A Common Atom Model for the Bayesian Nonparametric Analysis of Nested Data

Francesco Denti*
Department of Statistics
University of California, Irvine, CA
and
Federico Camerlenghi†
Department of Economics, Management and Statistics
University of Milano - Bicocca, Milan, Italy
and
Michele Guindani
Department of Statistics
University of California, Irvine, CA
and
Antonietta Mira
Università della Svizzera italiana, Lugano, Switzerland
Università dell'Insubria, Varese, Italy

August 18, 2020

*During the development of this article, F. Denti was also supported as a Ph.D. student by University of Milano - Bicocca, Milan, Italy and Università della Svizzera italiana, Lugano, Switzerland.
†Also affiliated to Collegio Carlo Alberto, Piazza V. Arbarello 8, Torino and BIDSA, Bocconi University, Milano, Italy.

The use of high-dimensional data for targeted therapeutic interventions requires new ways to characterize the heterogeneity observed across subgroups of a specific population. In particular, models for partially exchangeable data are needed for inference on nested datasets, where the observations are assumed to be organized in different units and some sharing of information is required to learn distinctive features of the units. In this manuscript, we propose a nested Common Atoms Model (CAM) that is particularly suited for the analysis of nested datasets where the distributions of the units are expected to differ only over a small fraction of the observations sampled from each unit. The proposed CAM allows a two-layered clustering at the distributional and observational level and is amenable to scalable posterior inference through the use of a computationally efficient nested slice-sampler algorithm. We further discuss how to extend the proposed modeling framework to handle discrete measurements, and we conduct posterior inference on a real microbiome dataset from a diet swap study to investigate how the alterations in intestinal microbiota composition are associated with different eating habits. We further investigate the performance of our model in capturing true distributional structures in the population by means of a simulation study.

Keywords: Common Atoms Model, Microbiome Abundance Analysis, Nested Dataset, Nested Dirichlet Process, Partially Exchangeable Data

1. Introduction

The use of high-dimensional data for targeted therapeutic interventions requires new ways to characterize the heterogeneity observed across subgroups of a specific population. In particular, models for partially exchangeable data are needed for inference on nested datasets, where the observations are assumed to be organized in different, though related, units. The borrowing of strength across units induced by these probabilistic structures is tailored to several applied problems. Here we deal with a microbiome dataset made up of count measurements for 38 subjects (units) from both the U.S.A. and rural Africa, and the interest is to describe the different patterns of microbial diversity observed across the individuals since those patterns could inform future nutritional interventions. The description of microbial diversity requires investigating the structure, concentration, and richness of microbiota in each subject and how the distributions of microbiota abundances vary across subgroups of subjects. As the groups are typically unknown, they need to be estimated from the data.

A few approaches have been proposed in the literature for clustering distributional features directly. For example, Irpino and Verde (2015) have recently proposed clustering methods in symbolic statistics, by employing the Wasserstein distance on histograms treated as units. Similarly, Batagelj et al. (2015) have proposed generalized leaders and Wards hierarchical methods to cluster modal valued symbolic data. These are exploratory tools, which extend usual multivariate clustering methods to the analysis of (empirical) probability distributions, but they do not allow for a probabilistic assessment of cluster uncertainty.

The Nested Dirichlet process (nDP, Rodríguez et al., 2008) and its extensions have been widely employed to identify distributional groups in Bayesian nonparametric model-based approaches. For example, Rodriguez and Dunson (2014) have proposed a generalization of the nDP for functional data analysis; Graziani et al. (2015) have investigated how the distribution of the changes of a targeted biomarker varies due to treatment and whether it is associated with a clinical outcome; Zuanetti et al. (2018) have discussed a marginal nDP for clustering genes related to DNA mismatch repair via the

distribution of gene-gene interactions with other genes. The nDP leads to a two-layered clustering: first, it allows grouping together similar units (distributional clustering), and then, within each distributional cluster, it clusters similar observations (observational clustering). However, Camerlenghi et al. (2019a) have recently proved that the inference obtained using the nDP may be affected by a *degeneracy* property: if two distributions share even only one atom in their support, the two distributions are automatically assigned to the same cluster. To overcome this drawback, Camerlenghi et al. (2019a) propose a class of latent nested processes, which relies on estimating a latent mixture of shared and idiosyncratic processes across the subgroups. However, the computational burden of the resulting sampling scheme becomes demanding when the number of units increases.

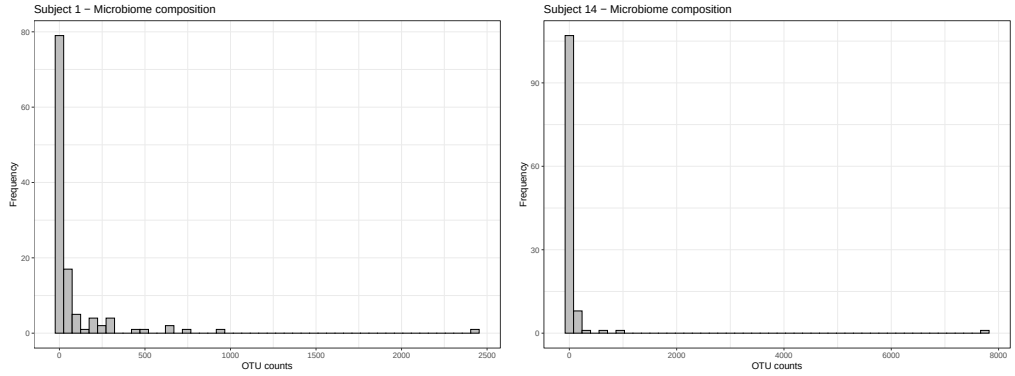
The degeneracy of the nDP is particularly problematic when analyzing high-dimensional data in genomics and microbiome studies. Here, the distribution profiles of sequencing data are expected to be quite similar across individuals and to vary only for a small fraction of differentially abundant sequences, which directly intervene to regulate the biological processes and their dysfunctions. Figure 1 reports a snapshot of the observed microbial distributions for two representative individuals from the dataset we analyze in Section 4. In addition to the typical skewness and zero-inflation of microbial distributions, we note that the two distributions considerably overlap, and they are quite similar except for the presence of a small set of sequences which appear with high frequency. In those applications, the nDP may provide unreliable inferences when comparing distributional patterns across individuals.

In this paper, we propose a nested Common Atoms Model (CAM) that is particularly suited for the analysis of nested data sets, where the distributions of the units are expected to differ only over a small fraction of the observations. Although our proposal could be described as a constrained modification of the nDP, where atoms are allowed to be shared across all subgroups, the CAM i) does not suffer from the degeneracy issue of the nDP, and ii) allows scalable inference with high-dimensional data. Furthermore, in the nDP, unit-level measurements can be clustered together only within units that are

assigned to the same group. Thus, while the within-group clustering still contributes to a compact representation of the data, unit-level inference across subgroups is precluded. Instead, the proposed CAM framework naturally allows unit-level inference and clustering of observations across groups, since the structure of the common atoms allows mapping group-specific distributional patterns to a shared support. Compared to the proposal of Camerlenghi et al. (2019a), the proposed CAM is computationally more efficient, as it allows to conduct inference on a larger number of observations and population subgroups. To this purpose, we develop a novel nested slice sampler algorithm (Kalli et al., 2011), which allows to target the true posterior distribution, without employing the standard truncation-based approximation, which is typically used for posterior inference with nDP models.

In the microbiome literature, ad-hoc solutions are sometimes adopted to address the challenges put forward by the analysis of microbiome data. For example, when dealing with the excess of zero counts, some authors simply add a small number (e.g. 1) to each count, thus generating "pseudo counts". Here, we embed the proposed CAM framework within a rounded mixture of Gaussian (RGM) model (Canale and Dunson, 2011). In this way, we effortlessly obtain a BNP nested model for count data that can naturally handle the sparsity and the zero-inflation typical of microbiome abundance tables. The resulting discrete CAM allows to cluster rows of an abundance table according to their distributional characteristics, providing a partition of patients with similar microbiome distribution. For example, the proposed CAM assigns the two subjects of Figure 1 to two different population subgroups with high probability.

