Hindawi
Complexity
Volume 2021, Article ID 9951817, 16 pages
https://doi.org/10.1155/2021/9951817

Research Article

A Dirichlet Autoregressive Model for the Analysis of Microbiota Time-Series Data

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 24 March 2021; Accepted 7 July 2021; Published 19 July 2021

Academic Editor: Misako Takayasu

Growing interest in understanding microbiota dynamics has motivated the development of different strategies to model microbiota time series data. However, all of them must tackle the fact that the available data are high-dimensional, posing strong statistical and computational challenges. In order to address this challenge, we propose a Dirichlet autoregressive model with time-varying parameters, which can be directly adapted to explain the effect of groups of taxa, thus reducing the number of parameters estimated by maximum likelihood. A strategy has been implemented which speeds up this estimation. The usefulness of the proposed model is illustrated by application to a case study.

1. Introduction

Recent studies suggest that microbiota, which denotes the collection of bacteria living either in or on the human body, plays a key role in the health status of individuals. In this respect, some studies have pointed out that the maintenance of a stable microbial ecosystem is necessary for a healthy life. In fact, it is known that a disruption of the stable state of the microbiota can be associated with different diseases such as obesity, diabetes, or cancer [1–3]. Therefore, analyzing stability of the microbiota and understanding how quickly it recovers and reaches a new stable state are key questions in the study of the human health status. In this context, longitudinal studies can help to both understand microbiota regularity over time in healthy individuals and study the response of the microbiota to perturbations in disease scenarios.

Many proposals for microbiota data longitudinal analyses use count-based strategies (see, for instance, Section 3.5 in [4] and the references therein). However, more recent approaches suggest considering compositional vectors of relative abundances [5–7]. The reason is that microbiota data are generated through DNA sequencing and they are constrained by an arbitrary constant sum. This is due to the fact that the sequencing instruments used have a fixed upper bound on the number of reads delivered. Therefore, the read count cannot be related to the absolute number of molecules in the input biological sample, and so microbiome datasets must be converted to either relative abundance values or normalized counts [8–11].

On the contrary, the compositional nature of microbiota longitudinal data forces the use of multivariate time series models that take into account the following two features: first, each time series is related to a bacterial taxon and, second, the vector corresponding to each time point represents nonnegative proportions that add up to one. A well-known approach to analyze compositional time series in different scenarios involves transforming the data in order to break the unit sum constraint, and so the use of standard time series techniques is appropriate. Within this strategy, log-ratio or Box-Cox transformations have been considered and Gaussian distributions have been used [12]. However, alternative approaches can also be taken into account, for instance, those based on the use of the Dirichlet distribution [13–15].

Focusing on the analysis of the dynamics of microbial communities, autoregressive models have been considered, with some of them using a standard Lotka–Volterra structure [16–18]. However, these models are based on pairwise interactions and thus fail to capture effects that a third microbe may have on an interacting pair of microbes; see [19] for more details about limitations of models based on Lotka–Volterra structure. A nonparametric approach with an additive structure, which does not presuppose any underlying functional form for community dynamics, has also been proposed [20, 21]. Additive models have the advantage in that they do not need explicit specifications of the functional forms of the relationships between microbes. Nevertheless, this approach admits additivity in the relationships, which is not necessarily realistic for complex microbial communities. Also, a common practice in those works is to consider taxonomic averaging with the aim of reducing the number of parameters in the model, which can lead to inaccurate conclusions [22]. For instance, the role played by important community members can be missed if they are associated with a low abundance, and microbiome stability may be overestimated. Recent works also propose the use of state-space models, which assume that abundances are associated with a real-value hidden state variable vector that evolves through time based on a first-order Markov process and can identify the microbial interaction [23–25]. Other alternatives for the analysis of microbial community temporal dynamics are linear mixed models that provide flexibility in correlated longitudinal data [26, 27] or dynamic Bayesian networks, which are another class of state-space models appropriate to model the interaction of microbial taxa [28].

In this paper, we model relative abundances of microbial taxa with a Dirichlet distribution with time-varying parameters. We assume that these relative abundances, after a log-ratio transformation, can be explained by an autoregressive structure which takes into account the effect of the bacterial community as a whole. This proposal can be useful to understand the relationships between microbes and the identification of keystone members of the microbial ecosystem that may play an important role. It is worth noting that the Dirichlet distribution has a strong 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 [29]. However, it has also been found useful when used as a conditional distribution; see, for instance, [30, 31].

One important feature of our proposal is the consideration of the bacterial community effect as a whole, by recurring to the geometric mean, in order to reduce dimensionality. This formulation allows us to encapsulate information and decrease the number of parameters to estimate without removing or grouping microbial taxa. To the best of our knowledge, this is the first model developed for microbiota time series based on Dirichlet distribution with time-varying parameters.

The paper proceeds as follows: In Section 2, we present some basic definitions and describe the proposed model and in Section 3 we illustrate the performance of the model with a case study. Finally, some conclusions are drawn and directions for future research are suggested.

2. Model

2.1. Basic Definitions and Preliminaries. Let $\mathbf{y} = (y_1, y_2, \ldots, y_K)$ be a K-dimensional random vector which satisfies that $\sum_{i=1}^{K} y_i = 1$ and $0 < y_i < 1$ for all $1 \leq i \leq K$. The random vector $\mathbf{y}$ follows a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, $\alpha_i > 0$, for all $i \in \{1, 2, \ldots, K\}$, if its probability density distribution is

$$f(\mathbf{y}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} y_i^{\alpha_i - 1}, \tag{1}$$

where the normalizing constant, $B(\boldsymbol{\alpha})$, is the multivariate beta function which can be defined in terms of the gamma function, $\Gamma(\cdot)$:

$$B(\boldsymbol{\alpha}) = \prod_{i=1}^{K} \frac{\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}. \tag{2}$$

Considering $\tau = \sum_{i=1}^{K} \alpha_i$, the expectation of each component, y_i, is defined as $E(y_i) = \alpha_i/\tau$ and the variance as $\mathrm{Var}(y_i) = \alpha_i(\tau - \alpha_i)/\tau^2(\tau + 1)$. The covariance for $i \neq j$ is $\mathrm{Cov}(y_i, y_j) = -\alpha_i \alpha_j/\tau^2(\tau + 1)$.

Compositional time series are multivariate time series in which the observation vector at time t is nonnegative proportions whose sum is 1. Historically, such time series have been modeled by considering transformation of the observations and modeling them with standard multivariate techniques using Gaussian distributions. However, alternative approaches have also been considered in which the original data are modeled directly, that is, how they are observed experimentally. In this case, probability distributions on the simplex must be considered and a very popular distribution is the Dirichlet distribution.

Longitudinal data on relative taxa abundances can be regarded as a compositional time series where the vector of relative abundances corresponding to each time point is an element of the simplex:

$$S^K = \left\{ (y_1, y_2, \ldots, y_K) \colon y_i > 0 \text{ and } \sum_{i=1}^{K} y_i = 1 \right\}. \tag{3}$$

Taking into account the fact that the Dirichlet distribution with covariates and time-varying parameters can provide an adaptable covariance structure [15, 30, 31], our proposal is based on this probability distribution and on a reparameterization defined by the well-known additive log-ratio transformation (alr transformation). Note that alternative forms of transformation can also be proposed, such as the centered log-ratio transformation (clr transformation) or the isometric log-ratio transformation (ilr transformation). See [29, 32], for details related to these transformations. The additive log-ratio transformation of index K is the one-to-one linear transformation from S^K to $\mathbb{R}^{K-1}$ defined as

$$\operatorname{alr}(y_1, y_2, \ldots, y_K) = \left(\ln\left(\frac{y_1}{y_K}\right), \ln\left(\frac{y_2}{y_K}\right), \ldots, \ln\left(\frac{y_{K-1}}{y_K}\right) \right). \tag{4}$$

Zheng and Chen analyze the consideration of the additive log-ratio (alr) transformation and the centered log-ratio (clr) transformation as link function in an autoregressive moving-average model with the Dirichlet distribution [15]. In our case, we have also taken into consideration the alr transformation as the link function in an autoregressive structure but have redefined the effect of the relative abundance of the rest of the microbial community in order to reduce the number of parameters. Note that, in a standard autoregressive model (VAR model), the effect of the remaining taxa on each taxon must be defined by adding each of them as an additive term and this option increases model dimensionality. We have considered the effect of the rest of the taxa on average. Compared with the approach proposed by Zheng and Chen, we have not considered a moving-average component to reduce dimensionality. The alr transformation has been considered for biological purposes. This option allows the comparison of two particular taxa. Taking into account y_K as reference, we can consider that

$$\ln\left(\frac{y_i}{y_{i'}}\right) = \ln\left(\frac{y_i}{y_K} : \frac{y_i'}{y_K}\right) = \ln\left(\frac{y_i}{y_K}\right) - \ln\left(\frac{y_i'}{y_K}\right). \tag{5}$$

