\right) \text{where } i = 1, \ldots, D-1. \tag{10}$$

Note that $\tau_t = \alpha_{1t} + \alpha_{2t} + \ldots + \alpha_{Dt}$ is the concentration parameter of the Dirichlet distribution and must also be estimated.

5.5.2. Comparison with AR(1)-clr Model. This consideration is similar to the AR(1)-alr model. In this case, the centered log-ratio (clr) transformation is used. Note that this transformation scales each component vector by the geometric mean, $g(\cdot)$.

$$\log\left(\frac{\alpha_{it}}{g(\boldsymbol{\alpha}_t)}\right) := \mu_{it},$$

$$\mu_{it} = a_{i0} + \sum_{j=1}^{D-1} a_{ij} \cdot \log\left(\frac{y_{jt-1}}{g(\mathbf{y}_{t-1})}\right) \text{where } i = 1, \ldots, D. \tag{11}$$

In this alternative formulation, τ_t must also be estimated.

5.5.3. Comparison with alr-Bal Model. This model has the same structure as the AR(1)-alr model, but in this case:

$$\mu_{it} = a_{i0} + a_{i1} \cdot \log\left(\frac{g\left(y_{1,t-1}, y_{2,t-1}, \ldots, y_{D-1,t-1}\right)}{g\left(y_{D,t-1}\right)}\right). \tag{12}$$

Note that $\log(g(\cdot)/g(\cdot))$ is a balance expression and τ_t must also be estimated.

5.5.4. Comparison with alr-alr-Bal Model. Following the structure of AR(1)-alr, this model presents an alternative formulation of the autoregressive component, combining the alr transformation and the balance used in the alr-Bal model.

$$\mu_{it} = a_{i0} + a_{i1} \cdot \log\left(\frac{y_{i,t-1}}{y_{D,t-1}}\right) + a_{i2} \cdot \log\left(\frac{g\left(y_{1,t-1}, y_{2,t-1}, \ldots, y_{D-1,t-1}\right)}{g\left(y_{D,t-1}\right)}\right). \tag{13}$$

Note that $\log(g(\cdot)/g(\cdot))$ is a balance expression and τ_t must also be estimated.

We compare the proposed model with these four alternatives to assess the ability of principal balances compared to AR models or other balance options. In fact, analysing Table 5, we can thoroughly compare the five methods. The proposed model is the one that achieves the lowest DIC values not only at the 2004 DIC but also at the 2002 DIC. If we focus on the 2004 DIC, we observe that the proposed model has by far the smallest DIC value; in addition, alr-Bal and alr-alr-Bal models obtain similar results, but they enhance the values obtained with the AR methods (except in female datasets). In spite of the fact that AR methods describe similar behaviours, in general, we can observe an improvement in the AR(1)-alr model. This dynamic is also matched at the 2002 DIC values; however, the values of the 2002 DIC are smaller than the values of the 2004 DIC in all models.

We can find the smallest 2002 DIC and 2004 DIC in the female dataset analysed at genus taxonomic level. The proposed model minimizes the 2002 and 2004 DICs with the values of $-5.16 \cdot 10^3$ and $-5.15 \cdot 10^3$, respectively. On the other hand, we observe the highest values of both DICs in the simulation with $D = 60$. The AR(1)-clr model maximizes the 2002 DIC with a value of $9.86 \cdot 10^5$, and AR(1)-alr maximizes the 2004 DIC with the value of $1.92 \cdot 10^{11}$. Note that in the female dataset analysed at genus taxonomic level, we can find the biggest difference between the DIC values obtained with the proposed model and the DICs achieved with other datasets. We must mention that the AR(1)-alr model did not achieve convergence with the dataset at genus taxonomic level. In conclusion, the difference between the proposed model and the other models is reflected far more clearly in this dataset.