The remainder of the article is as follows. In Section 2 we introduce our model for continuous measurements, and we discuss its properties. In Section 2.3 we discuss how to adapt the model to count data. In Section 3, we face posterior inference and outline the nested version of the slice sampler. Section 4 applies our model to a publicly available microbiome dataset in a diet swap study. Section 5 presents a simulation study to assess the clustering behavior of the model as the number of observations and groups grow in different scenarios. Section 6 summarizes our contributions and discusses some



(a) Subject 1 (b) Subject 14

Figure 1: Histograms of the microbiome populations of two subjects in the study of O'Keefe et al. (2015). The distributions of the two units appear very similar and extremely skewed.

future directions. We defer proofs, additional algorithms, and simulation studies to the Supplementary Material.

2. Common Atoms Model for Continuous Measurements

We consider a *nested* dataset, where we are provided with continuous measurements $\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{n_j,j})$ observed over J experimental units. We assume that each observation $y_{i,j}$, $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, takes values in a suitable Polish space $\mathbb{X}$ endowed with the respective Borel σ-field $\mathcal{X}$. Similarly as in the nDP (Rodríguez et al., 2008), our goal is to achieve a partition of the vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J$ into a few, say $K \leq J$, distributional clusters. However, Camerlenghi et al. (2019a) have shown that the partially exchangeable partition probability function of the nDP implies that distributions collapse into a common cluster when they share even only one atom. This unappealing behavior can be avoided if the prior explicitly models the commonality of atoms between groups. Here, we propose a Common Atoms Model (CAM) such that distributions belonging to different clusters are characterized by specific weights assigned to a common

set of atoms. In this section, we define the model and investigate its properties for analyzing high-dimensional data. More specifically, let G_j, as $j = 1, \ldots, J$, denote the distribution of the j-th experimental unit, so that

$$(y_{i_1,1}, \ldots, y_{i_J,J}) | \, G_1, \ldots, G_J \overset{ind.}{\sim} G_1 \times \cdots \times G_J \, , \qquad (i_1, \ldots, i_J) \in \mathbb{N}^J. \tag{1}$$

Then, similarly as in the nDP formulation, we assume that the G_j's are a sample from an almost surely discrete distribution Q over the space of probability distributions on $\mathcal{X}$, namely

$$G_1, \ldots, G_J | Q \overset{i.i.d.}{\sim} Q, \qquad\qquad Q \;\; = \sum_{k \geq 1} \pi_k \, \delta_{G_k^*}. \tag{2}$$

where $G_k^* = \sum_{l \geq 1}, \omega_{l,k} \, \delta_{\theta_l}$, $k \geq 1$, and the common atoms $\theta_1, \theta_2, \ldots$ are drawn from a non-atomic base measure H on $(\mathbb{X}, \mathcal{X})$. We further assume the Griffiths-Engen-McCloskey (GEM) distribution for the weights, which characterizes the stick-breaking (or Sethuraman's) construction of the Dirichlet process (Sethuraman, 1994), i.e. we consider $V_k \sim Beta(1, \alpha)$, $k \geq 1$, and then set $\pi_1 = V_1$, and $\pi_k = V_k \prod_{r=1}^{k-1}(1 - V_r)$, $k > 1$, indicated as $\boldsymbol{\pi} = \{\pi_k\}_{k \geq 1} \sim GEM(\alpha)$. Similarly, $\boldsymbol{\omega}_k = \{\omega_{l,k}\}_{l \geq 1} \sim GEM(\beta)$ for all $k \geq 1$.

The distribution defining G_k^* can be seen as a single-atom dependent DP as defined in Barrientos et al. (2012, Definition 3), indexed by a categorical covariate with support on $\mathbb{N}$. Hatjispyros et al. (2016) have previously investigated the use of a common atoms structure to model pairwise-dependent Dirichlet processes across m *known* subpopulations. Our CAM similarly employs common atoms to induce dependence across the G_k^*'s, but further allows clustering of distributional units, leading to a new model of nested random probability measures. Due to the commonality of the atoms at the unit level, our construction is also reminiscent of the Hierarchical Dirichlet process (HDP) by Teh et al. (2006). However, there are crucial differences between the two constructions. More specifically, the HDP does allow a flexible representation of each unit-level distribution G_j, $j = 1, \ldots, J$, but does not induce distributional clusters among the units. Our formulation preserves a two-layered clustering structure, across units (distributional clustering) and between observations (observational clustering). Thus, the proposed

CAM is closer in spirit to recently developed hierarchical topic models, where an HDP is adopted as a base measure of an (outer) DP, in symbols $Q \sim DP(\alpha, HDP(\beta, H))$ (Paisley et al., 2015; Tekumalla et al., 2015). However, those nested HDP formulations aim at describing topic distributions which can be obtained as mixtures of separate topics (i.e. a document may contain words typical of both medicine and sports news), whereas our objective is to cluster individual distributions and the observations wherein (a patient-specific distribution is not obtained as a mixture of other patients' distributions). Hence, our proposal closely mimics the intended purpose of the original nDP model. Finally, we mention an alternative semi–parametric model recently developed by Beraha et al. (2020) that also avoids the degeneracy issue of the nDP and allows for distributional clustering by extending the hierarchical Dirichlet process of Teh et al. (2006). With respect to the work by Beraha et al. (2020), our proposal is fully nonparametric, yet computationally efficient, and it easily accommodates extensions to the clustering of count data.

2.1. Partition structure and correlation

In the following, we investigate some important properties of the proposed CAM in terms of partition structure and correlation across groups. In particular, we show how the model does not suffer from the theoretical degeneracy of the nDP. We also discuss the implied dependence between pairs of observations and distributions.

The discreteness of the random probability measures in our model (1)–(2) induces ties at the observational level, whose corresponding partition can be described via the so-called partially Exchangeable Partition Probability Function (pEPPF) (see, e.g., Camerlenghi et al. (2019b) and references therein). For notational simplicity, we illustrate the main results by focusing on $J = 2$, but our strategy easily extends to the general case. We further assume that there are $s > 0$ distinct values out of a sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J$, which will be denoted by $y_1^*, \ldots, y_s^*$, with corresponding frequencies $\boldsymbol{n}_j = (n_{1,j}, \cdots, n_{s,j})$, where $n_{i,j}$ indicates the number of times that the i-th distinct value y_i^* has been observed out of the initial sample in unit j. We denote by $\mathsf{P}_{\mathbb{X}}$ the space of all random probability

measures on $\mathbb{X}$. Our first result characterizes the mixed moments of the random probability measures G_1 and G_2 as a convex combination of the fully exchangeable case and a situation of independence across samples (see also Proposition 2 in Camerlenghi et al., 2019a).

Proposition 1. *Let f_1 and f_2 be two measurable functions defined on $\mathsf{P}_{\mathbb{X}}$ and taking values in $\mathbb{R}^+$, then*

$$\mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1)f_2(g_2)Q(dg_1)Q(dg_2)\right] = q_1\mathbb{E}[f_1(G_1^*)f_2(G_1^*)] + (1-q_1)\mathbb{E}[f_1(G_1^*)f_2(G_2^*)] \quad (3)$$

where we have set $q_1 := \mathbb{P}(G_1 = G_2)$.

Following Camerlenghi et al. (2019b), we formally define the pEPPF as the probability of the observed allocation $\{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J\}$ of $s > 0$ distinct observations out of the available sample, i.e.

$$\Pi_N^{(s)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) := \mathbb{E}\int_{\mathbb{X}^s} \prod_{j=1}^{J}\prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*), \quad (4)$$

with $N = \sum_{j=1}^{J} n_j$. We point out that the i-th distinct value is shared by any two units j and κ if and only if $n_{i,j}\, n_{i,\kappa} \geq 1$. If $J = 1$ one obtains the usual exchangeable partition probability function (EPPF) for an individual sample, defined by (Pitman, 1995), and denoted here as $\Phi_{n_j}^{(s)}(\boldsymbol{n}_j)$. In the case of the Dirichlet process, this coincides with the well–known Ewens sampling formula, $\Phi_{n_j}^{(s)}(\boldsymbol{n_j}) = \frac{\alpha^s\Gamma(\alpha)}{\Gamma(\alpha+n_j)}\prod_{i=1}^{s}(n_{i,j}-1)!$ (Ewens, 1972). The pEPPF for the CAM is described by the following theorem, for the case $J = 2$.

