Metodi e analisi statistiche 2021 ISBN: 978-88-6629-066-7
DIPARTIMENTO DI ECONOMIA E FINANZA p. 7-20

A Brief Review on Compositional Inference of Microbiome Data

Massimo Bilancia[1]*, Fabio Manca[2], Gianvito Pio[3]

[1]Ionian Department (DSJGE) – University of Bari Aldo Moro, Via Duomo 259, 74123 Taranto IT, [2]Department of Education Science, Psychology, Communication Science (FOR.PSI.COM) – University of Bari Aldo Moro, Palazzo Chiaia/Napolitano, Via Scipione Crisanzio 42, 70122 Bari IT, [3]Department of Computer Science – University of Bari Aldo Moro, Via Orabona 4, 70125 Bari IT.

Summary: The microbiota can be defined as the collection of all microbes, bacteria, viruses, and fungi living in a host. Recent evidence suggests that the bacterial microbiota of the human gut plays an important role in the pathogenesis of many diseases by promoting inflammation. Historically, the first structured model for data analysis of the human gut microbiota was proposed in (Holmes et al., 2012). The central idea of this model is that the microbiota of a given individual represents a sample of a bacterial metacommunity, also called enterotype. An enterotype therefore represents a particular composition of the microbiota in terms of the relative abundances of different bacterial species. A similar specification was recently proposed in the unsupervised setting in (Anderlucci & Viroli, 2020). In both cases, the focus is on classification, and it was proposed to marginalize with respect to probability distributions over enterotypes to obtain a class-conditional likelihood that depends only on the specific parameters of each component of the mixture and the latent indicator variable of the metacommunity (enterotype). We review these two models available in the literature to highlight their similarities and differences, also discussing possible future research.

Keywords: Human Microbiome Data; Bayesian Statistics; Hierarchical Modeling; Multinomial-Dirichlet Finite-Mixture Model.

1. Introduction

The microbiome can be defined as the collection of all microbes, bacteria, viruses, and fungi living in a host. Recent evidence suggests that the bacterial microbiota of

* Corresponding author: massimo.bilancia@uniba.it
The three authors reviewed and revised the manuscript, contributing equally to the draft, and approved the final version by agreeing to submit it for publication.

the human gut plays an important role in the pathogenesis of many diseases by promoting inflammation (Cammarota et al., 2020; Duvallet et al., 2017; Lu et al., 2019; Valdes et al., 2018). Furthermore, gut bacterial communities change after transplantation and immunosuppressive therapy, and the prevalence of certain pathogens in these communities may be indicative of the establishment of an alloimmune response and graft rejection (Annavajhala et al., 2019).

Gut microbial communities are profiled by sequencing specific regions, such as 16S ribosomal RNA (or 16S rRNA). This biological macromolecule is part of the prokaryotic ribosome, and the genes that encode it evolve very slowly. Therefore, this sequence is largely similar across bacteria, but it also contains hypervariable subdomains that are specific to each bacterial species. These sequences are grouped into operational taxonomic units (OTUs), i.e. clusters of similar sequences, to represent the abundance of a particular bacterial species in the microbiota while avoiding the excessive sparseness that would result from grouping only identical sequences (Wang et al., 2007).

Historically, the first structured model for data analysis of the human gut microbiota was proposed in (Holmes et al., 2012). The central idea of this model is that the microbiota of a given individual represents a sample of a bacterial metacommunity, also called enterotype. An enterotype therefore represents a particular composition of the microbiota in terms of the relative abundances of different bacterial species and can be associated with particular physiological, pathological or experimental conditions. Since we do not know which enterotype each individual belongs to, it is natural to treat the data generating process as a finite mixture of probability distributions, each of which is a probability distribution over bacterial species and determines their frequency of occurrence in samples belonging to the same enterotype. This model has interesting performance when used for classification, both in supervised and unsupervised settings.

A similar specification was recently proposed in the unsupervised setting in (Anderlucci & Viroli, 2020), which considered marginalization with respect to probability distributions over enterotypes to obtain a class-conditional likelihood that depends only on the specific parameters of each component of the mixture and the latent indicator variable of the metacommunity (enterotype). In this way, the parameters can be easily estimated using the EM algorithm fitted to the Dirichlet-Multinomial distribution.