2.2. Model Formulation. Let $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{Kt})$ be the vector of relative abundances at time t, so that $y_{it} \in (0, 1)$ and $y_{1t} + y_{2t} + \cdots + y_{Kt} = 1$. We assume that the vector $\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1$ follows a Dirichlet distribution with positive parameters $\boldsymbol{\alpha}_t = (\alpha_{1t}, \alpha_{2t}, \ldots, \alpha_{Kt})$:

$$\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1 \sim \operatorname{Dir}(\boldsymbol{\alpha}_t). \tag{6}$$

In order to link $\boldsymbol{\alpha}_t$ with $\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1$, we propose $\ln(\alpha_{jt}/\alpha_{Kt}) = \eta_{jt}$, where η_{jt} is defined as

$$\eta_{jt} = a_{j0} + \sum_{p=1}^{P} a_{jp} \cdot \ln\left(\frac{y_{jt-p}}{y_{Kt-p}}\right) + \sum_{p=1}^{P} b_{jp} \cdot \ln \frac{\left(\prod_{s=1 s \neq j}^{K-1} y_{st-p}\right)^{1/K-2}}{y_{Kt-p}}, \quad j = 1, 2, \ldots, K-1. \tag{7}$$

Note that additive log-ratio data transformation and first-order Taylor approximation enable us to state that

$$E\left(\ln\left(\frac{y_{jt}}{y_{Kt}}\right)\right) = E\left(\operatorname{alr}\left(y_{jt}\right)\right) \approx \operatorname{alr}\left(E\left(y_{jt}\right)\right) = \operatorname{alr}\left(\frac{\alpha_{jt}}{\tau}\right) = \ln\left(\frac{\alpha_{jt}}{\alpha_{Kt}}\right), \tag{8}$$

and thus to assume that taxon relative abundance over time is dependent on their own relative abundance as well as on the effect of the relative abundance of the rest of the microbial community at the previous time points. Note that a_{jp} are coefficients associated with relative abundance of taxa j at the previous time points and b_{jp} are those associated with the effect of the relative abundance of the rest of the microbial community on (geometric) average. The coefficient P is the order of the model and a_{j0} represents the expected mean value of $E(\ln(y_{jt}/y_{Kt}))$ when significance is lacking for both the effect of the relative abundance of taxa j at the previous time points and the effect of the remaining relative abundance of the rest of the microbial community on (geometric) average. It is worth emphasizing that in equation (7) each taxon abundance is evaluated with respect to the abundance of taxon y_K and thus positive values signify that the taxon in the numerator has more weight than taxon y_K and conversely for negative values.

The parameter $\tau_t = \alpha_{1t} + \alpha_{2t} + \cdots + \alpha_{Kt}$ is the concentration parameter of the Dirichlet distribution and must also be estimated. We consider that it is a time invariant parameter; that is, $\tau_t = \tau$ for each t. In order to reduce dimensionality, equal concentration for probability mass has been admitted at each time point.

Therefore, using the expression $\ln(\alpha_{jt}/\alpha_{Kt}) = \eta_{jt}$, we can determine α_{jt} and α_{Kt} as follows:

$$\alpha_{jt} = \frac{e^{\eta_{jt}} \cdot \tau_t}{1 + e^{\eta_{1t}} + \cdots + e^{\eta_{K-1t}}}, \tag{9}$$

$$\alpha_{Kt} = \frac{\tau_t}{1 + e^{\eta_{1t}} + \cdots + e^{\eta_{K-1t}}}. \tag{10}$$

From these equalities, it is easy to calculate the expected value $E(y_{it}) = \alpha_{it}/\tau_t$ and variance $\operatorname{Var}(y_{it}) = \alpha_{it}(\tau_t - \alpha_{it})/\tau_t^2(\tau_t + 1)$ for each taxon i at time t.

3. Case Study

In order to demonstrate the utility of our proposal, we analyze an available time series microbiome dataset. In this section, we show a summary of the results obtained. We have studied the applicability on prediction, variability analysis, and trends clustering. There are other works taking similar approaches, which also provide applications to predict or detect relevant taxa, albeit only partially, and have different features. See, for instance, [33–35]. Conversely, our proposal is an alternative integrated framework, applicable to more than one context.

In the case study performed, we have selected family as the taxonomic level and y_K is defined as the family with the greatest relative abundance. We must point out that only a first-order autoregression has been considered, $P = 1$ in equation (7), because higher-order autoregressions were not significant. Given the concern of the presence of zeros, we have chosen only families whose relative abundance is bigger

than 1% for each single time point but this cutoff value can be arbitrarily defined.

3.1. Dataset. We have analyzed the database generated by the 16S rRNA gene sequencing of stool samples of a healthy male and a female with irritable bowel syndrome (IBS). Fecal bacteria from the healthy male were monitored for 15 consecutive days. The IBS patient's fecal samples were collected in the morning on alternate days the first week and once a week thereafter. This longitudinal dataset was studied by Durbán et al. in [36, 37].

Figure 1 shows the relative abundances of the microbial families in samples of the above-mentioned individuals. Note that, at family level, samples were quite similar. The families Bacteroidaceae, Porphyromonadaceae, Rikenellaceae, and Ruminococcaceae are present in both samples, while the families Erysipelotrichaceae and Lachnospiraceae were present in the individual with IBS. We must point out that in both the healthy and the IBS individuals several undefined families were also present, which we classified collectively as *Other*. Taxa that did not meet the 1% threshold have also been rolled into the *Other* category. This group of undefined detections has been taken as reference, y_K. It is should be pointed out that equivalent results would be obtained if we chose another component (another *family*) as reference. Our proposal satisfies the permutation invariance principle. See [29] for details.

3.2. Predicting Temporal Behavior of Microbial Taxa. In this section, we describe how the proposed model can effectively predict the future dynamics of a microbial community. Modeling microbiota time series data and using models for predicting temporal dynamics of a future state can help to gain a better understanding of the different roles played by microbes.

Figure 2 displays both the predicted relative abundance, $E(y_{it})$, and the experimentally reported values for each biological family in the healthy male. The same fit was also performed for the IBS patient, the results of which are shown in Figure 3.

In order to predict the future evolution of families, $E(y_{it+h})$, and variance, $\mathrm{Var}(y_{it+h})$, expressions (9) and (10) must be evaluated at time $t + h$. Table 1 shows the estimated values for the parameters a_{j0}, a_{j1}, and b_{j1}. The estimation procedure is detailed in the Appendix.

To analyze the predictive performance of our approach, we have compared it with a simple method which predicts the same value as the previous time point. This simple method will provide us with a baseline value. For this comparative purpose, the data for the last three time points ($L = 3$) have been left to evaluate the ability of our model to predict the relative abundance of taxa, and the first twelve time points have been used to estimate the parameters. As a measure of predictive efficiency, the sum of absolute errors (SAE) has been considered. This index is defined as

$$\mathrm{SAE} = \sum_{i=1}^{K} \sum_{l=1}^{L} \left| y_{il} - \widehat{y}_{il} \right|, \tag{11}$$

where y_{il} is the relative abundance of taxa i at time points which have been left to evaluate the predictive efficiency and $\widehat{y}_{il}$ is the corresponding predicted value.

In Table 2, we can see that, under the proposed model, SAE is greater than the sum of absolute errors under the baseline approach in the healthy male. The analysis indicates that the predictions for the families *Other* and Rikenellaceae calculated by our proposal are not accurate for the healthy male. However, this does not happen for the remaining families. For the IBS patient, the SAE index shows that our proposal performs well. In this case, all the predictions calculated with our proposal are more accurate than the ones provided by the baseline method.

We have also evaluated the predictive efficiency of our model on other real datasets. Lloyd-Price et al. [38] tracked 132 subjects (Crohn's diseases or ulcerative colitis patients) for one year each to analyze microbial activity during the disease (up to 24 time points each). We have analyzed two of them. They are two school-aged subjects diagnosed with Crohn's disease recruited from Cedars-Sinai Medical Center in Los Angeles and Cincinnati Children's Hospital, respectively. The subject recruited from Cedars-Sinai Medical Center reported antibiotic use. Caporaso et al. [39] present a human microbiota time series analysis of two healthy individuals over 396 time points at four body sites. We have used their available longitudinal time series data of gut microbiome samples from both subjects over 80 time points.

We have also investigated predictive efficiency of the proposed model using two simulated datasets. The simulation study has been carried out considering five and eight microbial entities, respectively. Following the proposal established in [40], we have generated its time series over 30 time points with a generalized Lotka–Volterra structure. This approach allows us to simulate the temporal dynamics of a bacterial community considering the interentity interactions as input. To generate the interaction matrix, we have taken into account the algorithm proposed by Klemm and Eguiluz in [41]. This algorithm generates a modular and scale-free interaction matrix that reproduces properties of a microbial network. We assigned interaction considering diagonal values to −1 and off-diagonal values by sampling from a uniform distribution between 0 and 1. We have also set the modularity parameter of the Klemm and Eguiluz algorithm to 4 and 6, for the database which considers five taxa and eight taxa, respectively, and the interaction matrix connectance to 0.4 and 0.3, respectively. Note that these parameters allow us to both measure the strength of division of a network into subcommunities and define the interaction probability between entities, respectively. The interaction matrix has been generated with a positive interactions percentage equal to 64%. We have used the R package presented in [40] called *seqtime*.