Theorem 1. *Let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be samples from $J = 2$ experimental units under the CAM (1)–(2). Then, the induced random partition of $s > 0$ distinct observations may be expressed as*

$$\Pi_N^{(s)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = q_1\Phi_{n_1+n_2}^{(s)}(\boldsymbol{n}_1 + \boldsymbol{n_2}) + (1-q_1)\int_{\mathbb{X}^s} \mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G_j^*)^{n_{i,j}}(dy_i^*). \quad (5)$$

Although a closed form expression is not available, due to the presence of the integral over $\mathbb{X}^s$ on the right hand side, the result is fundamental to show that the proposed CAM does not reduce to the fully exchangeable case in the presence of common observations across the two samples. Indeed, we can prove the following:

Proposition 2. *Assume that two samples* $\boldsymbol{y}_1$ *and* $\boldsymbol{y}_2$ *share* $s_0 > 0$ *distinct observations, then*

$$\int_{\mathbb{X}^s} \mathbb{E} \prod_{j=1}^{2} \prod_{i=1}^{s} (G_j^*)^{n_{i,j}} (dy_i^*) > 0.$$

Theorem 1 and Proposition 2 clarify that the pEPPF (5) of our proposal does not reduce to the EPPF of the full exchangeable model. The proofs of the previous results are deferred to the Supplementary Material, where we also provide an explicit expression for the integral in (5) (see Equation (24)).

Of course, ties among distributions at the outer level are still possible in view of the discreteness of Q in (2). Indeed, if $j \neq j'$ we have

$$\mathbb{P}\left(G_j = G_{j'} | Q\right) = \sum_{k \geq 1} \pi_k^2 > 0, \quad \text{and} \quad \mathbb{P}\left(G_j = G_{j'}\right) = \frac{1}{1+\alpha}. \tag{6}$$

Moreover, the probability of a tie between two data points in two separate units j and j', with $j \neq j'$, can be computed as

$$\mathbb{P}\left[y_{i,j} = y_{i',j'}\right] = \frac{1}{1+\alpha}\left[\frac{1}{1+\beta} + \alpha \frac{1}{2\beta+1}\right]. \tag{7}$$

This shows that CAM induces a two-fold clustering structure: it clusters together experimental units characterized by similar distribution profiles, and it also clusters together observations, allowing for borrowing information across the two layers. The determination of (6)–(7) is also deferred to the Supplementary Material.

We conclude this section providing an explicit expression of the correlation between G_j and $G_{j'}$ on different Borel sets, as $j \neq j'$; the covariance and correlation are useful quantities to investigate the dependence across random probability measures and their suitability for practical applications. For any two Borel sets $A, B \in \mathcal{X}$ one has

$$\begin{aligned} &Cov\left(G_j(A), G_{j'}(B)\right) \\ &\quad = H(A \cap B)\left(\frac{q_1}{1+\beta} + \frac{1-q_1}{1+2\beta}\right) - H(A)H(B)\left(\frac{q_1}{1+\beta} + \frac{1-q_1}{1+2\beta}\right), \end{aligned} \tag{8}$$

where $q_1 = (1+\alpha)^{-1}$. In particular the correlation on the same set A equals

$$\rho_{j,j'} := Corr(G_j(A), G_{j'}(A)) = 1 - \frac{\beta}{2\beta+1}\frac{\alpha}{1+\alpha}. \tag{9}$$

See Section A of the Supplementary Material for the derivation of (8) and (9). It is interesting to note that $\rho_{j,j'} \in (1/2, 1)$, due to the commonality of the atoms. In many applications, especially in genomics, distribution profiles are expected to be quite similar across experimental units (e.g., subjects), and to vary only for a small fraction of the observations (e.g., genes). For the nDP, we have that $Corr\left(G_j(A), G_{j'}(B)\right) = (1+\alpha)^{-1} > 0$, where the expression does not depend on β: this is because the nDP assumes independence between atoms in separate distributions.

2.2. Common Atoms Mixture Model

The model defined through Equations (1)–(2) assumes a.s. discrete distributions. For modeling continuous distributions, one could follow established literature (Ferguson, 1983; Lo, 1984) and consider a nonparametric mixture model where (1) is substituted by

$$\begin{aligned} (y_{i_1,1}, \ldots, y_{i_J,J}) | \, f_1, \ldots, f_J &\overset{ind.}{\sim} f_1 \times \cdots \times f_J \qquad i_j = 1, \ldots, n_j, \; j = 1, \ldots, J \\ f_j(\cdot) &= \int_\Theta p(\cdot|\theta) \, G_j(d\theta), \quad j = 1, \ldots, J, \end{aligned} \tag{10}$$

where $p(\cdot|\theta)$ denotes an appropriate parametric continuous kernel density, and $G_j|Q \overset{i.i.d.}{\sim} Q$ as in (2). In the rest of the paper, we will adopt Gaussian kernels, i.e. we assume $p(\cdot|\theta)$ to be Normal and $\theta = \left(\mu, \sigma^2\right)$ is a vector of location and scale parameters.

To simplify the computational algorithm, we can introduce an alternative representation using two sequences of latent variables, $\boldsymbol{S} = \{S_j\}_{j \geq 1}$ and $\boldsymbol{M} = \{M_{i,j}\}_{i \geq 1, j \geq 1}$, describing – respectively – the clustering process at the distributional level and the observational level i.e. $S_j = k$ and $M_{i,j} = l$ if the observation i in unit j is assigned to the l-th observational cluster and the k-th distributional cluster. Thus we deal with the following model:

$$\begin{aligned} y_{i,j}|\boldsymbol{M}, \boldsymbol{\theta} &\sim N\left(\cdot|\theta_{M_{i,j}}\right), & M_{i,j}|\boldsymbol{S}, \boldsymbol{\omega} &\sim \sum_{l=1}^{\infty} \omega_{l,S_j} \delta_l(\cdot), \\ \boldsymbol{\omega}_k|\boldsymbol{S} = \boldsymbol{\omega}_k &\sim GEM(\alpha), & S_j|\boldsymbol{\pi} &\sim \sum_{k=1}^{\infty} \pi_k \delta_k(\cdot), \\ \boldsymbol{\pi} &\sim GEM(\beta), & \theta_l &\sim \pi(\theta_l), \; l \geq 1, \end{aligned} \tag{11}$$

where we denoted with $\boldsymbol{\theta} = \{\theta_l\}_{l\geq 1}$. In the following, we consider a Normal-Inverse Gamma distribution for $\theta_l = \left(\mu_l, \sigma_l^2\right) \sim NIG(m_0, \kappa_0, \alpha_0, \beta_0)$, i.e. $\mu_l|\sigma_l^2 \sim N\left(m_0, \sigma_l^2/\kappa_0\right)$ and $\sigma_l^2 \sim IG\left(\alpha_0, \beta_0\right)$.

2.3. Common Atoms Model for Count Data

In Section 4, we consider an application to microbiome data, which can be represented by abundance tables containing the observed frequency of a particular microbial sequence in a sample - or subject (unit). Here, we describe how the CAM can be adapted to count data, characterized by skewness and zero-inflation typically observed in microbiome studies. Let $z_{i,j} \in \mathbb{N}$ be the observed count of microbial sequence $i = 1, \ldots, n_j$ in subject $j = 1, \ldots, J$. Consequently, the vector $\boldsymbol{z}_j = \left(z_{1,j}, \ldots, z_{n_j,j}\right)$ will denote the observed microbiome abundance vector of individual j. We embed model (1)–(2) in the rounded mixture of Gaussian framework of Canale and Dunson (2011). See also Bandyopadhyay and Canale (2016) and Canale and Prünster (2017), where the rounded mixture framework is compared to less flexible nonparametric mixtures of Poisson densities for count data. In order to define a probability mass function for the discrete measurements z, Canale and Dunson (2011) consider a data augmentation framework by latent continuous variables y, such that

$$f\left(Z = j\right) = \int_{a_j}^{a_{j+1}} g\left(y\right) dy, \quad j \in \mathbb{N}$$

for a fixed sequence of thresholds $a_0 < a_1 < a_2 < \ldots < a_\infty$ and for some density function $g(\cdot)$, such that $\int_{a_0}^{a_{+\infty}} g\left(y\right) dy = 1$. Typically, the sequence of thresholds is set as $\mathbf{a} = \{a_j\}_{j=0}^{+\infty} = \{-\infty, 0, 1, 2, \ldots, +\infty\}$ and $g(\cdot)$ is a Dirichlet Process mixture density, to ensure a flexible representation of the table of counts. We propose a novel nested formulation, where $g(\cdot)$ is modeled as a CAM mixture (11). More specifically, we consider

$$z_{i,j}|y_{i,j} \sim \sum_{g=0}^{+\infty} \delta_g(\cdot)\mathbf{1}_{[a_g, a_{g+1})}\left(y_{i,j}\right), \tag{12}$$

where $y_{i,j}$ is distributed as in (11). We will refer to this new setting as the Discrete Common Atoms Model (DCAM).