In Table 5, we also can study the RMSD values of the fitting and the predicting parts. Notice that the values of the

fitting part are smaller than the values of the predicting part in all models. Focusing on the fitting part, we can observe that alr-Bal and alr-alr-Bal have larger RMSE values than the proposed model in all datasets; however, AR(1)-alr and AR(1)-clr have in general smaller values than the proposed model. Despite this, the proposed model has advantages over these models: it reduces the number of parameters and this enhances the estimation of large datasets. This is reflected in the simulated dataset with $D = 60$ where the proposed model has the smallest RMSE value, improving on the AR(1)-alr and AR(1)-clr RMSE values. Shifting the focus from fitting to predicting, we can observe that the AR(1)-alr or the AR(1)-clr has a larger RMSE value in all the datasets with $D > 9$ except for the simulated dataset with $D = 60$, where the difference between the RMSE values of the proposed model and the AR(1)-alr or AR(1)-clr is small. As a result, in this dataset, the models with AR structure and the proposed model have a similar prediction capacity; however, the proposed model describes the bacterial dynamics better, as we have seen before with the RMSE values of the fitting part. The same is true of the alr-Bal and alr-alr-Bal models; they have a smaller RMSE value in the predicting part than the proposed model, but they have the worst RSME value in the fitting part, so these models are the ones that worst describe the bacterial dynamics.

In conclusion, the proposed model is the one with the smallest DIC values, and this leads to us conclude that the proposed model (PM) is the most suitable approach to model microbiome dynamics. We must take into account that the proposed model does not have the smallest RMSE value in general; in these cases, however, the proposed model still has an advantage over the other models: it better describes the bacterial dynamics of large datasets.

6. Conclusion

In this paper, we develop a compositional autoregressive model in Bayesian framework that is able to describe microbiome time series dynamics. In order to model the relative abundance of microbiome longitudinal data, we use the Dirichlet distribution with time-varying parameters. We propose that the logarithm of the Dirichlet parameters can be explained by an autoregressive structure, which contains the principal balances that account for at least 80% of the variability. A comparison study selecting principal balances that account for a lower percentage of variance is provided in this article. Even though we have developed the model for the analysis of microbiome longitudinal data, it can be adapted for other scenarios defined by longitudinal data with compositional nature.

Principal balances are a compositional tool that enables us to extract the balances that maximize the variance. As balances with more variability will better explain the variability of the data, this formulation allows us to reduce the number of parameters to be estimated without removing or grouping microbial taxa. A direct consequence of reducing the number of parameters is being able to model datasets with a higher number of microbial taxa. In fact, we demonstrated the ability of our model with six datasets, three of which are publicly available and present the data of a healthy female and a healthy male, while the rest are simulated. Two of the publicly available datasets have 9 families to model and the other has 18 genera. The simulated datasets have 15, 20, and 60 microbial taxa to model, respectively.

In conclusion, we have tested the ability of our model with datasets that have different numbers of microbial taxa, different taxonomic levels, healthy people of different sexes, and datasets that are publicly available and simulated. This paper presents the results obtained by analysing these datasets using the proposed model and four alternative models with which we compare the proposed model. These alternative models differ from the proposed model because they change the reparameterization of the Dirichlet parameters using the alr and the clr transformations, and they modify the autoregressive component using an AR model structure or other balance options. After comparing them, we conclude that models with balances are the ones that worst describe the bacterial dynamics but obtain the best prediction results. Models with AR structure describe the data better than the proposed model, except for big datasets, where they also present the worst prediction results. Finally, the proposed model better describes the bacterial dynamics of big datasets due to reducing the number of parameters and obtains the lowest DIC values. In addition, it can be effective for short-term prediction of the future dynamics of a microbial community. This approach can be an interesting alternative to predict the future dynamics of a microbial community.