In all these alternative datasets, we have also selected family as the taxonomic level and families whose relative abundance is bigger than 1% for each single time point have also been considered. The last three time points have also been left to evaluate the predictive ability of our proposal. Table 2 shows the predictive efficiency of our approach on the alternative datasets. The results display the predictive

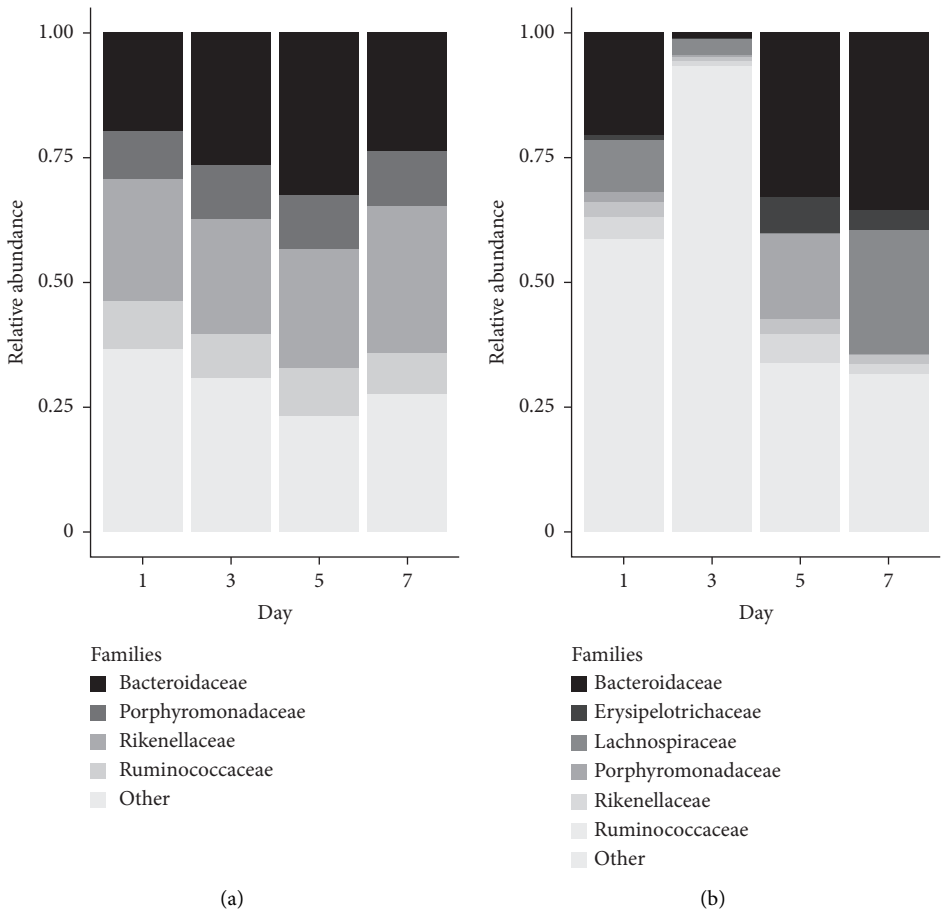


Figure 1: Relative abundances of microbial families in samples of a healthy male and a female IBS patient. In this figure, the fecal samples collected on alternate days during the follow-up period have been considered. (a) Healthy male. (b) Female IBS patient.

performance of our proposal compared to the baseline method. Additionally, in order to increase the evidence of our proposal, we have also compared it against the well-known TGP-CODA method proposed by Äijö et al. in [33]. The results obtained on the real and simulated datasets also demonstrate validity of our model for predicting temporal behavior of microbial taxa.

3.3. Analyzing Variations in Temporal Behavior of Microbial Taxa. The aim of this case study is to illustrate how our model can be useful to show the relationship between microbiome variability and host health status. In this respect, the estimates of the variances for each microbial taxon, $\mathrm{Var}(y_{it})$, differentiate between healthy and dysbiotic microbiota (Figure 4). In addition, boxplots combining all the families also show a clear difference across time between healthy and unhealthy microbiota. We can appreciate that median, interquartile range, and whiskers are larger in the IBS patient. In the healthy individual, all families present a stable variance. As we mentioned before, the estimation procedure for the parameters of the model, α_{it}, and thus for the variance, $\mathrm{Var}(y_{it})$, is described in the Appendix.

In order to evaluate how our variance analysis performs in terms of indicating host health status, we have compared our proposal to a microbial trend analysis (MTA). Wang et al. [42] propose an MTA framework for longitudinal microbiome data analysis. This proposal can capture the common dynamic patterns on the microbial community and identify the dominant taxa. Additionally, MTA can also classify individuals based on its longitudinal microbial profiling. We have used MTA because this toolbox is similar to our proposal. Note that both are integrated frameworks which allow several applications to microbial longitudinal data. In the IBS case study, the distance-based classification algorithm proposed in MTA framework also classifies the subjects (IBS patient and healthy subject) into different groups.

We have also analyzed the ability of our proposal to distinguish the microbial dynamics between subjects in alternative scenarios. We have also used the simulated datasets described before and the real datasets studied by Lloyd-Price et al. and Caporaso et al., respectively. Figure 5 displays the variance time series for the alternative scenarios considered. We can also appreciate the differences, although in these cases they are lower than those observed in the IBS scenario. For these aforementioned alternative datasets, we have also compared the performance of our variance analysis to that of the MTA framework. In all scenarios, the MTA approach also classifies the subjects in different groups.

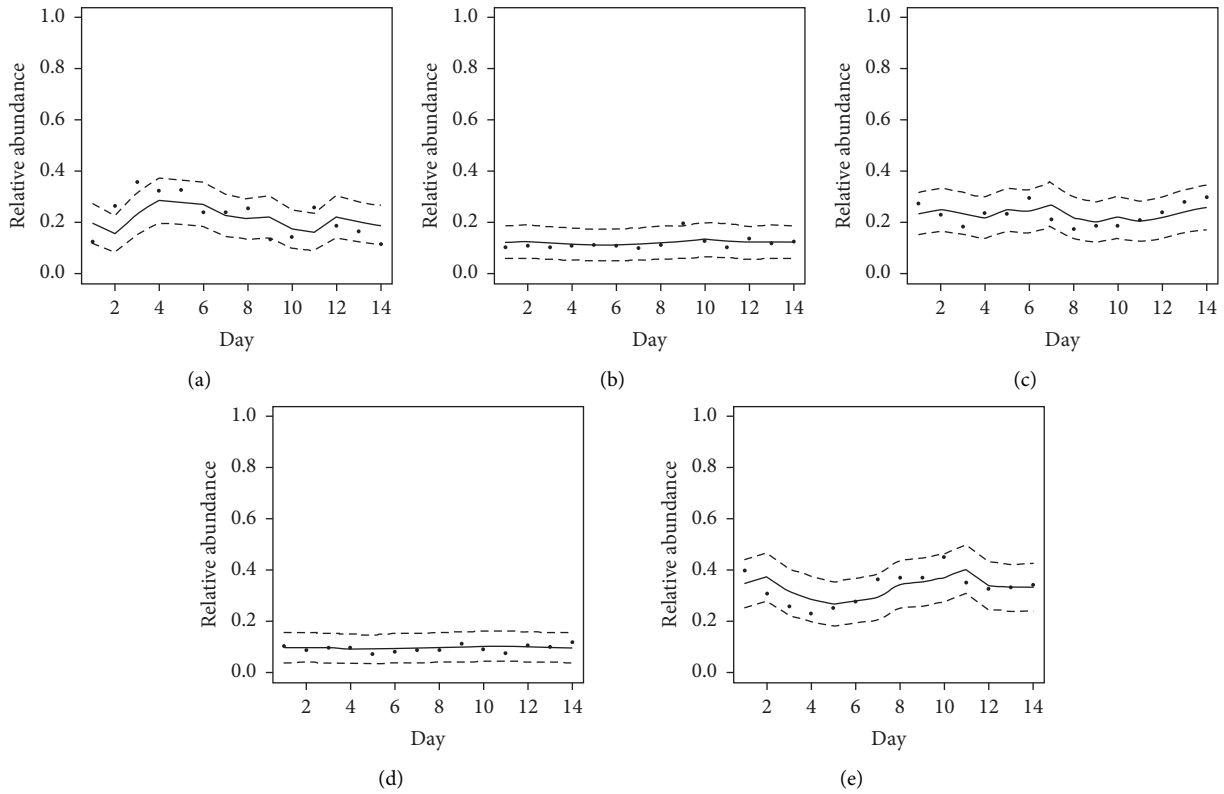


Figure 2: Model fitting each family in the healthy male. Relative abundances reported are represented by solid points. The solid line shows the estimated values. Standard deviation interval has also been plotted with dashed lines, $E(y_{it}) \pm 2 \cdot \sqrt{\mathrm{Var}(y_{it})}$. (a) Bacteroidaceae. (b) Porphyromonadaceae. (c) Rikenellaceae. (d) Ruminococcaceae. (e) Other.

Additionally, we have also compared our proposal against MITRE, a supervised machine learning method proposed by Bogart and others in [34]. This proposal also allows predicting or inferring the status of the host analyzing microbiota time series data. Table 3 displays the probability associated with the host's status. As in our approach, MITRE also clearly discriminates the host's status on each scenario analyzed.