3. Posterior Inference

Typically, posterior samples for the nDP process have been obtained using a truncated version of the Blocked-Gibbs Sampler (Ishwaran and James, 2001), i.e. by choosing proper upper bounds for the infinite sums that appear in (11). Specifically, the model representation in (11) is useful to obtain such an algorithm, which we detail in Section B of the Supplementary Material, where we also provide useful upper bounds to control the resulting truncation error. Here we present a novel nested version of the independent slice-efficient algorithm (Walker, 2007; Kalli et al., 2011). Compared to truncation-based algorithms, the proposed slice sampler has two main advantages: it allows to target the true posterior distribution and it considerably decreases the computational time by stochastically truncating the model at the needed number of mixture components. The proposed slice sampling scheme can be easily extended to the nDP, and is related to the sampling scheme in Banerjee et al. (2013), although their model is essentially different from ours. In the following, we focus on the Common Atoms Mixture model (11), as variations to accommodate for count data are straightforward.

Let $p\left(y_{i,j}|\theta_l\right)$ denote a generic density function for the observation $y_{i,j}$, conditionally given θ_l, let $\boldsymbol{\pi} = \{\pi_k\}_{k\geq 1}$ and $\boldsymbol{\omega} = \{\omega_{l,k}\}_{l,k\geq 1}$ be the two sets of weights, one referred to the distributional clusters, the other one referred to the observational clusters. Then, we can write:

$$f\left(y_{i,j}|\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\pi}\right) = \sum_{k\geq 1}\pi_k \sum_{l\geq 1}\omega_{l,k}\, p\left(y_{i,j}|\theta_l\right).$$

As in the classic slice sampler, we augment the model introducing two sets of latent variables controlling which components of the mixture are "active" and which can be ignored. More specifically, we introduce $\boldsymbol{u^D} = \{u_j^D\}_{j=1}^J$– where the D in the superscript indicates the distributional level – and, within every unit $j = 1,\ldots,J$, we define an inner sets of latent variables, $\boldsymbol{u_j^O} = \{u_{i,j}^O\}_{i=1}^{n_j}$, at the level of the observations. Moreover, we also consider the following deterministic sequences: $\boldsymbol{\xi}^D = \{\xi_k^D\}_{k\geq 1}$ and, for every k, $\boldsymbol{\xi_k^O} = \{\xi_{l,k}^O\}_{l\geq 1}$. Then the model can be rewritten as

$$f_{\boldsymbol{\xi}^D,\boldsymbol{\xi}^O}\left(y_{i,j},u_j^D,u_{i,j}^O|\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\pi}\right) = \sum_{k\geq 1}\mathbb{1}_{\{u_j^D<\xi_k^D\}}\frac{\pi_k}{\xi_k^D}\sum_{l\geq 1}\mathbb{1}_{\{u_{i,j}^O<\xi_{l,k}^O\}}\frac{\omega_{l,k}}{\xi_{l,k}^O}p\left(y_{i,j}|\theta_l\right). \tag{13}$$

Notice that if we assume $\xi_k^D = \pi_k$ and $\xi_{l,k}^O = \omega_{l,k}$, we recover the nested version of the efficient-dependent slice sampler. By introducing two sets of latent labels that identify the distributional ($\boldsymbol{S}$) and observational ($\boldsymbol{M}$) cluster in which the observation is allocated, we get rid of the infinite sums in the previous equations. The complete likelihood for the entire dataset becomes

$$\begin{aligned} & f_{\boldsymbol{\xi}^D,\boldsymbol{\xi}^O}\left(\boldsymbol{y},\boldsymbol{u^D},\boldsymbol{u^O},\boldsymbol{M},\boldsymbol{S}|\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\pi}\right) \\ & \qquad = \prod_{j=1}^{J} \mathbb{1}_{\{u_j^D < \xi_{S_j}^D\}} \frac{\pi_{S_j}}{\xi_{S_j}^D} \prod_{i=1}^{n_j} \mathbb{1}_{\{u_{i,j}^O < \xi_{M_{i,j},S_j}^O\}} \frac{\omega_{M_{i,j},S_j}}{\xi_{M_{i,j},S_j}^O} p\left(y_{i,j}|\theta_{M_{i,j}}\right). \end{aligned} \tag{14}$$

Let $\phi(\cdot|\theta)$ and $\Phi(\cdot|\theta)$ denote the p.d.f. and the c.d.f. of a normal random variable with location-scale parameter θ, respectively. Then, if we assume $p(y_{i,j}|\theta_{M_{i,j}}) = \phi(y_{i,j}|\theta_{M_{i,j}})$ we recover the CAM model listed in (10). Alternatively, to recover the DCAM model for discrete data $\boldsymbol{z}$ as in (12), it is sufficient to adopt the mixing kernel $p\left(z_{i,j}|\theta_{M_{i,j}}\right) = \Delta\Phi(a_{z_{i,j}};\theta_{M_{i,j}}) = \Phi\left(a_{z+1};\theta_{M_{i,j}}\right) - \Phi\left(a_z;\theta_{M_{i,j}}\right)$, obtained by integrating out the latent continuous variable. In a general framework, the nested slice sampler is obtained by looping over the full conditionals for T iterations, according to the pseudo-code reported in Algorithm 1. For the DCAM, an additional step is added to update the latent continuous variable (see Step 1 of the algorithm in the Supplementary Material). The computation of Steps 5, 6, and 7 is feasible, as we stochastically truncate the number of mixture components to a sufficiently high integer to ensure that the two steps can be carried out exactly. Additional details for this procedure are reported in the Supplementary Material.

4. Analysis of microbial distributions of African Americans and rural Africans

We apply the proposed modeling framework to the analysis of a microbiome dataset. Here, a primary goal is to study *microbial diversity*, i.e. how the distribution of microbial units varies across subgroups of a population. Typically, summary statistics are used to

Algorithm 1: Nested Slice-Efficient Sampler for the Common Atoms Model

for $i = 1, \ldots, T$ **do**

1. Sample each u_j^D from a uniform distribution $\mathcal{U}\left(0, \xi_{S_j}^D\right)$.
2. Sample each $u_{i,j}^O$ from a uniform distribution $\mathcal{U}\left(0, \xi_{M_{i,j},S_j}^O\right)$.
3. Sample the proportions $\boldsymbol{v}$ for the SB weights independently from $v_k \sim Beta\left(a_k, b_k\right)$, where $a_k = 1 + \sum_{j=1}^J \mathbb{1}_{\{S_j=k\}}$ and $b_k = \alpha + \sum_{j=1}^J \mathbb{1}_{\{S_j>k\}}$. This full conditional is obtained marginalizing $\boldsymbol{u^D}$ out.
4. For each k, sample the proportions $\boldsymbol{u_k}$ independently from $u_{l,k} \sim Beta\left(a_l^k, b_l^k\right)$, where $a_l^k = 1 + \sum_{i=1}^N \mathbb{1}_{\{M_{i,j}=l, S_j=k\}}$ and $b_l^k = \beta + \sum_{i=1}^N \mathbb{1}_{\{M_{i,j}>l,\, S_j=k\}}$. This full conditional is obtained collapsing both $\boldsymbol{u^D}$ and $\boldsymbol{u^O}$.
5. Following Banerjee et al. (2013); Porteous et al. (2006), we obtain more efficient updates trough partial collapsing, integrating over the inner level slice variables $\boldsymbol{u^O}$. Then, we sample from

$$\mathbb{P}\left(S_j = k | \cdots\right) \propto \mathbb{1}_{\{u_j^D < \xi_k^D\}} \frac{\pi_k}{\xi_k^D} \prod_{i=1}^{n_j} \omega_{M_{i,j},k}.$$

6. Sample the observational labels from the following full conditional distribution:

$$\mathbb{P}\left(M_{i,j} = l | \cdots\right) \propto \mathbb{1}_{\{u_{i,j}^O < \xi_{l,S_j}^O\}} \frac{\omega_{l,S_j}}{\xi_{l,S_j}^O} p\left(y_{i,j} | \theta_l\right).$$

7. Sample θ_l from a conjugate NIG.

end

capture characteristics of species' distributions, e.g. α-diversity and β-diversity metrics such as Shannon's entropy and Bray-Curtis dissimilarity indexes, respectively (Whittaker, 2006). However, those metrics do not fully capture the complexity of microbiome data, which poses distinctive statistical challenges (Mao et al., 2020). In particular, the data are recorded as counts of the observed microbial genome sequences. The resulting histograms are highly skewed and sparse, due to the many low- or zero- frequency counts and to the presence of a few dominant sequences (see Figure 1). Indeed, when compared across subjects, microbiota abundance data show a characteristic zero-inflation.