Additionally, the proposed model allows us to analyse the microbial interactions using the value of the estimated parameters. Indeed, the estimated value of the intercept represents the abundance of the microbial taxa in the community; as a matter of fact, small intercepts are related to bacterial families of low abundance, whereas large intercepts indicate taxa of high abundance. In addition, the absolute value of the rest of the estimated parameters represents the weight that the selected principal balance has in describing the relative abundance of the family; as a result, we can say that the absolute value of a_{ij} with $j \neq 0$ represents the influence of the balance SPBal_j at the relative abundance of the family i. We must mention that balances provide information about the similarity between their numerator and denominator components; in fact, balances that are near zero indicate that the two groups are similar, and the higher the absolute value of the balance, the larger is the dissimilarity between them.

Nevertheless, our proposal also has some limitations. Even if we choose the method for obtaining the principal balances that requires least computation time, estimating the models in big datasets still requires a significant amount of time. As future development, in order to improve the usefulness of the model, we propose to extend it to consider external covariates. The incorporation of covariates affecting microbiota dynamics, such as antibiotics or diet variations, may enable us to study the effectiveness of some medical treatments. Moreover, including random effects will allow us to study different subjects at the same time, such as healthy and unhealthy individuals, for example.

Data Availability

The data can be found at https://drive.google.com/drive/folders/1xIbHcjT0P-TJ6MQ5P91FwoEbu3gRkM2n?usp=sharing.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by grants from the Spanish Ministry of Economy and Competitiveness (projects MTM2017-83850-P, SAF2012-31187, SAF2013-49788-EXP, and SAF2015-65878-R), Carlos III Institute of Health (projects PIE14/00045 and AC15/00022), Generalitat Valenciana (projects PrometeoII/2014/065 and Prometeo/2018/A/133), and Asociación Española Contra el Cancer (project AECC 2017-1485) and was co-financed by the European Regional Development Fund (ERDF).

Supplementary Materials

We present the Tables S1–S3 as supplementary material. Table S1: potential scale reduction factor (labelled psrf) and their upper 95% confidence limits (labelled upper C.I.) of all the estimated parameter a_{ij}. Table S2: estimated parameters (a_{ij}) and credible interval obtained applying the Proposed Model to the female dataset at genus taxonomic level. The table also contains the percentage of influence that each SPBal_j has on each genus and, in bold, it is represented the percentage of sway that each genus has on the bacterial community. Table S3: estimated parameters (a_{ij}) and credible interval obtained applying the Proposed Model to the Male dataset. The table also contains the percentage of influence that each SPBal_j has on each family and, in bold, it is represented the percentage of influence that each family has on the bacterial community. (*Supplementary Materials*)

References

[1] J. Lederberg and A. T. McCray, "Ome Sweet" Omics—a genealogical treasury of words," *The Scientist*, vol. 15, no. 8, 2001.

[2] J. F. Cryan and T. G. Dinan, "Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour," *Nature Reviews Neuroscience*, vol. 13, no. 10, pp. 701–712, 2012.

[3] J. Ahn, R. Sinha, Z. Pei et al., "Human gut microbiome and risk for colorectal cancer," *Journal of the National Cancer Institute: Journal of the National Cancer Institute*, vol. 105, no. 24, pp. 1907–1911, 2013.

[4] J. F. Vázquez-Castellanos, S. Serrano-Villar, N. Jiménez-Hernández et al., "Interplay between gut microbiota metabolism and inflammation in HIV infection," *The ISME Journal*, vol. 12, no. 8, pp. 1964–1976, 2018.

[5] R. Xu and Q. Wang, "Towards understanding brain-gut-microbiome connections in Alzheimer's disease," *BMC Systems Biology*, vol. 10, no. S3, pp. 63–712, 2016.

[6] N. Larsen, F. K Vogensen, F. W. J. van den Berg et al., "Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults," *PLoS One*, vol. 5, no. 2, Article ID e9085, 2010.

[7] J. M. Brown and S. L. Hazen, "The gut microbial endocrine organ: bacterially derived signals driving cardiometabolic diseases," *Annual Review of Medicine*, vol. 66, no. 1, pp. 343–359, 2015.

[8] A. Durbán, J. J. Abellán, N. Jiménez-Hernández et al., "Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome," *FEMS Microbiology Ecology*, vol. 86, no. 3, pp. 581–589, 2013.