The aim of this paper is therefore to review in more detail the two models available in the literature and to highlight their similarities and differences. The paper is organized as follows. In Section 2, we describe the Unigram model as a basic

building block for human microbiota data analysis. In Section 3, we present a simple hierarchical extension of the Unigram model, where we have a set of k Multinomial parameters, each representing a particular enterotype. Section 4 is specifically devoted to the Holmes-Harris-Quince (HHQ) model and the description of its hierarchical structure, while Section 5 focuses on mixtures of Dirichlet-Multinomial distributions and, in particular, on the above-mentioned model of Anderlucci and Viroli, of which we present a simplified version (AV2 model) that has a simpler biological interpretation. Section 6 is devoted to the differences between the two models, comparing their relative hierarchical structures. Section 7 reports a simple case study with real data. Finally, Section 8 discusses possible future research approaches.

2. Unigram models

The basic data structure is an OTU table of size $n \times p$, where the generic element $y_{i\ell}$ denotes the number of copies of the ℓ-th OTU ($\ell = 1,2,\dots,p$) detected in the i-th sample ($i = 1,2,\dots,n$). The natural model for the likelihood of OTU abundances detected in each experimental sample is the p-dimensional Multinomial distribution:

$$p(y_i|\beta) = \frac{\left(\sum_{l=1}^{p} y_{i\ell}\right)!}{\prod_{\ell=1}^{p} y_{i\ell}!} \prod_{\ell=1}^{p} \beta_\ell^{y_{i\ell}}, \tag{1}$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_p) \in \mathbb{R}^p$ is the vector of multinomial parameters that must satisfy the constraints $\beta_j > 0$ for $j = 1,2,\dots,p$ and $\sum_{\ell=1}^{p} \beta_{j\ell} = 1$. The multinomial coefficient is usually abbreviated as:

$$\frac{\left(\sum_{l=1}^{p} y_{i\ell}\right)!}{\prod_{\ell=1}^{p} y_{i\ell}!} \equiv \binom{y_{i+}}{y_i}, \tag{2}$$

with $y_{i+} = y_{i1} + y_{i2} + \cdots + y_{ip}$. The parameter y_{i+} is a nuisance parameter that only enters the normalization constant and is therefore irrelevant for the inference from the posterior distribution of the parameters.

From the perspective of textual data analysis, the Multinomial model can be interpreted as a Unigram model (Nigam et al., 2000). Suppose we have a vocabulary V with $p = |V|$ terms, where in this interpretation an OTU corresponds to a vocabulary term, while the index i denotes the i-th document of a corpus of N documents. A document can be viewed as a stream of S terms from this vocabulary. The ℓ-th term in V is represented by a p-dimensional vector w of the unit basis, such that the

only non-zero component of w is the ℓ-th component. In the Unigram model, the conditional likelihood has the expression:

$$w_{is}|\beta \overset{\text{ind.}}{\sim} \text{Multinouilli}_p(\beta)\,, \quad s = 1,2,\dots,S. \tag{3}$$

Since the generic w-vector has expression $w_{is} = (0,\dots,0,1,0,\dots,0)$, with $w_{i+} = 1$, the Multinouilli distribution in (3) indicates a Multinomial distribution on a single trial. It can also be viewed as a generalization of the Bernoulli distribution for the case of $S > 2$ groups. The specification (3) is infinitely exchangeable (Gelman et al., 2013), and the data-generating mechanism can be viewed as a random mechanism that generates streams of terms such that, regardless of the length of the sequence, two sequences that differ only in the order of occurrence of the terms have the same probability of being generated. This model captures, at a low level, the outcome of the process of classifying 16S rRNA sequences into a bacterial taxonomy and encodes the fact that the only relevant information is whether a particular bacterial taxon is present in the gut bacterial community of an individual. The frequency counts of each OTU can be expressed as:

$$y_i = \sum_{s=1}^{S} w_{is}, \tag{4}$$

since by summing the unit vectors w_{is} over all possible terms in the document, we obtain the counts of each term in the vocabulary V.

The vector $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})$ can obviously correspond to several different configurations of the indicator vectors w_{is}. Given the independence and expression of the Multinouilli pmf in (3), the probability of observing y_i for a single configuration is:

$$\underbrace{\beta_1 \times \beta_1 \times \cdots \times \beta_1}_{y_{i1}\text{ times}} \times \underbrace{\beta_2 \times \beta_2 \times \cdots \times \beta_2}_{y_{i2}\text{ times}} \times \cdots \times \underbrace{\beta_p \times \beta_p \times \cdots \times \beta_p}_{y_{ip}\text{ times}} = \beta_1^{y_{i1}} \beta_2^{y_{i2}} \cdots \beta_p^{y_{ip}}, \tag{5}$$

and summing this probability over all possible configurations gives the Multinomial likelihood (1), showing complete equivalence between the two representations.