The taxonomical classification of microbial species is typically conducted based on sequence alignments, e.g. through the use of 16S rRNA sequences: "practically identical" sequenced tags ($\geq$ 95% of degree of similarity) are clustered together into the same *phylotype*, and referred to as an *operational taxonomic unit* (OTU). Thus, for each specimen (e.g. fecal sample) obtained from a particular ecosystem (e.g. the gut), the number of recurrences of each OTU is recorded (Jovel et al., 2016; Kaul et al., 2017). Collecting samples from distinct individuals leads to the construction of an *abundance table*, a matrix formed by the OTU counts (taxa) observed in each sample. Let $\boldsymbol{Z}$ indicate a $n \times J$ abundance table where each entry $z_{i,j} \in \mathbb{N}$ is the frequency of the i-th OTU observed in the j-th subject, $i = 1, \ldots, n$, $j = 1, \ldots, J$, where n represents the total number of OTUs. Thus, the vector $\boldsymbol{z}_j = (z_{1,j}, \ldots, z_{n,j})'$ denotes the observed microbiome sample of individual j.

To understand the varying composition of the microbiome in the population, we apply the DCAM model proposed in Section 2.3 to the dataset from the study of O'Keefe et al. (2015), publicly available in the R package `microbiome`. The dataset contains the OTU counts of both healthy middle-aged African Americans (AA) and rural Africans (AF). The participants to the experiments were asked to follow their characteristic diet – "rural" (low-fat and high-fiber) for AF and "western" (high fat and low-fiber) for AA – for two weeks and then swap their diet regimes for other two weeks. During these two weeks, fecal samples were regularly collected to investigate the role of fat and fiber in the association between a specific diet and colon cancer risk. For our application, we focus

on the abundance table obtained at the beginning of the experiment. Once we restrict our attention to the first time point, we find that 11 OTUs are absent across all the individuals. Therefore, they are removed from the dataset. However, since our model is designed to handle sparsity, we do not discard any underrepresented taxa, to avoid potential statistical power loss (McMurdie and Holmes, 2014). Our abundance table consists of 119 taxa measured for 38 patients. The heatmap of the data in log-scale, stratified by nationality, is shown in Figure 7 in the Supplementary Material.

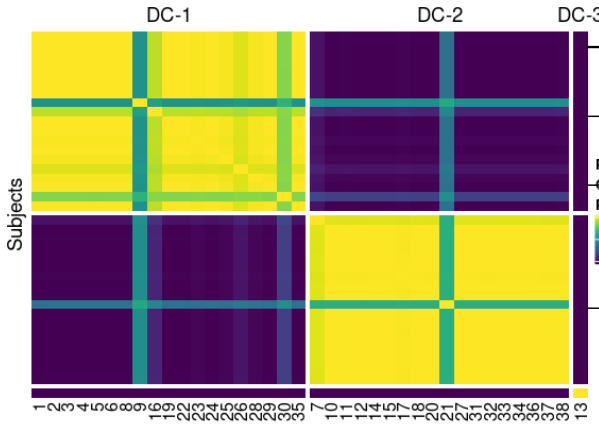
The varying sequencing depths also affect the so-called *library size*, i.e. the total frequencies of the observed species (OTUs) in each subject sample. Let $X_j = \sum_{i=1}^{n} z_{i,j}$ indicate the library size for subject j and let $\gamma_j = \bar{X}_j$ denote the corresponding average of the OTU frequencies. We incorporate the library sizes as a scaling factor in the latent level of the DCAM, i.e.,

$$y_{i,j}|\boldsymbol{M}, \boldsymbol{\mu}, \boldsymbol{\sigma^2} \sim N\left(\gamma_j \cdot \mu_{M_{i,j}}, \gamma_j^2 \cdot \sigma^2_{M_{i,j}}\right) \iff \frac{y_{i,j}}{\gamma_j}|\boldsymbol{M}, \boldsymbol{\mu}, \boldsymbol{\sigma^2} \sim N\left(\mu_{M_{i,j}}, \sigma^2_{M_{i,j}}\right). \quad (15)$$

Both the mean and the variance of the latent continuous random variable are decomposed multiplicatively into the deterministic term γ_j that describes the depth of the sequencing, and two stochastic terms that capture the intensity $\mu_{M_{i,j}}$ and the uncertainty $\sigma^2_{M_{i,j}}$ behind the OTU counts, respectively.

We adopt standard prior settings for all the hyperparameters $(m, \kappa, \alpha_0, \beta_0, a_\alpha, b_\alpha, a_\beta, b_\beta)$. Following an empirical Bayes rationale, we set m and κ to be equal to the grand mean and the inverse of the overall sample variance. According to Rodríguez et al. (2008), we then set $\beta_0 = 1$ and $\alpha_0 = a_\alpha = b_\alpha = a_\beta = b_\beta = 3$. A MCMC sample of 100,000 iterations was collected after a burn in period of the same length. Convergence of the MCMC was assessed based on visual inspection and standard convergence diagnostics (Plummer et al., 2006).

Distributional cluster analysis. To obtain an estimate for the distributional clustering, we first compute the posterior pairwise co-clustering matrix. From this matrix, we estimate the optimal partition by considering a decision-theoretic approach and minimizing the expected posterior loss under a specific loss function. We follow Wade and Ghahramani (2018), who propose to rely on the minimization of the Variation of Infor-



Cluster	DC-1	DC-2	DC-4
Cardinality	19	18	1
Africans	2	14	1
Americans	17	4	0
Female	11	6	0
Male	8	12	1

Figure 2: Left: pairwise posterior probability matrix of coclustering among the 38 subjects. A partition of the subjects' distributions into three clusters is obtained after minimization of the posterior expected Variation of Information loss function. Right: Table reporting the clusters' characteristics.

mation loss function developed by Meilúa (2007). The results are reported in Figure 2, where we also summarize the main characteristics of these distributional clusters (DCs) in terms of cardinality, nationality, and gender. It is remarkable how the different subpopulations of microbiome populations are captured by our model: in fact, Cluster DC-1 contains almost all the AA subjects, while Cluster DC-2 is composed mostly of AF. Cluster DC-3 contains only one subject, whose microbiome distribution is substantially unique. The resulting DCs capture relevant distributional characteristics and the diversity of the microbiomes. In particular, the Shannon index (Shannon, 1948) or the Simpson index are often used to measure the α-diversity of a microbiome community, i.e. the richness (number) and evenness (frequencies' similarity) of the different OTUs observed in a sample. Conditionally on the optimal configuration, we compute 9 summary statistics for each subject. The DCs capture the different levels of α-diversity of the microbiome subpopulations. Indeed, the Shannon index and the Simpson Index vary substantially across the groups. In detail, the distributional cluster DC-1 is characterized by microbiome distributions with shorter range, lower standard deviations, skewness, and kurtosis than DC-2. However, DC-2 also show less richness/diversity