3.4. Clustering Groups of Taxa Sharing a Similar Pattern over Time. Another important goal while analyzing microbiota time series data is the detection of groups of taxa which present similar trends over a timeframe. The detection of taxa with similar temporal dynamics versus taxa with alternative patterns can help to understand the principles that define the microbiome in health or cause dysbiosis in disease.

It should be remembered that, in equation (7), a_{j0} represents a family-specific intercept that picks up the average relative abundance of family j versus the relative abundance of family y_K over time; a_{j1} are related to the intrinsic dynamics for each family versus family y_K and b_{j1} to the dynamics of the remaining families considered in geometric mean.

Figure 6 displays the PCA biplots with the first two principal components that together explain 98.93% and 99.45% of the total variance of the IBS patient and healthy individual, respectively. We can observe that a_{j1} and b_{j1} are positioned on opposed quadrants. They are negatively correlated. Note that families close to each other in the biplot represent observations with similar values. Therefore, we can determinate which families have measurements that are the most similar to each other.

In the PCA biplot for the IBS patient, we observe that Porphyromonadaceae and Rikenellaceae are associated with the highest values of a_{j1} and the lowest values of b_{j1}. We can also note that Lachnospiraceae is associated with the highest values of b_{j1} and the lowest ones of a_{j1}. On the other hand, in the healthy subject, we note that Rikenellaceae and Ruminococcaceae are now the families that are associated with the highest values of b_{j1} and the lowest values of a_{j1}, respectively.

Taking into account the fact that a_{j1} are related to the intrinsic dynamics for each family and b_{j1} are related to the dynamics of the remaining families in geometric mean, we are be able to cluster families sharing a similar pattern over time. For instance, in the IBS patient, the temporal dynamics of the relative abundance of Lachnospiraceae (versus family y_K) is strongly related to that of the remaining families in geometric mean. However, the dynamics of the relative abundance of both Porphyromonadaceae and Rikenellaceae (versus family y_K) are not so closely associated with it. Remember that in our proposal all relative abundances are analyzed with respect to the relative abundance of y_K and this one clusters undefined taxa.

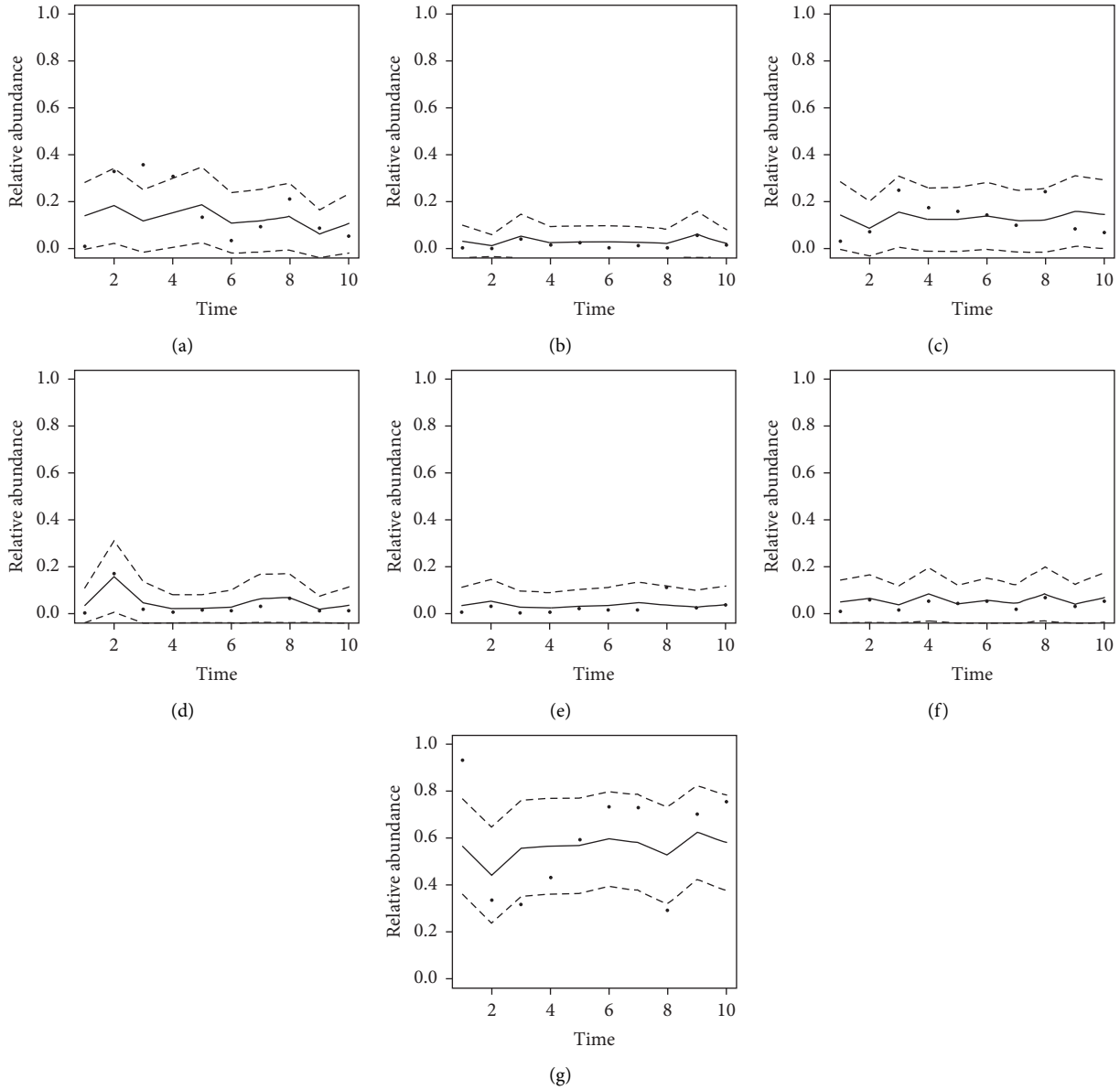


Figure 3: Model fitting each family in the IBS patient. Relative abundances reported are represented by solid points. The solid line shows the estimated values. Standard deviation interval has also been plotted with dashed lines, $E(y_{it}) \pm 2 \cdot \sqrt{\mathrm{Var}(y_{it})}$. (a) Bacteroidaceae. (b) Erysipelotrichaceae. (c) Lachnospiraceae. (d) Porphyromonadaceae. (e) Rikenellaceae. (f) Ruminococcaceae. (g) Other.

Table 1: Estimated parameter values for IBS patient (top) and healthy male (bottom).

	a_{j0}	a_{j1}	b_{j1}
IBS patient			
Bacteroidaceae	−1.831	−0.536	0.186
Erysipelotrichaceae	−4.614	−0.571	0.420
Lachnospiraceae	−1.455	−3.701	2.473
Porphyromonadaceae	−5.417	0.358	−1.412
Rikenellaceae	−3.906	0.457	−0.910
Ruminococcaceae	−3.833	−0.555	0.052
Healthy male			
Bacteroidaceae	−3.595	2.518	−4.920
Porphyromonadaceae	0.116	1.705	−1.519
Rikenellaceae	0.803	−1.665	1.610
Ruminococcaceae	−2.843	−2.127	1.486

Table 2: Predictive efficiency. The sum of absolute errors (SAE $\times 10^1$) is shown.

Dataset	Proposed model	Baseline approach	TGP-CODA
Durbán et al.			
Healthy male	5.48	3.78	6.05
IBS patient	12.74	19.76	14.39
Lloyd-Price et al.			
Crohn's disease patient (with antibiotic use)	5.30	10.56	8.10
Crohn's disease patient (without antibiotic use)	2.62	11.28	8.60
Caporaso et al.			
Female	3.83	5.22	5.11
Male	6.23	9.27	5.78
Simulated datasets			
With 5 simulated taxa	6.61	8.78	7.93
With 8 simulated taxa	2.50	4.49	3.65

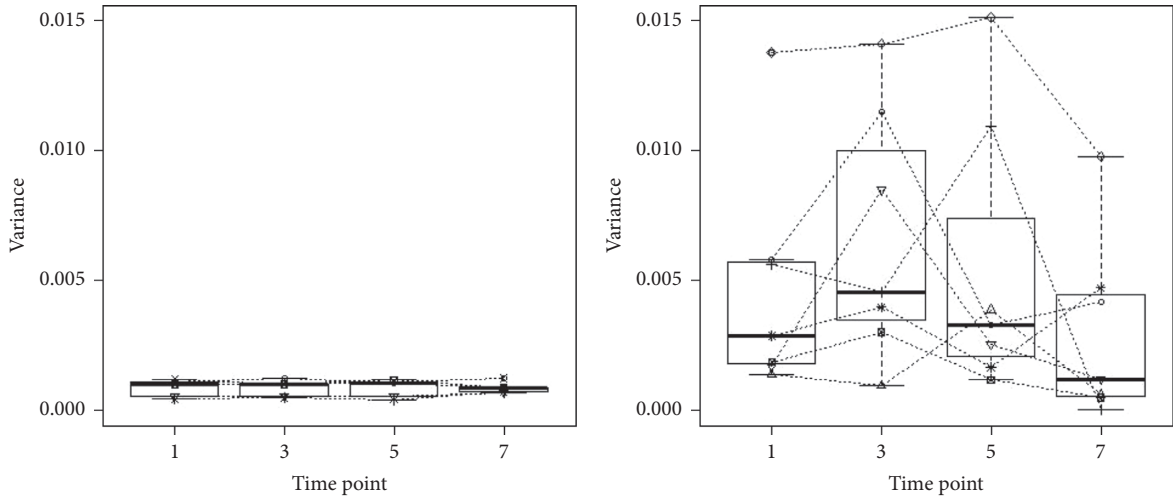


Figure 4: Variance time series for the healthy male and the female IBS patient. The families have been represented by the following markers: Bacteroidaceae (∘), Porphyromonadaceae (◇), Rikenellaceae (▽), Ruminococcaceae (⊠), Erysipelotrichaceae (△), Lachnospiraceae (+), and Other (×). (a) Healthy patient. (b) Unhealthy patient.