3. Hierarchical mixtures of unigrams

We can construct a finite mixture of Unigrams, that is a mixture of Multinomial distributions, to have greater flexibility. In this case, a microbial host community can

belong to one of k different enterotypes. An enterotype is a probability distribution over bacterial taxa, and different enterotypes therefore have different relative compositions, meaning that bacterial taxa that are highly expressed in one enterotype may be of little importance in another enterotype. In a finite mixture of Unigrams we have k probability distributions over the vocabulary of terms, each of which represents a different physiological or experimental condition (Cheng et al., 2014). In this case, the Multinomial parameter vector is replaced by a matrix $\beta \in \mathbb{R}^{k\times p}$, i.e.:

$$\beta = \{\beta_{j\ell}\}, \quad j = 1,2,\dots,k, \quad \ell = 1,2,\dots,p. \tag{6}$$

In other words, each row of the β matrix is a probability distribution representing a Multinomial probability vector (an enterotype). Since we do not know which enterotype each sample belongs to, we can use a finite mixture of Unigrams written in hierarchical form as follows (Gelman et al., 2013):

$$\begin{aligned} y_i|\beta, z_i &\overset{\text{ind.}}{\sim} \text{Multinomial}_p(\beta_j), \quad i = 1,2,\dots,n, \\ z_i|\lambda &\overset{\text{ind.}}{\sim} \text{Multinouilli}_k(\lambda), \quad i = 1,2,\dots,n, \end{aligned} \tag{7}$$

where $\beta_j \in \mathbb{R}^p$ denotes the row of the matrix β corresponding to the index of the single element in the latent indicator vector $z_i \in \mathbb{R}^k$ such that $z_{ij} = 1$, while $\lambda \in \mathbb{R}^k$ denotes the mixture weights. Thus, the probability distribution over the j-th enterotype ($j = 1,2,\dots,k$) is described by a Multinomial distribution with enterotype-specific Multinomial probabilities selected by the latent indicator vector z_i specific to the i-th microbial community present in the i-th individual likelihood (1),.

4. The Holmes-Harris-Quince (HHQ) model

In a mixture of Unigrams, the Multinomial parameters β_j are constant for all samples generated from the same enterotype. However, this hypothesis may not be valid due to variability in the underlying biological material. To account for this additional source of variability, we can treat the multinomial parameters as random variables by specifying a prior distribution for each row of the β matrix (Subedi et al., 2020; Xia et al., 2013). The first formulation of a hierarchical mixture model for the analysis of the human microbiota goes back to (Holmes et al., 2012), who proposed a model with the following hierarchical structure:

$$\begin{aligned} y_i|\beta_i &\overset{\text{ind.}}{\sim} \text{Multinomial}_p(\beta_i), \quad i = 1,2,\dots,n, \\ z_i|\lambda &\overset{\text{ind.}}{\sim} \text{Multinouilli}_k(\lambda), \quad i = 1,2,\dots,n, \\ \beta_i|z_i,\theta &\overset{\text{ind.}}{\sim} \text{Dirichlet}_p(\theta_j), \quad i = 1,2,\dots,n, \end{aligned} \tag{8}$$

with $\theta_j \in \mathbb{R}^p$, for $j = 1,2,\dots,k$, and $\theta_{j\ell} > 0$ for $\ell = 1,2,\dots,p$. In a full Bayesian context, this structure can be completed, e.g., by setting independent Gamma priors over the mixture parameters θ_j and a Uniform distribution over each component of the mixing weights λ. However, this hierarchical structure is quite peculiar, since we have a set of Multinomial parameters specific to each sample, and a mixture-like prior distribution over such parameters that depends on enterotype-specific hyperparameters θ_j. To better understand this formulation, we observe that by integrating the multinomial parameters into the conditional likelihood in (8) we obtain:

$$\begin{aligned} p(y_i|z_i,\theta) &= \int p(y_i,\beta_i|z_i,\theta)d\beta_i = \int p(y_i|\beta_i)p(\beta_i|z_i,\theta)d\beta_i = \\ &= \int \binom{y_{i+}}{y_i} \prod_{\ell=1}^{p} \beta_{i\ell}^{y_{i\ell}} \frac{\Gamma\left(\sum_{\ell=1}^{p}\theta_{j\ell}\right)}{\prod_{\ell=1}^{p}\Gamma(\theta_{j\ell})} \prod_{\ell=1}^{p} \beta_{i\ell}^{\theta_{j\ell}-1} d\beta_{i\ell} = \\ &= \binom{y_{i+}}{y_i} \frac{\Gamma\left(\sum_{\ell=1}^{p}\theta_{j\ell}\right)}{\prod_{\ell=1}^{p}\Gamma(\theta_{j\ell})} \int \prod_{\ell=1}^{p} \beta_{i\ell}^{y_{i\ell}+\theta_{j\ell}-1} d\beta_{i\ell} = \\ &= \binom{y_{i+}}{y_i} \frac{\Gamma\left(\sum_{\ell=1}^{p}\theta_{j\ell}\right)}{\prod_{\ell=1}^{p}\Gamma(\theta_{j\ell})} c^{-1} \underbrace{\int c \prod_{\ell=1}^{p} \beta_{i\ell}^{y_{i\ell}+\theta_{j\ell}-1} d\beta_{i\ell}}_{=1}, \end{aligned} \tag{9}$$

where the inverse of the normalization constant c in (9) has expression:

$$c^{-1} = \frac{\prod_{\ell=1}^{p}\Gamma(y_{i\ell}+\theta_{j\ell})}{\Gamma\left(\sum_{\ell=1}^{p}(y_{i\ell}+\theta_{j\ell})\right)}. \tag{10}$$

Using standard notation for the multivariate Beta function:

$$B(x) = B(x_1,x_2,\dots,x_p) = \frac{\prod_{\ell=1}^{p}\Gamma(x_\ell)}{\Gamma\left(\sum_{\ell=1}^{p}x_\ell\right)}, \tag{11}$$

the conditional likelihood (9) can be rewritten as:

$$p(y_i|z_i,\theta) = \binom{y_{i+}}{y_i} \frac{B\left(y_i+\theta_j\right)}{B\left(\theta_j\right)}. \tag{12}$$

The pmf (12) defines the Dirichlet-Multinomial distribution. It was studied, among others, by Mosimann (1962), who showed that the variance of each marginal component of the class-conditional distribution is:

$$\mathrm{Var}(y_{i\ell}|z_i,\theta) = y_{i+}\mathrm{E}(\beta_{j\ell})\mathrm{E}(1-\beta_{j\ell})\left(\frac{y_{i+}+\theta_{j+}}{1+\theta_{j+}}\right), \tag{13}$$

with $\theta_{j+} = \sum_p \theta_{j\ell}$. Thus, as a consequence of explicitly accounting for variability in the multinomial probabilities, the variance of the class-conditional likelihood exhibits overdispersion with respect to the variance of the class-conditional likelihood in (7). The extent of this overdispersion, which depends on the heterogeneity of the underlying biological material, is controlled by the term θ_{j+}, with higher values of θ_{j+} corresponding to lower overdispersion.

5. Mixtures of Dirichlet-Multinomial distributions

The generative structure of the HHQ model introduces a finite mixture prior, since the Multinomial parameters of each observation are generated by a finite mixture of Dirichlet distributions. A possible alternative specification is the following:

$$\begin{aligned} y_i|\beta, z_i &\overset{\text{ind.}}{\sim} \mathrm{Multinomial}_p(\beta_j), \quad i = 1,2,\dots,n, \\ z_i|\lambda &\overset{\text{ind.}}{\sim} \mathrm{Multinouilli}_k(\lambda), \quad i = 1,2,\dots,n, \\ \beta_j|\theta &\overset{\text{ind.}}{\sim} \mathrm{Dirichlet}_p(\theta_j), \quad j = 1,2,\dots,k. \end{aligned} \tag{14}$$