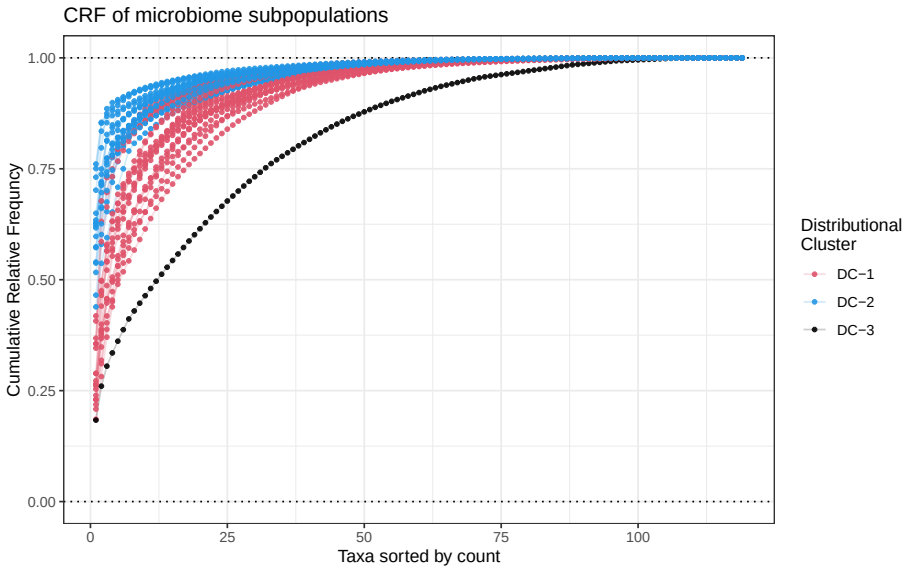


Figure 3: Cumulative Relative Frequency of the OTU abundances, sorted by decreasing order. Each color represents a DC. The lower the line, the richer and more diverse is the microbiome.

than DC-1. See Figure 11 in the Supplementary Material. Therefore, we expect that the microbiomes clustered in DC-2 are more likely to contain a small fraction of highly prominent OTUs. To confirm this intuition, let $z_{(i),j}$ represent the i-th most frequently observed OTU in subject j. We define the cumulative relative frequency (CRF) for subject j as $CRF_j(i) = \sum_{l=1}^{i} z_{(l),j} / \sum_{i=1}^{n} z_{i,j}$. Figure 3 shows the CRFs for all the subjects colored by the DCs. The CRF curves in DC-2 tend to get very close to 1 within the first 25 most abundant OTUs, showing that the relative frequencies are dominated by few, but highly expressed taxa. At the same time, the CRF curves in DC-1 increase with a slower pace, meaning more heterogeneity in the microbiome subpopulations. The CRF curve of the single subject in DC-3 increases much more slowly, indicating a peculiar microbiome, richer and more diverse than any other. We compute the median abundance of each OTU stratified by DC. In both cluster DC-2 and cluster DC-3, the leading OTU is the *Prevotella melaninogenica*. On average, it represents 60% of the observed counts in each individual in DC-2 and the 18% in DC-3. Cluster DC-1 is more diverse: the two

most expressed OTUs are the *Bacteriodes vulgatus* and the *Oscillospira guillermondii* that on average represent the 15% and the 12% of the subjects' library size, respectively. Cluster DC-3 is also characterized by a high proportion of *Faecalibacterium prausnitzii* (7%).

Observational cluster analysis. We further investigate the observational clusters (OC) induced by DCAM. Minimizing the Variation of Information we find 8 OCs, representing different intensities of the latent process underlying the counts. For a visual comparison, we report in Figure 4 the boxplots of the taxa counts grouped by OC, with the value of the median superimposed. For simplicity, we group the 8 OCs in three macro clusters representing the *abundance classes* (Low, Medium, and High). Heatmaps showing the prevalence of each OTU in every abundance class are reported in the Supplementary Material.

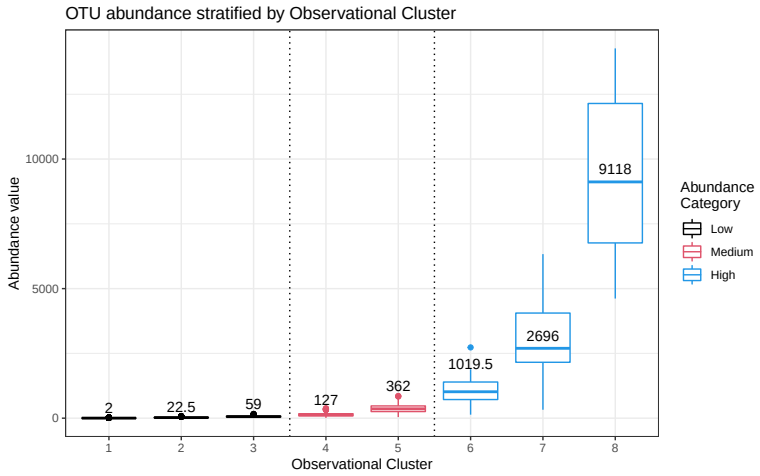


Figure 4: Boxplots of microbiome abundance counts stratified by observational clusters. We can recover three macro-clusters, with Low, Medium and High level of expression. The count median of each category is superimposed.

Finally, the distributional and observational results can be combined to discover more informative patterns, relating OTUs and subjects. Here, we investigate the co-expression structure among the most expressed OTUs in DC-1 and DC-2. To do so, we first

stratify the subjects by distributional clusters (DC-1 and DC-2) and remove the OTUs that, across all individuals, are always assigned to the Low abundance class. With the remaining 12 OTU, we compute two pairwise co-occurrence matrices (PCM_k) as $PCM_k(l,g) = \sum_{h=1}^{n_k} \mathbb{1}_{\{AC(g)=AC(l)\}}/n_k$, i.e. the percentage of times that OTU l and OTU g have been assigned to the same abundance class (AC) across the n_k individuals assigned to DC $k = 1, 2$. We plot two co-occurrence networks among the selected OTUs in Figure 5. Taxa l and g are linked if $PCM_k(l,g) = PCM_k(g,l) > 0.5$. The nodes are colored according to the modal abundance class. Again, the *Prevotella malaninogenica* and the *Prevotella oralis* are both highly expressed and co-occurent in DC-2, while in DC-1 they fall in the Low abundance class and are not linked. In DC-1, highly and co-occurent taxa are the *Bacteriodes vulgatus*. These results are in line with well-established results in the literature, since subjects with a preponderance of *Prevotella spp.* are more likely to consume fibers, while diets richer protein and fat diet - typical of western diets - lead to a predominance of *Bacteroides spp.* (Graf et al., 2015; Preda et al., 2019).

5. Simulation study

We test the performances of the proposed methodology for continuous (CAM) and discrete measurements (DCAM) within a simulation study comprised of three scenarios. For every scenario, we generate the units containing the observations from highly overlapping mixture densities. We want to assess our model's ability to recover the ground truth by recognizing the units sampled from the same mixture density (i.e. identify the distributional clusters – DC) and the observations generated from the same mixture component (i.e. identify the observational clusters – OC), for increasing number of observations in each unit, n_j, or for increasing number of units, J. We adopt the same prior specification as in the case study and estimate the best partitions by minimizing the Variation of Information given the MCMC output. We now describe the three scenarios:

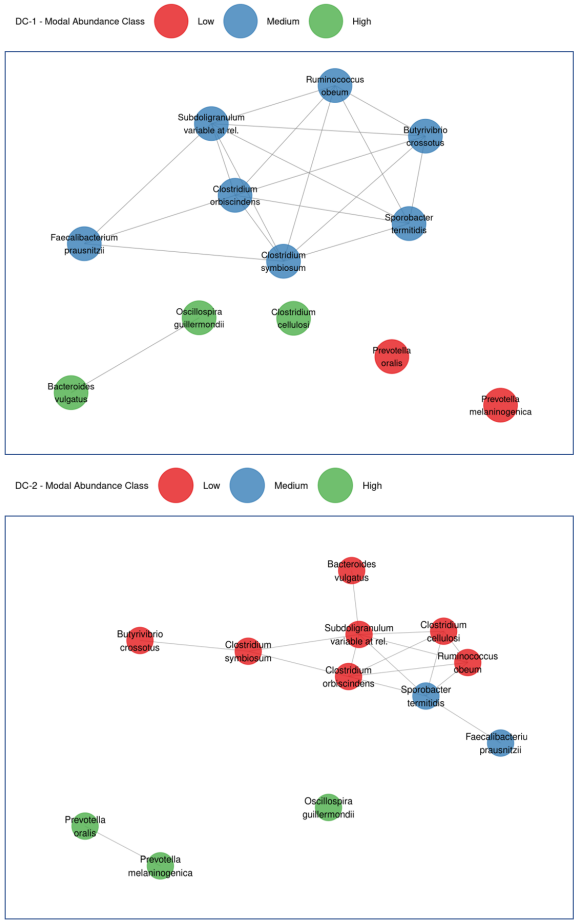


Figure 5: Co-expression networks among OTUs reporting a subset of most expressed microbes for DC-1 (left panel) and DC-2 (right panel).