We have also evaluated the identification of taxa by comparing our proposal to the microbial trend analysis (MTA). Table 4 shows the output of MTA related to the contribution of each taxon to the longitudinal microbial profiling for each individual.

We can observe a strong correspondence between the classification of taxa defined by our proposal and the contribution values to the longitudinal microbial profiling presented by MTA. In the IBS patient, our model points to Bacteroidaceae and Lachnospiraceae as relevant taxa compared to the rest. Note that Bacteroidaceae is associated with the highest values of a_{j0} and Lachnospiraceae to the highest values of b_{j1}. We can corroborate this using the MTA analysis. In this case, Bacteroidaceae and Lachnospiraceae are the families with greater contribution to the longitudinal microbial profiling of this patient with a contribution of 0.200 and 0.180, respectively. The other families grouped by the PCA biplot present a similar contribution value around 0.04. Note that *Other*, which has a contribution of 0.959, has been taken as reference in our proposal. In the healthy individual, we can observe that our approach detects Bacteroidaceae and Rikenellaceae as the families associated with the highest values of a_{j1} and b_{j1}, respectively. The MTA analysis indicates that these families are the most relevant in the longitudinal microbial profiling of this subject, with a contribution of 0.459 and 0.476, respectively.

In the PCA biplot for the subjects reported by Lloyd-Price et al. (Figure 7), we observe the relevant role of Clostridiaceae, which is clearly associated with higher values of coefficient a_{j1} (intrinsic dynamics for each family) in the subject with antibiotic use and with higher values of coefficient a_{j0} (average relative abundance of family j versus the relative abundance of family y_K over time) in the subject without antibiotic use. Its relevant role is also highlighted by the MTA approach. See Table 5. It should be remembered that a higher absolute value indicates a stronger contribution to a common trend.

With respect to the healthy female studied by Caporaso et al., we note the relevant association between *Other* and a_{j0}, which is also emphasized by the MTA analysis with a high

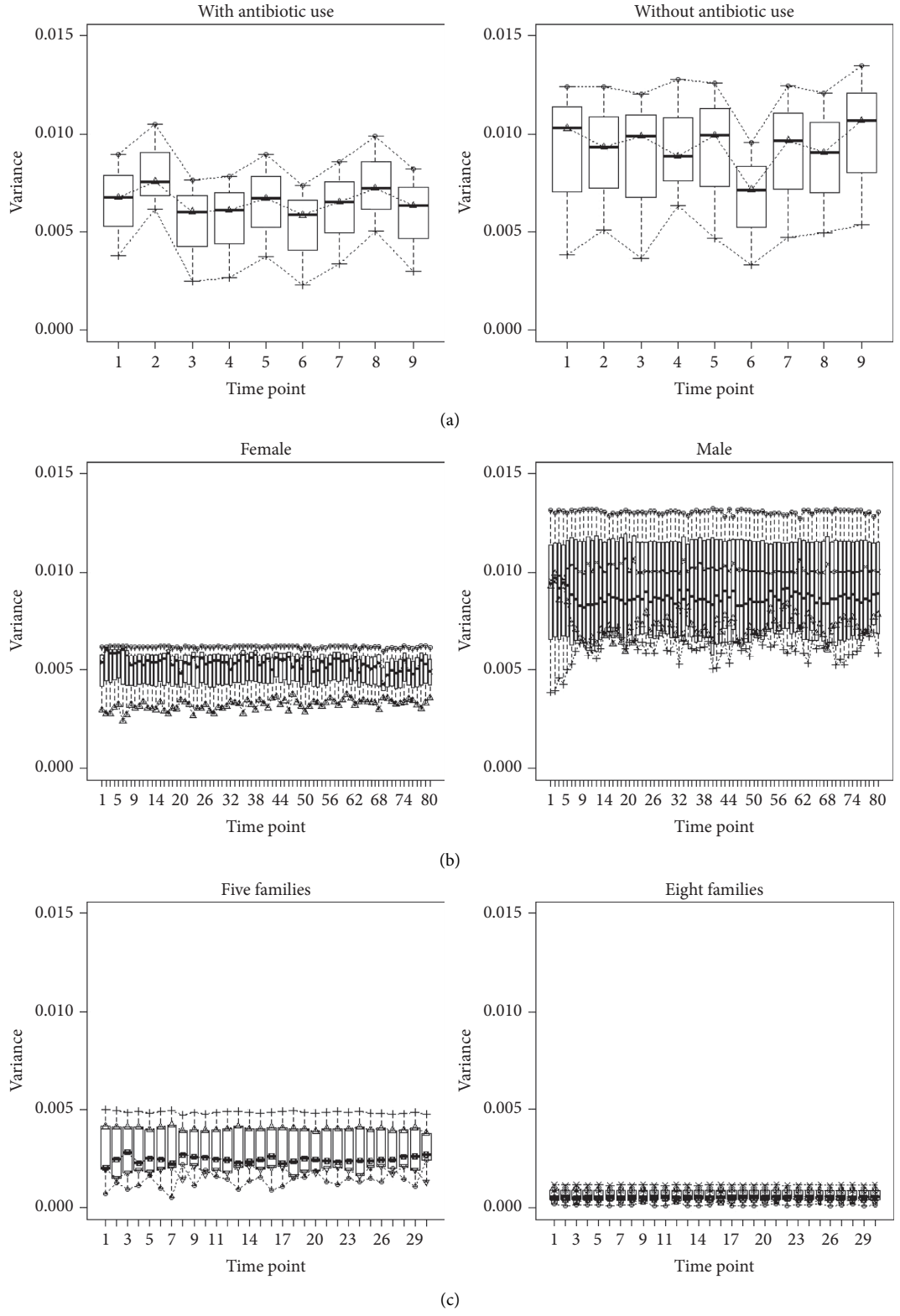


Figure 5: Variance time series for the subjects considered by Lloyd-Price et al. (a), Caporaso et al. (b), and simulated datasets (c), respectively. In this case, the markers are as follows: Other (+), Clostridiaceae (△), and Bacteroidaceae (°), for subjects reported by Lloyd-Price et al., and Other (×), Lachnospiraceae (△), and Bacteroidaceae (°), for individuals considered by Caporaso et al.

Table 3: MITRE exploratory analysis on the real and simulated datasets considered. The probabilities are associated with host status.

Dataset	Probability	Status
Durbán et al.	0.947	Healthy
Lloyd-Price et al.	0.998	Without antibiotic use
Caporaso et al.	0.960	Female
Simulated datasets	0.993	With 5 simulated taxa

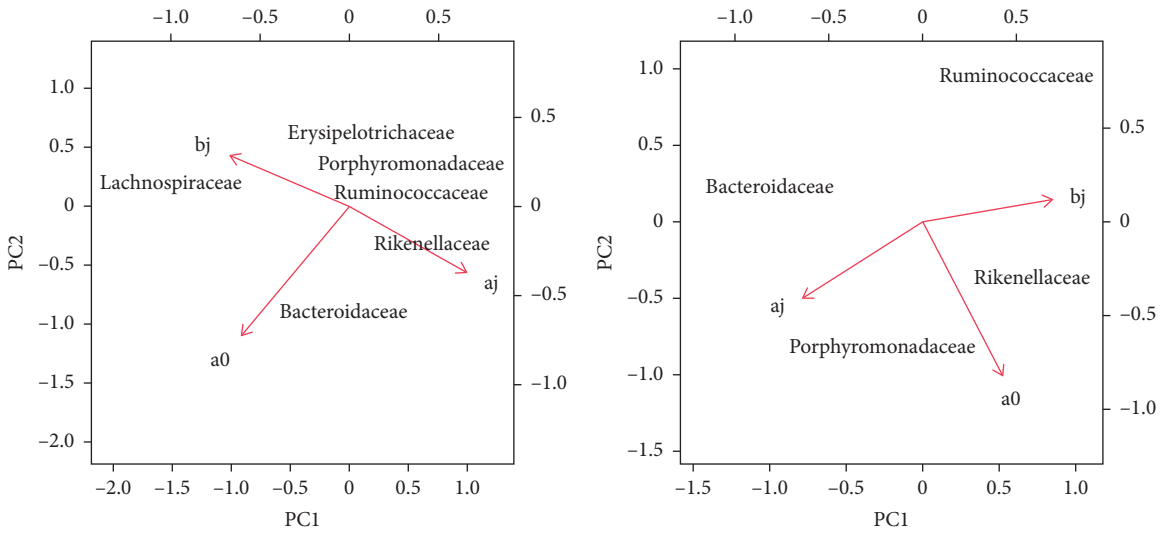


Figure 6: PCA Biplot for IBS patient (a) and healthy male (b).