In this model we have k distinct probability distributions β_j over the enterotypes that are explicitly part of the model. As in the case of the HHQ model:

$$
\begin{aligned}
p(y_i|z_i,\theta) &= \int (y_i,\beta|z_i,\theta)d\beta = \int p(y_i|\beta,z_i)p(\beta|\theta)d\beta = \\
&= \int p(y_i|\beta_j)\prod_{s=1}^{K} p(\beta_s|\theta_s)\,d\beta_s = \\
&= \int p(y_i|\beta_j)p(\beta_j|\theta_j)\,d\beta_j \underbrace{\int \prod_{-j} p(\beta_{-j}|\theta_{-j})\,d\beta_{-j}}_{=1} = \\
&= \int \binom{y_{i+}}{y_i} \prod_{\ell=1}^{p} \beta_{j\ell}^{y_{i\ell}} \frac{\Gamma(\sum_{\ell=1}^{p}\theta_{j\ell})}{\prod_{\ell=1}^{p}\Gamma(\theta_{j\ell})} \prod_{\ell=1}^{p} \beta_{j\ell}^{\theta_{j\ell}-1}\, d\beta_{j\ell} = \\
&= \binom{y_{i+}}{y_i} \frac{\Gamma(\sum_{\ell=1}^{p}\theta_{j\ell})}{\prod_{\ell=1}^{p}\Gamma(\theta_{j\ell})} \int \prod_{\ell=1}^{p} \beta_{j\ell}^{y_{i\ell}+\theta_{j\ell}-1}\, d\beta_{j\ell},
\end{aligned}
\tag{15}
$$

and then continuing the proof in the same way that applies to the HHQ model, the class-conditional result to be a Dirichlet-Multinomial:

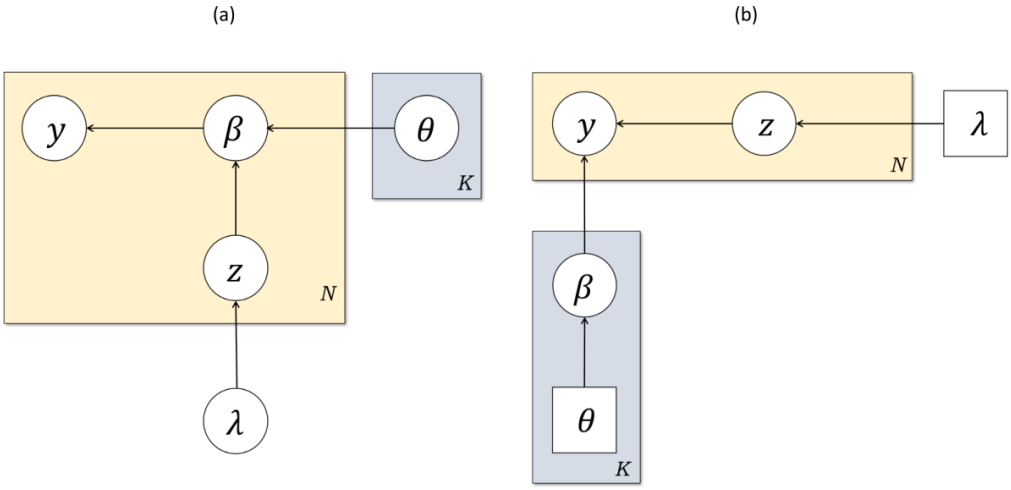
The hierarchical model defined by (14) is a suitably refactored version of the model presented in (Anderlucci & Viroli, 2020). In the originally proposed specification of each β_j prior in (14) also depended on the latent indicator variable z_i. However, this formulation was redundant since the exchangeable prior distribution in (14) induces a class-conditional likelihood equivalent to that of the model originally proposed in Anderlucci & Viroli (2020). In what follows, we refer to the model we presented in this section as the AV2 model to highlight the subtle differences that exist in our hierarchical specification.

6. A comparison of the AV2 and HHQ models

The representation in terms of oriented graphs of the two hierarchical models HQQ and AV2 is shown in Figure 1. The specification of *k* probability distributions over the multinomial parameters of the AV2 model is equivalent to a subset of the generative structure of the Latent Dirichlet Allocation (LDA) model introduced in Blei et al. (2003), in which we have *k* probability distributions over the multinomial probabilities of occurrence of terms from a vocabulary V. However, the LDA model is a mixed membership model (Airoldi et al., 2014), where each OTU can be associated with one of the k probability distributions β_j. Therefore, the starting point of the LDA model is the Multinouilli conditional likelihood (3), and the bacterial community of each individual can be viewed as a mixture of enterotypes with assignment

probabilities determined by the prior distribution over the latent variables z_i. The view based on the LDA model is discussed in detail in Sankaran & Holmes (2019) and will not be pursued here. In contrast, the AV2 model is based on a biologically simpler assumption, namely that each community can belong to a unique enterotype because the configuration of the latent indicator variable z_i selects a row of the β-matrix for the entire sample.