[9] D. Gevers, S. Kugathasan, L. A. Denson et al., "The treatment-naive microbiome in new-onset Crohn's disease," *Cell Host & Microbe*, vol. 15, no. 3, pp. 382–392, 2014.

[10] L. Giloteaux, J. K. Goodrich, W. A. Walters, S. M. Levine, R. E. Ley, and M. R. Hanson, "Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome," *Microbiome*, vol. 4, no. 1, pp. 30–681, 2016.

[11] H. J. Flint, "Obesity and the gut microbiota," *Journal of Clinical Gastroenterology*, vol. 45, pp. 128–132, 2011.

[12] S. Subramanian, S. Huq, T. Yatsunenko et al., "Persistent gut microbiota immaturity in malnourished Bangladeshi children," *Nature*, vol. 510, no. 7505, pp. 417–421, 2014.

[13] P. Kundu, E. Blacher, E. Elinav, and S. Pettersson, "Our gut microbiome: the evolving inner self," *Cell*, vol. 171, no. 7, pp. 1481–1493, 2017.

[14] J. Peterson, S. Garges, M. Giovanni et al., "The NIH human microbiome project," *Genome Research*, vol. 19, no. 12, pp. 2317–2323, 2009.

[15] J. M. Martí, D. Martínez-Martínez, T. Rubio et al., "Health and disease imprinted in the time variability of the human," *MicrobiomeAmerican Society for Microbiology Journals*, vol. 2, no. 2, 2018.

[16] E. C. Dinleyici, D. Martínez-Martínez, A. Kara et al., "Time series analysis of the microbiota of children suffering from acute infectious diarrhea and their recovery after treatment," *Frontiers in Microbiology*, vol. 9, p. 1230, 2018.

[17] J. A. Gilbert and S. V. Lynch, "Community ecology as a framework for human microbiome research," *Nature Medicine*, vol. 25, no. 6, pp. 884–889, 2019.

[18] L. Proctor, "What's next for the human microbiome?" *Nature*, 2019.

[19] P. G. Larsson, E. Brandsborg, U. Forsum et al., "Extended antimicrobial treatment of bacterial vaginosis combined with human lactobacilli to find the best treatment and minimize the risk of relapses," *BMC Infectious Diseases*, vol. 11, no. 1, p. 223, 2011.

[20] Z. Kassam, C. H. Lee, Y. Yuan, and R. H. Hunt, "Fecal microbiota transplantation for *Clostridium difficile* infection: systematic review and meta-analysis," *American Journal of Gastroenterology*, vol. 108, no. 4, pp. 500–508, 2013.

[21] E. van Nood, M. G. Dijkgraaf, and J. J. Keller, "Duodenal infusion of feces for recurrent *Clostridium difficile* N," *Engl. J. Med.*vol. 368, pp. 401–415, 2013.

[22] S. Serrano-Villar, A. Talavera-Rodríguez, M. J. Gosalbes et al., "Fecal microbiota transplantation in HIV: a pilot placebo-controlled study," *Nature Communications*, vol. 12, no. 1, p. 1139, 2021.

[23] S. Marino, N. T. Baxter, G. B. Huffnagle, J. F. Petrosino, and P. D. Schloss, "Mathematical modeling of primary succession of murine intestinal microbiota," *Proceedings of the National Academy of Sciences*, vol. 111, no. 1, pp. 439–444, 2014.

[24] R. R. Stein, V. Bucci, N. C. Toussaint et al., "Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota," *PLoS Computational Biology*, vol. 9, no. 12, Article ID e1003388, 2013.

[25] B. K. Kuntal, C. Gadgil, and S. S. Mande, "Web-gLV: a web based platform for lotka-volterra based modeling and simulation of microbial populations," *Frontiers in Microbiology*, vol. 10, p. 288, 2019.

[26] V. Bucci and J. B. Xavier, "Towards predictive models of the human gut microbiome," *Journal of Molecular Biology*, vol. 426, no. 23, pp. 3907–3916, 2014.