Table 4: MTA output related to the contribution of each taxon to the longitudinal microbial profiling in IBS scenario. Note that a higher absolute value indicates a stronger contribution to the common trend.

IBS patient		Healthy male	
Bacteroidaceae	0.200	Bacteroidaceae	0.459
Erysipelotrichaceae	0.026	Porphyromonadaceae	0.244
Lachnospiraceae	0.180	Rikenellaceae	0.476
Porphyromonadaceae	0.043	Ruminococcaceae	0.192
Rikenellaceae	0.037	Other	0.682
Ruminococcaceae	0.058		
Other	0.959		

absolute value. We also appreciate that this association is not as narrow as in the male studied. This fact is corroborated by the MTA analysis. Its absolute associated value is lower in the male than in the female. In the PCA biplot for the male individual, we also observe that Lachnospiraceae and Ruminococcaceae are grouped close to a_{j1} and this fact is also detected by our MTA analysis with similar values around −0.20. In these alternative scenarios, Bacteroidaceae has been considered as reference family, y_K.

The resulting PCA biplot for five simulated taxa clearly shows that taxon 1 a_{j0} correlates with having a high MTA contribution value equal to 0.515. Additionally, it displays that taxon 3 and taxon 4 are associated with a_{j1} at a similar level, with contribution values to the longitudinal microbial profiling equal to 0.209 and 0.258, respectively. From this biplot, we can also conclude that taxon 5 is associated with the coefficient b_{j1}. By contrast, the biplot for eight simulated taxa indicates that taxa 4 and 7 are grouped and closely related to a_{j0} and taxon 1, taxon 2, and taxon 5 are also grouped and related to a_{j1}. This evidence has also been confirmed by its higher MTA contribution values equal to 0.449 and 0.428, and 0.212, 0.257 and 0.242, respectively. In this simulated case, taxon 2 and taxon 3 have been considered as reference, respectively.

The results of all these complementary analyses have also corroborated that the taxa grouped by the PCA biplot contribute similarly to the MTA values, and the families clearly located close to the axes present higher values.

In order to address the relationship between taxa, we have also analyzed association using correlation approach. Figures 8 and 9 show the correlation matrices provided by *microbiome* R package [43]. Note that these correlation indexes are provided considering the clr transformation data.

In the PCA biplot for the IBS patient (Figure 6), we observe that Porphyromonadaceae and Rikenellaceae had a similar pattern (defined by the highest values of a_{j1} and the lowest values of b_{j1}) and it can also be observed in the correlation analysis. We can note that Porphyromonadaceae and Rikenellaceae present a significant positive correlation;

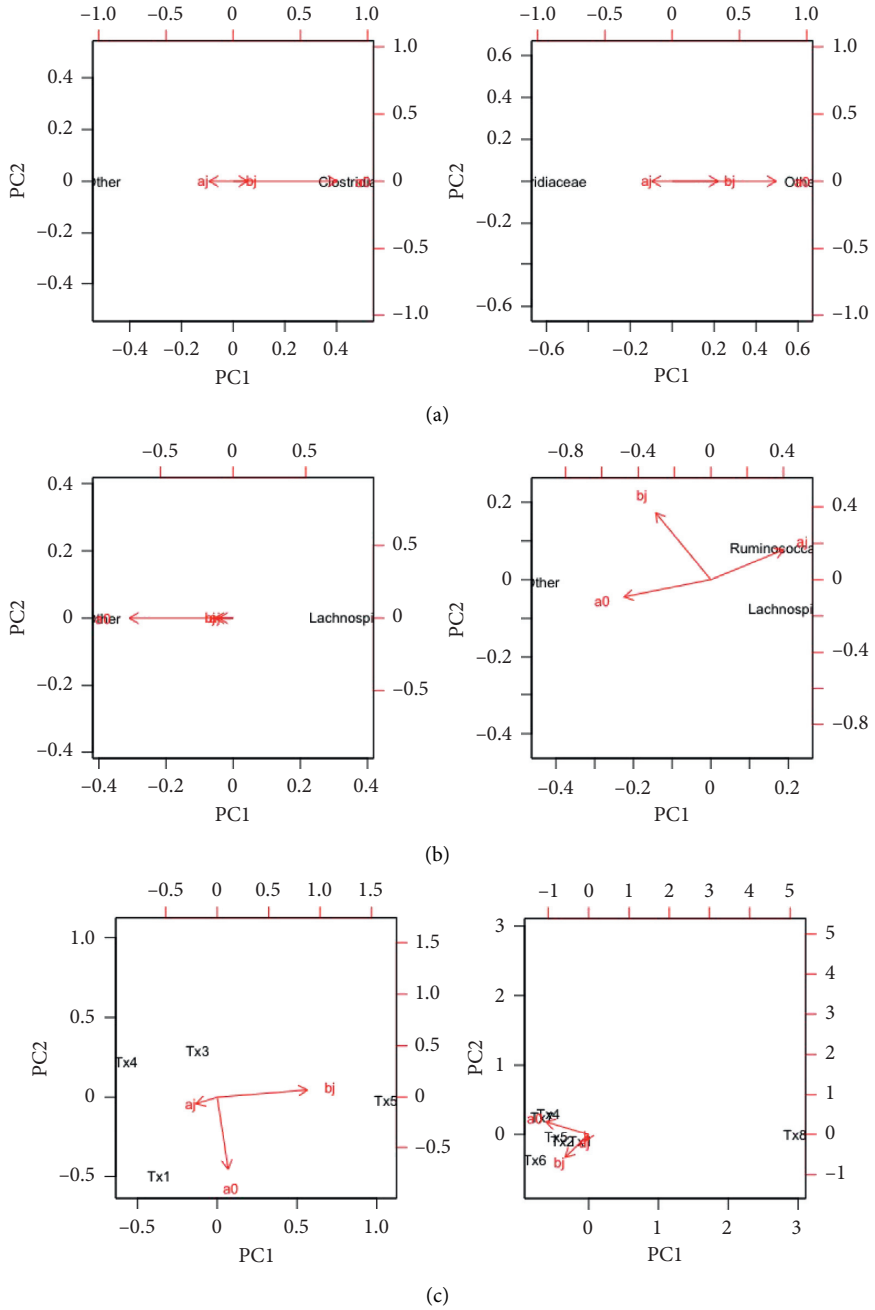


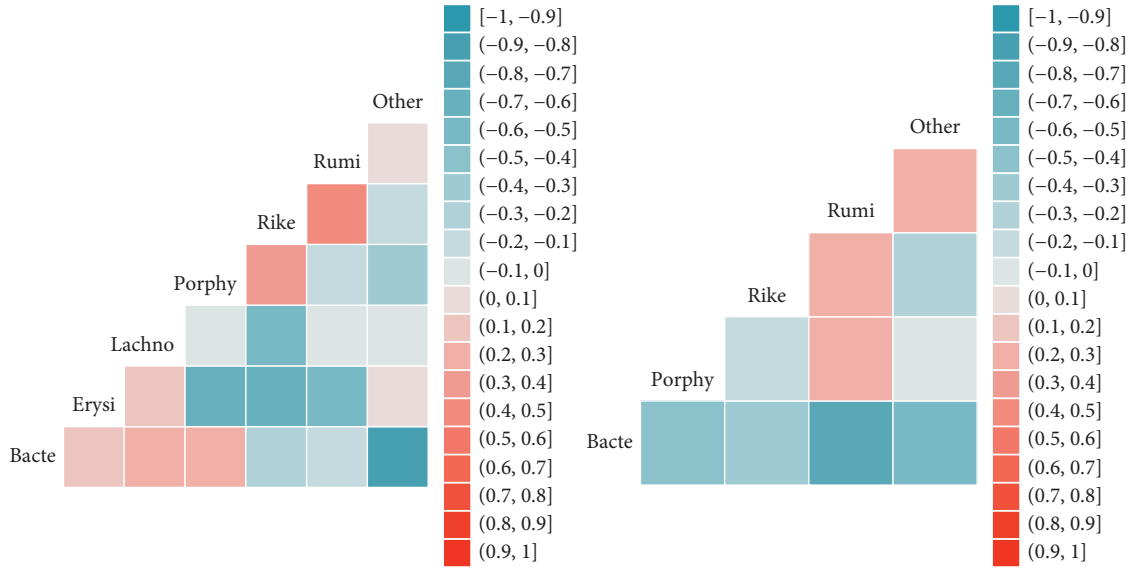
FIGURE 7: PCA biplot for the subjects reported by Lloyd-Price et al. ((a); (A) and (B)), individuals recorded by Caporaso et al. ((b); (C) and (D)) and simulated datasets ((c); (E) and (F)).

see Figure 8. We can also note that Lachnospiraceae presents an alternative pattern which is associated with the highest values of b_{j1} and the lowest ones of a_{j1}. The correlation matrix displays negative correlation between Lachnospiraceae and Porphyromonadaceae and between Lachnospiraceae and Rikenellaceae.