Figure 1. *The Dirichlet-Multinomial mixture model in graphical form. (a) HHQ model (b) AV2 model. The Bayesian specification of the HHQ model can be completed, for example, by setting independent Gamma priors over the mixture parameters θ_j and a Uniform distribution over each component of the mixing weights λ. In the AV2 model, the vector parameters θ_j and λ are assumed to be fixed but unknown.*



Enterotype distributions do not appear explicitly in the model specification, but can be approximated using estimates of the Dirichlet parameters θ_js. To be more precise, define (for $j = 1,2,\dots,k$):

$$\begin{aligned} \psi_j &= \sum_{\ell=1}^{p} \theta_{j\ell} \in \mathbb{R}, \\ \alpha_j &= \psi_j m_j \in \mathbb{R}^p, \end{aligned} \tag{16}$$

with:

$$m_j = (m_{j1}, m_{j2}, \dots, m_{jp}), \qquad \sum_{\ell=1}^{p} m_{j\ell} = 1, \tag{17}$$

that is each m_j is a probability measure (normalized to 1). That every m_j is actually a probability measure is easy to prove, because:

$$\mathrm{E}(\beta_{i\ell}|z_i,\alpha) = \frac{\alpha_{j\ell}}{\sum_{\ell=1}^{p}\alpha_{j\ell}} = \frac{\psi_j m_{j\ell}}{\sum_{\ell=1}^{p}\psi_j m_{j\ell}} = \frac{\psi_j m_{j\ell}}{\psi_j \sum_{\ell=1}^{p} m_{j\ell}} = m_{j\ell}, \qquad (18)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ and the j-index is determined by the value assumed by the latent indicator z_i. Thus, the elements of the m_j vector can be interpreted as the expected values of the Multinomial probabilities of samples belonging to the j-th enterotype, where ψ_j acts like a precision with a large ψ_j giving a small variance around these expected values.

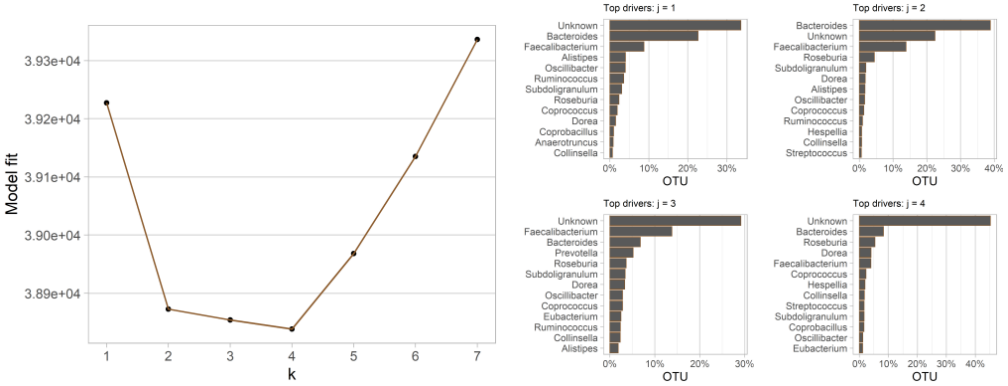
The AV2 model explicitly introduces probability distributions β_j for the enterotypes. However, in Anderlucci & Viroli (2020) a EM-like algorithm for parameter estimation is proposed based on the class-conditional Dirichlet-Multinomial likelihood, where the parameters β_j have been integrated out. A possible alternative approach is not to integrate over the enterotype distributions. In this way, we have more flexibility in modeling because we can structure the distributions over the enterotypes in a more realistic way that better fits the underlying biological processes without the limitation of analytical integrability. The obvious drawback of this approach is that the analytical simplicity of the pure Dirichlet-Multinomial model is lost, and the marginal likelihood is no longer available in closed form.

7. The Twin study

In this section, we present a brief illustrative example of the capabilities of the methods we presented using a dataset originally presented in Holmes et al. (2012). The original data were sampled from 154 different individuals characterized by family and body mass index. Each individual was sampled at two time points approximately two months apart. Of the 308 possible samples, some contained no reads after preprocessing, leaving a total of 278 samples available.