Scenario 1 - CAM. We define 6 different distributions of the simulated data $\boldsymbol{Y}_h$, as

$$\boldsymbol{Y}_h \sim \sum_{g=1}^{h} \frac{1}{g} N(m_g, 0.6), \text{ where } m_g \in \{0, 5, 10, 13, 16, 20\} \text{ and } h = 1, \ldots, 6.$$

From each of these distributions, we sample two units, therefore $J = 12$. The true number of DCs and OCs is 6 in both cases. To assess how the model behaves with asymmetries in the units' sample sizes, we follow two different approaches. **Case A**: all the units have the same cardinality $n_j = n_A$, where $n_A \in \{25, 50, 75\}$. **Case B**: each unit has cardinality n_j proportional to the number of mixture components it contains. Specifically, $n_j = n_B \cdot j$ for $j = 1, \ldots, 6$ and $n_B \in \{5, 10, 20\}$.

Scenario 2 - CAM. Four highly overlapping mixtures are considered:

$$\boldsymbol{Y}_1 \sim 0.75N(0, 0.6) + 0.25N(3, 0.6), \quad \boldsymbol{Y}_2 \sim 0.25N(0, 0.6) + 0.75N(3, 0.6),$$
$$\boldsymbol{Y}_3 \sim 0.33N(0, 0.6) + 0.34N(-2, 0.6) + 0.33N(2, 0.6),$$
$$\boldsymbol{Y}_4 \sim 0.25N(0, 0.6) + 0.25N(-2, 0.6) + 0.25N(2, 0.6) + 0.25N(10, 1).$$

The true number of DCs is 4 and there are 5 OCs, corresponding to the 5 different normal distributions that constitute the mixtures. We keep the number of observation per unit constant, equal to $n_j = 40$ for any j. Instead, we vary the number of units sampled from each distribution, indicated as $r = 1, \ldots, 6$. Therefore, $J_r = 4 \cdot r$, i.e. the total number of considered units ranges from $J_1 = 4$ to $J_6 = 24$. In this way, we can investigate the estimated DC structures as the total number of units increases.

Scenario 3 - DCAM. First, let δ_x denote a point mass placed on point x and let $\mathcal{U}_d(q, Q)$ represent a Uniformly Discrete distribution over the set of integers $\{q, \ldots, Q\} \subset \mathbb{Z}$. We consider three different discrete mixtures, from which we sample $J = 10$ units:

$$\boldsymbol{Y}_g \sim \sum_{b=1}^{2} \omega_b \delta_{b-1} + \omega_3 \, \mathcal{U}_d(0, Q_g) \text{ with } g = 1, 2, 3 \text{ and } Q_g \in \{10, 50, 100\},$$

with $\omega_g = n_g / \sum_{l=1}^{3} n_l$, $g = 1, \ldots, 3$ denoting the mixture weights. We set $\omega_1 = \omega_2$ by generating the $n_1 = 50$ observations equal to zero and $n_2 = 50$ equal to one to simulate a case of low value inflation. We investigate the performance of the model in 6 cases, distinguished by the number of observations assigned to the third mixture component, i.e. $n_3 \in \{10, 15, 25, 50, 75, 100\}$. We design this simulation study to test how DCAM perform on distributions that are similar to typical microbiome samples, raising the same type of challenges. The number of true DCs is fixed equal to 3. However, there is no clear number of true OCs in this case. To assess the grouping at the level of the observations, we assume the following sets as ground truth, mimicking the segmentation in abundance levels of Section 4. We postulate 4 OCs, where the first set contains "low-expressed" observations (i.e. constituted of zeros and ones). The remaining 3 groups are obtained partitioning the support into abundance classes corresponding to the intervals $[2, 10]$,$(11, 50]$ and $(51, 100]$.

We report an illustrtion of the mixtures distributions of each scenario in the Supplementary Material.

For each scenario, we also run a nDP mixture model for the case with the highest number of observations. In Table 1 we assess the goodness of the estimated optimal partition by comparing the number of detected clusters, computing the Adjusted Rand Index (ARI - Hubert and Arabie, 1985) between the estimated optimal partition and the ground truth. Moreover, we report the normalized Frobenius distance (Horn et al., 2013) between the estimated posterior pairwise coclustering matrices and the true coclustering structures, defined as follows. Given two $p \times p$ matrices $A = \{a_{ij}\}_{i,j=1}^{p}$ and $B = \{b_{ij}\}_{i,j=1}^{p}$, we define $NFD(A, B) = \sum_{i,j=1}^{p}(a_{ij} - b_{ij})^2/p^2$. From Table 1, we can appreciate how the model can recover the ground truth, even for small sample sizes. In particular, the NFD between the distributional clustering structures approaches zero as the sample size increase. The same holds for the ARI index, that shows how the truth is recovered by the estimated best partition. We see how CAM misassigned a few observations in the wrong OCs in Scenario 2. This is due to the fact that the different mixture components are highly overlapping. Nevertheless, CAM and DCAM perform really well in Scenarios 1 and 3, respectively, where the true OC are well separated. Lastly, it is evident how the overlap of the data impacts the estimated partitions of the nDP, both at the distributional and at the observational level. In particular, when highly overlapping discrete data are considered (Scenario 3), it collapses all the units in a single DC.

6. Discussion

We have introduced a nested nonparametric model that allows investigating distributional heterogeneity in nested data. The proposed Common Atoms Model allows a two-layered clustering at the distributional and observational level, similarly to the nDP of Rodríguez et al. (2008). By construction, our model formulation allows the sharing of atoms with different weights across distributions, and it does not suffer from the degeneracy properties that occurs in the nDP, as noted by Camerlenghi et al. (2019a) whenever there is a tie between atoms. The Common Atoms Model specification is appealing and

Scenario 1	$n_A = 25$	$n_A = 50$	$n_A = 75$	$n_B = 5$	$n_B = 10$	$n_B = 20$	nDP
DC-D/T	4/6	6/6	6/6	4/6	5/6	6/6	5/6
DC-ARI	0.421	1.000	1.000	0.542	0.718	1.000	0.718
DC-NFD	0.123	0.007	0.004	0.094	0.058	0.002	0.056
OC-D/T	4/6	6/6	6/6	5/6	6/6	6/6	6/6
OC-ARI	0.925	0.988	0.973	0.964	0.970	0.964	0.353
OC-NFD	0.082	0.102	0.115	0.041	0.064	0.098	0.134
Scenario 2	$J_1 = 4$	$J_3 = 8$	$J_3 = 12$	$J_4 = 16$	$J_5 = 20$	$J_6 = 24$	nDP
DC-D/T	3/4	4/4	5/4	5/4	4/4	4/4	4/4
DC-ARI	0.000	1.000	0.891	0.918	1.000	1.000	1.000
DC-NFD	0.081	0.003	0.022	0.019	0.003	0.008	0.001
OC-D/T	5/5	5/5	5/5	4/5	5/5	5/5	2/5
OC-ARI	0.665	0.756	0.714	0.629	0.758	0.768	0.092
OC-NFD	0.113	0.124	0.143	0.152	0.129	0.143	0.149
Scenario 3	$n_3 = 10$	$n_3 = 15$	$n_3 = 20$	$n_3 = 50$	$n_3 = 75$	$n_3 = 100$	nDP
DC-D/T	7/3	2/3	3/3	5/3	4/3	3/3	1/3
DC-ARI	0.115	0.366	1.000	0.695	0.826	1.000	0.000
DC-NFD	0.259	0.251	0.035	0.076	0.057	0.000	0.640
OC-D/T	4/4	4/4	5/4	3/4	4/4	6/4	10/4
OC-ARI	0.999	0.945	0.966	0.973	0.953	0.937	0.534
OC-NFD	0.722	0.238	0.414	0.454	0.338	0.151	0.740

Table 1: Distributional (DC-) and observational (OC-) clustering performance for CAM, DCAM and nDP evaluated according the number of detected clusters over the truth (D/T), the Adjusted Rand Index (ARI) and the normalized Frobenius distance (NFD) between posterior pairwise coclustering matrices.

convenient for a variety of reasons: it is simple, allows a more refined description of distributional clusters, and it is computationally efficient thanks to the implementation of a nested version of the independent slice-efficient sampler. We have extended the methodology to take into account the modeling and clustering of discrete distributions, by considering a rounded mixture of Gaussian kernels as in Canale and Dunson (2011). We applied our methodology to a real microbiome dataset, aiming to cluster individuals characterized by similar taxa distributions. Controlling for each subject's library size, we grouped the data minimizing the Variation of Information loss function, and showed how the model detects clusters catching main differences among the distributions. In our application, the distributional clustering we recover distinguishes among dietary patterns, discriminating African high fiber from Western high fats diets. The observational clustering provides insights about the abundance levels among taxa and helps the identification of co-expression networks. We also assess the performance of our modeling approach through a simulation study where the data are simulated from highly overlapping distributions.