In the healthy subject, we note that Rikenellaceae and Ruminococcaceae are now the families that are associated with the highest values of b_{j1} and the lowest values of a_{j1}, respectively. This association between Rikenellaceae and Ruminococcaceae can also be observed in the correlation matrix with a positive correlation between these taxa. We can appreciate that the well-known correlation analysis corroborates the associations between taxa defined by our approach. In the PCA biplot for the subjects reported by Lloyd-Price et al. (Figure 7), we observe the difference between Clostridiaceae and *Other*. This difference is also pointed out in the correlation analysis, which is shown with a negative index. Analyzing the rest of scenarios, we can also appreciate that similar patterns defined by our approach are associated with positive correlations and nonsimilar patterns are associated with negative ones.

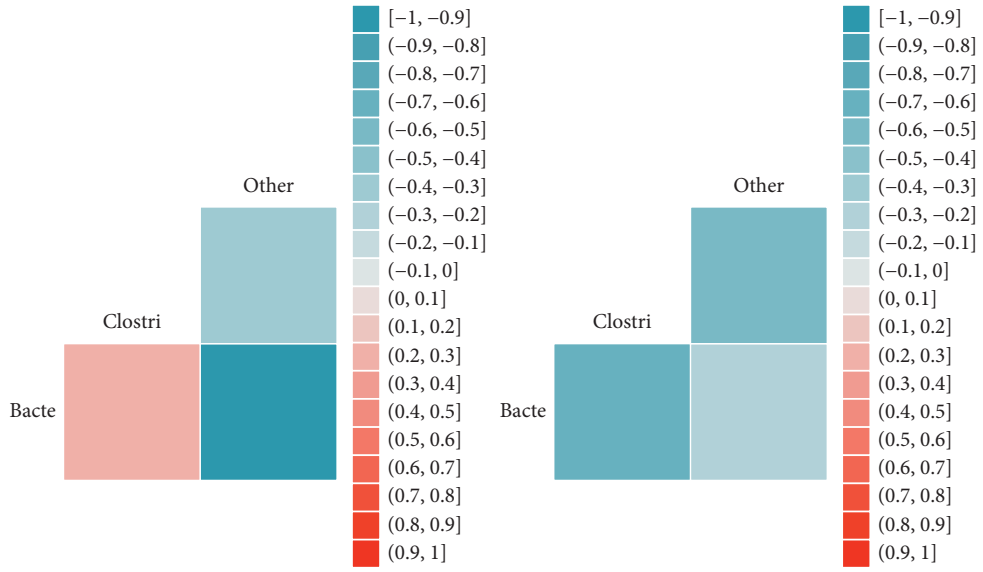
Table 5: MTA output related to the contribution of each taxon to the longitudinal microbial profiling in real and simulated datasets, respectively.

Lloyd-price et al.			
Without antibiotic use		With antibiotic use	
Bacteroidaceae	−0.968	Bacteroidaceae	−0.975
Clostridiaceae	−0.233	Clostridiaceae	−0.195
Other	−0.093	Other	−0.099
Caporaso et al.			
Female		Male	
Bacteroidaceae	−0.844	Bacteroidaceae	−0.832
Lachnospiraceae	−0.224	Lachnospiraceae	−0.258
Other	−0.487	Other	−0.450
		Ruminococcaceae	−0.193
Simulated datasets			
5 taxa		8 taxa	
Taxon 1	0.515	Taxon 1	0.212
Taxon 2	0.772	Taxon 2	0.257
Taxon 3	0.209	Taxon 3	0.642
Taxon 4	0.258	Taxon 4	0.449
Taxon 5	0.164	Taxon 5	0.242
		Taxon 6	0.168
		Taxon 7	0.428
		Taxon 8	0.074



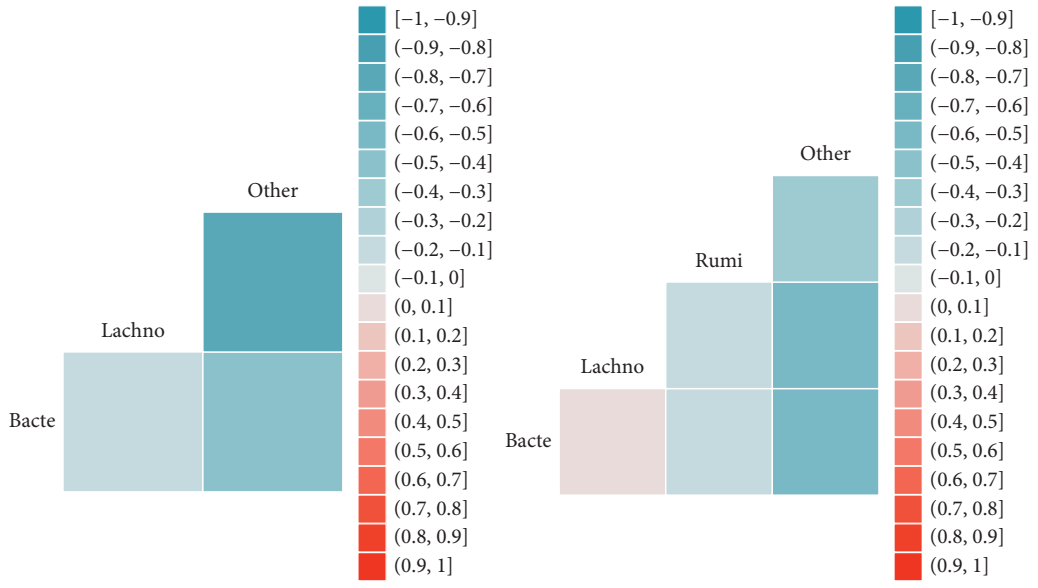
(a)

Figure 8: Continued.



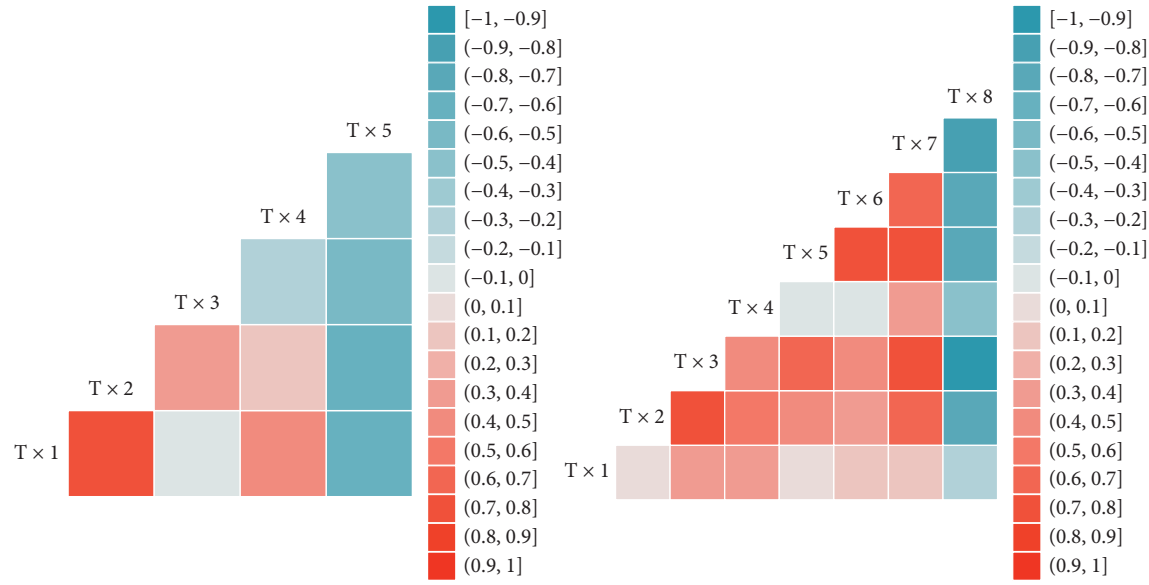
(b)

Figure 8: Correlation matrices to analyze the relationship between taxa for the subjects reported by Durbán et al. (a) and Lloyd-Price et al. (b), respectively.



(a)

Figure 9: Continued.



(b)

Figure 9: Correlation matrices to analyze the relationship between taxa for the subjects considered by Caporaso et al. (a) and simulated datasets (b), respectively.

4. Conclusions

In this paper, we develop an autoregressive model for the analysis of microbiota time series data considering a Dirichlet conditional distribution with time-varying parameters. In this approach, we assume that relative abundances, after a log-ratio transformation, can be explained by this autoregressive structure, which takes into account the effect of the bacterial community.

An empirical study has been carried out in order to show how our proposal can be useful to reveal the relationship between microbiome variability and host health status or to cluster groups of taxa sharing a similar pattern over time. Additionally, in order to increase convergence speed of the algorithm used in the optimization procedure, an accelerated strategy has been implemented. Note that a nonlinear optimization strategy requires a good initial value in order to significantly reduce the computational burden, which is a key factor in this high-dimensional scenario.