Using the HHQ model, the mixture of Dirichlet prior can be used to cluster samples at the metacommunity level. Assuming that each sample represents a unique community, we can infer which metacommunity that sample is most likely to have originated from. The number of mixture components k is estimated by minimizing the negative log-marginal distribution, approximated using the Laplace quadrature formula (Figure 2, left panel). In this way, we compute the model fit taking into account the model complexity, and the results show a minimum for $k = 4$, suggesting that a mixture of four Dirichlet distributions is more appropriate for these data.

Figure 2. *Left: negative log-marginal distribution for varying k, estimated using the Laplace quadrature formula. Right: Estimated enterotype distributions.*



Enterotype estimates were obtained by expression (18) and are shown in Figure 2 (right panel). They can be used to characterize the bacterial composition of the intestinal flora in the population we studied and possibly compare it with the composition of other populations (e.g., in individuals affected by a particular disease) or to assign each sample to the enterotype for which the respective posterior probability is highest.

8. Future developments

If we do not want to marginalize with respect to the β_j-distributions, it is natural to consider a fully Bayesian specification of the AV2 model consisting of the levels (14), and a symmetric Dirichlet prior distribution over the mixture weights:

$$\begin{aligned} y_i|\beta, z_i &\overset{\text{ind.}}{\sim} \text{Multinomial}_p(\beta_j), \quad i = 1,2,\dots,n, \\ z_i|\lambda &\overset{\text{ind.}}{\sim} \text{Multinouilli}_k(\lambda), \quad i = 1,2,\dots,n, \\ \beta_j|\theta &\overset{\text{ind.}}{\sim} \text{Dirichlet}_p(\mathbf{1}_p\theta), \quad j = 1,2,\dots,k, \\ \lambda|\alpha &\sim \text{Dirichlet}_k(\mathbf{1}_k\alpha) \end{aligned} \tag{19}$$

where $\alpha, \theta \in \mathbb{R}$ with $\alpha, \theta > 0$. We will assume that the hyperparameters θ and α are known and fixed by the researcher based on purely empirical considerations. Exploiting the conditional independence between the marginal components of the likelihood and the prior distributions of the model parameters, the marginal likelihood of the model is:

$$p(y|\theta,\alpha) = \int \prod_{i=1}^{n} \sum_{z_i} p(y_i|\beta,z_i)p(z_i|\lambda)p(\beta|\theta)p(\lambda|\alpha)d\lambda d\beta, \qquad (20)$$

which is clearly intractable. Therefore, the posterior distribution is not available in closed form, and we must resort to appropriate numerical methods for Bayesian estimation of the model parameters. One possibility, which was examined in Bilancia et al. (2021) is to use the methods of variational inference (Blei et al., 2017). The starting point is to approximate the posterior distribution $p(\beta, z, \lambda|\, y, \theta, \alpha)$ by a variational distribution $q(\beta, z, \lambda|\nu)$, which itself depends on the variational parameters ν. Our variational objective is:

$$\nu^{\star} = \text{argmin}_{\nu} \mathsf{KL}(q(\beta, z, \lambda|\nu)\big|p(\beta, z, \lambda|y, \theta, \alpha)\big), \qquad (21)$$

where the objective function is the reverse Kullback-Leibler (KL) divergence between the posterior distribution and the variational distribution. The solution of (21) provides the best possible approximation to the intractable posterior distribution with respect to the KL-divergence, and the optimized variational distribution $q(\beta, z, \lambda|\nu^{\star})$ is used to approximate the posterior inference of the model parameters. In this way, the posterior inference is treated as an optimization problem rather than a Monte Carlo sampling problem (Ghahramani, 2015).

The search for the optimal approximating distribution can be greatly simplified if we define the Evidence Lower Bound (ELBO) as follows:

$$\mathsf{ELBO}(q) = \mathsf{E}_q[\log p(y, z, \beta, \lambda|\theta, \alpha)] - \mathsf{E}_q[\log q(\beta, z, \lambda|\nu)], \qquad (22)$$

which is a functional of the variational distribution q. By making explicit the computation of the expected values in (22), the ELBO is a function of both the variational parameters and the fixed hyperparameters θ and α. It can then be shown that (Tran et al., 2021):

$$\log\, p(y|\theta,\alpha) = \mathsf{ELBO}(q) + \mathsf{KL}(q(\beta, z, \lambda|\nu)\big|p(\beta, z, \lambda|y, \theta, \alpha)\big). \qquad (23)$$

Since the KL-term is always positive, minimizing the KL-divergence with respect to the variational parameters is equivalent to maximizing the ELBO with respect to the variational parameters, and $\log p(y|\theta,\alpha) \geq \mathsf{ELBO}(q)$, that is the ELBO is a lower bound on the marginal log-likelihood. Therefore, minimizing the free energy of the variational distribution with respect to the variational parameters is equivalent to determining the tightest possible lower bound on the marginal log-likelihood. Using an appropriate variational distribution q, Bilancia et al. (2021) provide an explicit

expression of the ELBO that is maximized for parameter estimation using a coordinate ascent algorithm.