[27] C. K. Fisher and P. Mehta, "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression," *PLoS One*, vol. 9, no. 7, Article ID e102451, 2014.

[28] D. Gonze, K. Z. Coyte, L. Lahti, and K. Faust, "Microbial communities as dynamical systems," *Current Opinion in Microbiology*, vol. 44, pp. 41–49, 2018.

[29] P. Trosvik, K. Rudi, T. Naes et al., "Characterizing mixed microbial population dynamics using time-series analysis," *The ISME Journal*, vol. 2, no. 7, pp. 707–715, 2008.

[30] P. Trosvik, N. C. Stenseth, and K. Rudi, "Convergent temporal dynamics of the human infant gut microbiota," *The ISME Journal*, vol. 4, no. 2, pp. 151–158, 2010.

[31] I. Chen, Y. D. Kelkar, Y. Gu, J. Zhou, X. Qiu, and H. Wu, "High-dimensional linear state space models for dynamic microbial interaction networks," *PLoS One*, vol. 12, no. 11, Article ID e0187822, 2017.

[32] L. Shenhav, O. Furman, L. Briscoe et al., "Modeling the temporal dynamics of the gut microbial community in adults and infants," *PLoS Computational Biology*, vol. 15, no. 6, Article ID e1006960, 2019.

[33] A. Bodein, O. Chapleur, A. Droit, and K. A. Lê Cao, "A generic multivariate framework for the Integration of microbiome longitudinal studies with other data types," *Frontiers in Genetics*, vol. 10, no. 963, 2019.

[34] Y. Xia, J. Sun, and D. G. Chen, *Statistical Analysis of Microbiome Data with R*, Springer, Berlin, Germany, 2018.

[35] J. Aitchison, *The Statistical Analysis of Compositional Data. Monographs an Statistics and Applied Probability*Chapman Hall Ltd London, London, UK, 1986.

[36] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: and this is not optional," *Frontiers in Microbiology*, vol. 8, p. 2224, 2017.

[37] G. K. Pearson, "Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurements of organs," *Proceedings of the Royal Society*, vol. 60, 1897.

[38] J. Aitchison, *A Concise Guide to Compositional Data Analysis*, CDA Workshop, Girona Spain, 2003.

[39] P. Kynčlová, P. Filzmoser, and K. Hron, "Modeling compositional time series with vector autoregressive models," *Journal of Forecasting*, vol. 34, pp. 303–314, 2015.

[40] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, "A phylogenetic transform enhances analysis of compositional microbiota data," *Elife*, vol. 6, Article ID e21887, 2017.

[41] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle, "Balances: a new perspective for microbiome analysis," *mSystems*, vol. 3, no. 4, Article ID e00053-18, 2018.

[42] T. A. Joseph, L. Shenhav, J. B. Xavier, E. Halperin, and I. Pe'er, "Compositional Lotka-Volterra describes microbial dynamics in the simplex," *PLoS Computational Biology*, vol. 16, no. 5, Article ID e1007917, 2020.

[43] J. D. Silverman, H. K. Durand, R. J. Bloom, S. Mukherjee, and L. A. David, "Dynamic linear models guide design and analysis of microbiota studies within artificial human guts," *Microbiome*, vol. 6, no. 1, 2018.

[44] T. Äijö, C. L. Müller, and R. Bonneau, "Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing," *Bioinformatics*, vol. 34, no. 3, pp. 372–380, 2018.

[45] V. Pawlowsky-Glahn and J. J. Egozcue, "Principal balances," in *Proceedings of the 4th International Workshop on Compositional Data Analysis*, Feliu de Guíxols, Spain, June 2011.

[46] T. Zheng and R. Chen, "Dirichlet ARMA models for compositional time series," *Journal of Multivariate Analysis*, Elsevier, vol. 158, , pp. 31–46, 2017.