Although we have demonstrated that this novel proposal is useful to explain microbiota dynamics, several considerations should be taken into account. Microbiota longitudinal data are often temporally sparse with many zeros, and this should also be considered. Our proposal cannot be used for data without deleting or estimating the zero counts, but fortunately there are suitable methods of dealing with this type of value. One solution is to replace them with a small nonzero value. However, the zero values have multiple causes and thus there is no generalized treatment strategy. Therefore, it would be of great interest to extend the proposed model to allow the modeling of microbiota data when zero values are present without modifying any of them.

Additionally, in order make this model more widely applicable, we also propose its extension to contemplate external covariates, in order to consider extrinsic perturbations such as antibiotics or diets, seasonal terms, or multivariate random effects. Note that it can easily be adapted by adding these new regressors. Thus, this extension could be applied to vaginal microbiota data if a seasonal term was added to account for the menstrual cycle or to several individuals if random effects were to be considered.

Appendix

In order to estimate the parameters of the model, maximum likelihood estimation has been considered. In this case, the log-likelihood function is given by $L_T = \sum_{t=1}^{T} l_t$, where

$$l_t = \ln(\Gamma(\tau)) - \sum_{i=1}^{K} \ln(\Gamma(\alpha_{it})) + \sum_{i=1}^{K} (\alpha_{it} - 1)\ln(y_{it}). \quad (A.1)$$

We have maximized L_T using Nelder-Mead method with the function *optim* of R [44]. Note that this nonlinear optimization strategy requires a good initial value in order to significantly reduce the computational burden. The choice of the starting value plays a significant role in achieving rapid convergence of the algorithm. In our case, the initial points for the parameters of the model, except for τ, were chosen by solving the linear system defined by $\ln(y_{jt}/y_{Kt}) = \eta_{jt}$, where a linear least-squares algorithm was considered to solve for the unknown parameters. In this linear system optimization, L2 regularization was added to the cost function to prevent overfitting, and an optimal λ value was identified to minimize overfitting of the data. To do so, values of λ were analyzed from 0 to 1000. Computations for this optimization procedure with L2 regularization were carried out with *MASS* package of *R*, specifically with function lm.ridge. In order to assign an initial value for τ in likelihood

optimization, we carried out a second optimization procedure also considering function optim. In this case, with the parameter values obtained previously with the linear system optimization as initial values and τ in the parameter search interval [1, 30], we have calculated Akaike information Criterion (AIC) for each second optimization procedure and the τ value that minimizes Akaike information Criterion (AIC) has been selected. In summary, the initial value estimation procedure for likelihood optimization is as follows:

Step 1: solve the linear system defined by $\ln(y_{jt}/y_{Kt}) = \eta_{jt}$ taking into account a linear least-squares algorithm with L2 regularization. This step allows us to obtain the initial values for parameters except for τ. Remember that additive log-ratio data transformation and first-order Taylor approximation allow us to state that

$$E\left(\ln\left(\frac{y_{jt}}{y_{Kt}}\right)\right) = E\left(\operatorname{alr}\left(y_{jt}\right)\right) \approx \operatorname{alr}\left(E\left(y_{jt}\right)\right) = \operatorname{alr}\left(\frac{\alpha_{jt}}{\tau}\right) = \ln\left(\frac{\alpha_{jt}}{\alpha_{Kt}}\right). \tag{A.2}$$

Step 2:

(2a) Compute a second optimization procedure with the values obtained in Step 1 and τ in interval [1, 30].
(2b) Compute Akaike Information Criterion (AIC) for each optimization procedure carried out in Step (2a). This step provides the initial value for τ.

Data Availability

The data and code used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work 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 cofinanced by the European Regional Development Fund (ERDF).

References

[1] H. J. Flint, "Obesity and the gut microbiota," *Journal of Clinical Gastroenterology*, vol. 45, pp. 128–132, 2011.
[2] N. Larsen, F. K. Vogensen, F. K. Vogensen 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.
[3] J. Ahn, R. Sinha, Z. Pei et al., "Human gut microbiome and risk for colorectal cancer," *JNCI: Journal of the National Cancer Institute*, vol. 105, no. 24, pp. 1907–1911, 2013.
[4] Y. Xia, J. Sun, and D. G. Chen, *Statistical Analysis of Microbiome Data with R*, Springer, Berlin, Germany, 2018.
[5] 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 21887, 2017.
[6] C. Leong, J. J. Haszard, J. J. Haszard et al., "Compositional principal component analysis generates gut microbiota profiles that associate with children's diet and body composition," *Proceedings of the Nutrition Society*, vol. 79, no. OCE2, Article ID E284, 2020.
[7] 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.
[8] H. Li, "Microbiome, metagenomics, and high-dimensional compositional data analysis," *Annual Review of Statistics and Its Application*, vol. 2, no. 1, pp. 73–94, 2015.
[9] M. C. B. Tsilimigras and A. A. Fodor, "Compositional data analysis of the microbiome: fundamentals, tools, and challenges," *Annals of Epidemiology*, vol. 26, no. 5, pp. 330–335, 2016.
[10] G. B. Gloor, J. R. Wu, V. Pawlowsky-Glahn, and J. J. Egozcue, "It's all relative: analyzing microbiome data as compositions," *Annals of Epidemiology*, vol. 26, no. 5, pp. 322–329, 2016.
[11] 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.
[12] P. Kynclová and P. Filzmoser, "Modeling compositional time series with vector autoregressive models," *Journal of Forecasting*, vol. 34, pp. 303–314, 2015.
[13] G. K. Grunwald, A. E. Raftery, and P. Guttorp, "Time series of continuous proportions," *Journal of the Royal Statistical Society, B*, vol. 55, pp. 103–116, 1993.
[14] T. Zheng, H. Xiao, and R. Chen, "Generalized ARMA models with martingale difference errors," *Journal of Econometrics*, vol. 189, no. 2, pp. 492–506, 2015.
[15] T. Zheng and R. Chen, "Dirichlet ARMA models for compositional time series," *Journal of Multivariate Analysis*, vol. 158, pp. 31–46, 2017.
[16] 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.
[17] 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.
[18] 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. 228, 2019.

[19] 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.

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

[21] 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.

[22] G. K. Gerber, "The dynamic microbiome," *FEBS Letters*, vol. 588, no. 22, pp. 4131–4139, 2014.

[23] 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.

[24] 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, p. 202, 2018.

[25] A. J. Tyler, P. P. Amey, and I Pe'er, "Efficient and accurate inference of mixed microbial population trajectories from longitudinal count data," *Cell Systems*, vol. 10, no. 6, pp. 463–469, 2020.

[26] L. Shenvav, 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.

[27] 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, p. 963, 2019.

[28] J. Lugo-Martinez, D. Ruiz-Perez, G. Narasimhan, and Z. Bar-Joseph, "Dynamic interaction network inference from longitudinal microbiome data," *Microbiome*, vol. 7, no. 1, p. 54, 2019.

[29] J. Aitchison, *The Statistical Analysis of Compositional Data*, Chapman & Hall, London, UK, 1986.

[30] J. Brehm and S. Gates, "A comparison of methods for compositional data analysis," in *Proceedings of the Presented at the 1998 Political Methodology*, Society Annual Meeting, San Diego, USA, July 1998.

[31] R. H. Hijazi and R. W. Jernigan, "Modeling compositional data using dirichlet regression models," *Journal of Applied Probability and Statistics*, vol. 4, no. 1, pp. 77–91, 2009.

[32] V. Pawlowsky-GlahnJ. J. Egozcue et al., *Modeling and Analysis of Compositional Data*, Wiley, Hoboken, NJ, USA, 2015.

[33] 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.

[34] E. Bogart, R. Creswell, R. Creswell, and G. K. Gerber, "MITRE: inferring features from microbiota time-series data linked to host status," *Genome Biology*, vol. 20, no. 1, p. 186, 2019.

[35] T. E. Gibson and G. K. Gerber, "Robust and scalable models of microbiome dynamics," 2018, https://arxiv.org/abs/1805.04591.

[36] 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.

[37] 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.

[38] J. Lloyd-Price, C. Arze, C. Arze et al., "Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases," *Nature*, vol. 569, no. 7758, pp. 655–662, 2019.

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

[40] K. Faust, F. Bauchinger, B Laroche et al., "Signatures of ecological processes in microbial community time series," *Microbiome*, vol. 6, p. 120, 2018.

[41] K. Klemm and V. M. Eguíluz, "Growing scale-free networks with small-world behavior," *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, vol. 65, Article ID 057102, 2002.

[42] C. Wang and J. Hu, "Microbial trend analysis for common dynamic trend, group comparison and classification in longitudinal microbiome study," 2020, https://www.biorxiv.org/content/10.1101/2020.01.30.926824v1.

[43] L. Lahti and S. Shetty, "Tools for microbiome analysis in R," 2017, http://microbiome.github.com/microbiome.

[44] R. Core Team, "A language and environment for statistical computing. R foundation for statistical computing," 2013, http://www.R-project.org.