Several other extensions to the proposed model are indeed possible, through the use of nonparametric Bayesian models that would allow flexible modelling of associations between microbial communities and covariates such as environmental factors and clinical characteristics. These extensions are currently under investigation.

References

Airoldi, E. M., Blei, D., Erosheva, E. A., & Fienberg, S. E. (Eds.). (2014). *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC. https://doi.org/10.1201/b17520

Anderlucci, L., & Viroli, C. (2020). Mixtures of Dirichlet-Multinomial distributions for supervised and unsupervised classification of short text data. *Advances in Data Analysis and Classification*, *14*(4), 759–770. https://doi.org/10.1007/s11634-020-00399-3

Annavajhala, M. K., Gomez-Simmonds, A., Macesic, N., Sullivan, S. B., Kress, A., Khan, S. D., Giddins, M. J., Stump, S., Kim, G. I., Narain, R., Verna, E. C., & Uhlemann, A.-C. (2019). Colonizing multidrug-resistant bacteria and the longitudinal evolution of the intestinal microbiome after liver transplantation. *Nature Communications*, *10*(1), 4715. https://doi.org/10.1038/s41467-019-12633-4

Bilancia, M., Di Nanni, M., Manca, F., & Pio, G. (2021). Variational Bayes estimation of hierarchical Dirichlet-Multinomial mixtures for unsupervised classification. *Submitted*.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research*, *3*, 993–1022.

Cammarota, G., Ianiro, G., Ahern, A., Carbone, C., Temko, A., Claesson, M. J., Gasbarrini, A., & Tortora, G. (2020). Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature Reviews Gastroenterology and Hepatology*, *17*(10), 635–648. https://doi.org/10.1038/s41575-020-0327-3

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941. https://doi.org/10.1109/TKDE.2014.2313872

Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A., & Alm, E. J. (2017). Meta-

analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, *8*(1), 1784. https://doi.org/10.1038/s41467-017-01973-8

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis* (Third). Chapman and Hall/CRC.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459. https://doi.org/10.1038/nature14541

Holmes, Ian, Harris, K., & Quince, C. (2012). Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *PLoS ONE*, *7*(2), e30126. https://doi.org/10.1371/journal.pone.0030126

Lu, H.-F., Ren, Z.-G., Li, A., Zhang, H., Xu, S.-Y., Jiang, J.-W., Zhou, L., Ling, Q., Wang, B.-H., Cui, G.-Y., Chen, X.-H., Zheng, S.-S., & Li, L.-J. (2019). Fecal Microbiome Data Distinguish Liver Recipients With Normal and Abnormal Liver Function From Healthy Controls. *Frontiers in Microbiology*, *10*, 1518. https://doi.org/10.3389/fmicb.2019.01518

Mosimann, J. E. (1962). On the Compound Multinomial Distribution, the Multivariate β- Distribution, and Correlations Among Proportions. *Biometrika*, *49*(1/2), 65. https://doi.org/10.2307/2333468

Nigam, K., Mccallum, A. K., Thrun, S., & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, *39*(2/3), 103–134. https://doi.org/10.1023/A:1007692713085

Sankaran, K., & Holmes, S. P. (2019). Latent variable modeling for the microbiome. *Biostatistics*, *20*(4), 599–614.

Subedi, S., Neish, D., Bak, S., & Feng, Z. (2020). Cluster analysis of microbiome data by using mixtures of Dirichlet–multinomial regression models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1163–1187. https://doi.org/10.1111/rssc.12432

Tran, M.-N., Nguyen, T.-N., & Dao, V.-H. (2021). *A practical tutorial on Variational Bayes*. http://arxiv.org/abs/2103.01327

Valdes, A. M., Walter, J., Segal, E., & Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *BMJ (Online)*, *361*, 36–44. https://doi.org/10.1136/bmj.k2179

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Xia, F., Chen, J., Fung, W. K., & Li, H. (2013). A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biometrics*, *69*(4), 1053–1063. https://doi.org/10.1111/biom.12079