[47] J. Brehm, S. Gates, and B. Gomez, "A Comparison of Methods for Compositional Data Analysis," in *Proceedings of the 1998 Political Methodology Society annual meetings*, San Diego, CA, USA, April 1998.

[48] R. H. Hijazi and R. W. Jernigan, "Modeling compositional data using dirichlet regression models," *Journal of Applied Probability and Statistics*, vol. 4, pp. 77–91, 2009.

[49] V. Pawlowsky-Glahn and J. J. Egozcue, "BLU estimators and compositional data," *Mathematical Geology*, vol. 34, no. 3, pp. 259–274, 2002.

[50] J. J. Egozcue and V. Pawlowsky-Glahn, "Groups of parts and their balances in compositional data analysis," *Mathematical Geology*, vol. 37, no. 7, pp. 795–828, 2005.

[51] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, Wiley, Hoboken, NJ, USA, 2011.

[52] J. A. Martín-Fernández, V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosona-Delgado, "Advances in principal balances for compositional data," *Mathematical Geosciences*, vol. 50, no. 3, pp. 273–298, 2018.

[53] D. Müllner, "Fastcluster: fast hierarchical, agglomerative clustering routines for R and Python," *Journal of Statistical Software*, vol. 53, no. 9, pp. 1–18, 2013.

[54] F. J. Santonja, J. A. Martín-Fernández, and A. Moya, "Compositional PB-lagged regression models for microbiota longitudinal data," 2020.

[55] A. Gelman, "Prior distributions for variance parameters in hierarchical models," *Bayesian Analysis*, vol. 1, pp. 515–533, 2006.

[56] J. G. Caporaso, C. L. Lauber, E. K. Costello et al., "Moving pictures of the human microbiome," *Genome Biology*, vol. 12, no. 5, p. R50, 2011.

[57] A. Durbán, J. J. Abellán, N. Jiménez-Hernández, A. Latorre, and A. Moya, "Daily follow-up of bacterial communities in the human gut reveals stable composition and host-specific patterns of interaction," *FEMS Microbiology Ecology*, vol. 81, no. 2, pp. 427–437, 2012.

[58] K. Faust, F. Bauchinger, B. Laroche et al., "Signatures of ecological processes in microbial community time series," *Microbiome*, vol. 6, no. 1, p. 120, 2018.

[59] K. Klemm and V. M. Eguíluz, "Growing scale-free networks with small-world behavior," *Physical Review*, vol. 65, no. 5, Article ID 057102, 2002.

[60] L. Xu, A. D. Paterson, W. Turpin, and W. Xu, "Assessment and selection of competing models for zero-inflated microbiome data," *PLoS One*, vol. 10, Article ID e0129606, 2015.

[61] J. A. Martín-Fernández, C. Barceló-Vidal, and V. Pawlowsky-Glahn, "Zero replacement in compositional data sets," in *Data Analysis, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization*,

H. A. L. Kiers, JP. Rasson, P. J. F. Groenen, and M. Schader, Eds., Springer, Berlin, Heidelberg, 2009.

[62] V. Pawlowsky-Glahn and J. J. Egozcue, "Exploring compositional data with the CoDa-dendrogram," *Austrian Journal of Statistics*, vol. 40, no. 1& 2, pp. 103–113, 2011.

[63] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B*, vol. 64, no. 4, pp. 583–639, 2002.

[64] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman and Hall/CRC, Boca Raton, FA, USA, 2nd edition, 2004.

[65] M. Plummer, "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling DSC 2003 Working Papers," 2003, https://www.r-project.org/conferences/DSC-2003/.

[66] M. J. Denwood, "**Runjags**: An*R*Package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in**JAGS**," *Journal of Statistical Software*, vol. 71, no. 9, pp. 1–25, 2016.

[67] A. Gelman and D. B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–511, 1992.

[68] M. Plummer, N. Best, K. Cowles, and K. Vines, "CODA: convergence diagnosis and output analysis for MCMC," *R News*, vol. 6, no. 1, pp. 7–11, 2006.