The application of the proposed model to the real data set is limited by the type and number of clinical and demographic covariates that are available. If additional covariates were available, they could be used to define more complex dependencies, e.g. by constructing dependent random measures with covariate-dependent weights as in MacEachern (2000) (see, also Barrientos et al., 2012) or to build risk-prediction models. Another interesting extension considers the incorporation of a time dimension and the study of how distributional clusters vary across time. We leave these directions to future investigation. The code employed for this paper is openly available at `https://github.com/Fradenti/CommonAtomModel`

Supplementary Material

A. Proofs

A.1. Proof of Equation (6)

Let G_j and $G_{j'}$, with $j \neq j'$, be two random probability measures as defined in (1)-(2). Then,

$$\begin{aligned}\mathbb{P}\left(G_j = G_{j'}|Q\right) &= \sum_{k\geq 1}\mathbb{P}\left(G_j = G_{j'} = G_k^*|Q\right) = \sum_{k\geq 1}\mathbb{P}\left(G_j = G_k^*, G_{j'} = G_k^*|Q\right)\\ &= \sum_{k\geq 1}\mathbb{P}\left(G_j = G_k^*|Q\right)\mathbb{P}\left(G_{j'} = G_k^*|Q\right) = \sum_{k\geq 1}\pi_k^2 > 0.\end{aligned}$$

As a consequence we get

$$\mathbb{P}\left(G_j = G_{j'}\right) = \mathbb{E}\left[\mathbb{P}\left(G_j = G_{j'}|Q\right)\right] = \mathbb{E}\left[\sum_{k\geq 1}\pi_k^2\right] = \sum_{k\geq 1}\mathbb{E}\left[\pi_k^2\right],$$

exploiting the stick–breaking representation of the π_k's we have

$$\mathbb{P}\left(G_j = G_{j'}\right) = \sum_{k\geq 1}\mathbb{E}\left[V_k^2\prod_{i=1}^{k-1}(1-V_i)^2\right] = \sum_{k\geq 1}\frac{B(3,\alpha)}{B(1,\alpha)}\left[\frac{B(1,\alpha+2)}{B(1,\alpha)}\right]^{k-1}$$

where we denoted by $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ the beta function. Some simple calculations show that

$$\mathbb{P}\left(G_j = G_{j'}\right) = \frac{2}{(1+\alpha)(2+\alpha)}\sum_{k\geq 0}\left[\frac{\alpha}{\alpha+2}\right]^k = \frac{1}{\alpha+1},$$

and then (6) follows.

A.2. Proof of Equation (7)

Let $y_{i,j}|G_j \sim G_j$ and $y_{i',j'}|G_{j'} \sim G_{j'}$ be two observations coming from two probability measures both sampled from Q for $j \neq j'$. Then,

$$\begin{aligned}
\mathbb{P}\left[y_{i,j} = y_{i',j'}\right] &= \mathbb{E}\left[\mathbb{P}\left[y_{i,j} = y_{i',j'}|G_j, G_{j'}\right]\right] \\
&= \mathbb{E}\left[\frac{1}{1+\alpha}\mathbb{P}\left[y_{i,j} = y_{i',j'}|G_j = G_{j'}\right] + \frac{\alpha}{1+\alpha}\mathbb{P}\left[y_{i,j} = y_{i',j'}|G_j \neq G_{j'}\right]\right] \\
&= \frac{1}{1+\alpha}\mathbb{E}\left[\sum_{r\geq 1}\omega_{r,j}^2\right] + \frac{\alpha}{1+\alpha}\mathbb{E}\left[\sum_{r\geq 1}\omega_{r,j}\omega_{r,j'}\right]. \qquad (16)
\end{aligned}$$

To conclude the proof we evaluate the two expected values in the last equation. As for the first one, it is easy to observe that

$$\mathbb{E}\left[\sum_{r\geq 1}\omega_{r,j}^2\right] = \frac{1}{1+\beta}$$

using the stick-breaking representation of the weights as in Section A.1. As for the second expected value we exploit the independence across the $\omega_{r,j}$'s, for different values of j, and we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{r\geq 1}\omega_{r,j}\omega_{r,j'}\right] &= \sum_{r\geq 1}\mathbb{E}\left[\omega_{r,j}\right]\mathbb{E}\left[\omega_{r,j'}\right] = \sum_{r\geq 1}\mathbb{E}\left[\omega_{r,j}\right]^2 \\
&= \sum_{r\geq 1}\left[\mathbb{E}\left[V_r\right]\prod_{i=1}^{r-1}\mathbb{E}\left[1-V_i\right]\right]^2 = \sum_{r\geq 1}\left[\frac{1}{1+\beta}\left(\frac{\beta}{1+\beta}\right)^{r-1}\right]^2 \\
&= \frac{1}{2\beta+1}
\end{aligned}$$

where the last equality follows by straightforward calculations. Substituting the previous expressions in (16) we finally obtain

$$\mathbb{P}\left[y_{i,j} = y_{i',j'}\right] = \frac{1}{1+\alpha}\frac{1}{1+\beta} + \frac{\alpha}{1+\alpha}\frac{1}{2\beta+1}$$

and Equation (7) follows.

A.3. Proof of Equations (8)–(9)

Suppose that the G_j's are defined on a Polish space $(\mathbb{X}, \mathcal{X})$ and consider $A, B \in \mathcal{X}$. Recall that $G_j, G_{j'}|Q \overset{i.i.d.}{\sim} Q$, where $Q = \sum_{k\geq 1}\pi_k\delta_{G_k^*}$. In the following, for the sake of

notational simplicity and without loss of generality, we suppose that $j = 1$ and $j' = 2$. We now focus on the proof of (8), for this reason we first evaluate

$$\begin{aligned}\mathbb{E}\left[G_1(A)G_2(B)\right] &= \mathbb{E}\left[\mathbb{E}\left[G_1(A)\cdot G_2(B)|Q\right]\right] \\ &= \mathbb{E}\left[\sum_{k\geq 1}\pi_k^2 G_k^*(A)G_k^*(B) + \sum_{k_1\neq k_2}\pi_{k_1}\pi_{k_2}G_{k_1}^*(A)G_{k_2}^*(B)\right]\end{aligned}$$

Since the G_k^*'s are independent and identically distributed and thanks to the fact that

$$\mathbb{P}\left[G_1 = G_2\right] = \mathbb{E}\left[\sum_{k\geq 1}\pi_k^2\right],$$

we can equivalently write

$$\mathbb{E}\left[G_1(A)G_2(B)\right] = \mathbb{P}\left[G_1 = G_2\right]\mathbb{E}\left[G_1^*(A)G_1^*(B)\right] + \mathbb{P}\left[G_1 \neq G_2\right]\mathbb{E}\left[G_1^*(A)G_2^*(B)\right].$$

In view of Equation (6), the previous expression boils down to the following one

$$\mathbb{E}\left[G_1(A)G_2(B)\right] = \frac{1}{\alpha+1}\mathbb{E}\left[G_1^*(A)G_1^*(B)\right] + \frac{\alpha}{\alpha+1}\mathbb{E}\left[G_1^*(A)G_2^*(B)\right]. \tag{17}$$

We now focus on the evaluation of the two expected values in (17). The first one can be expressed as

$$\begin{aligned}\mathbb{E}\left[G_1^*(A)G_1^*(B)\right] =& \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,1}\delta_{\theta_l}(A)\cdot\sum_{l\geq 1}\omega_{l,1}\delta_{\theta_l}(B)\right] \\ =& \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,1}^2\delta_{\theta_l}(A\cap B)\right] + \mathbb{E}\left[\sum_{l\geq 1}\sum_{r\neq l}\omega_{l,1}\omega_{r,1}\delta_{\theta_l}(A)\delta_{\theta_r}(B)\right] \\ =& \mathbb{E}\left[\sum_{l\geq 1}\omega_{l,1}^2\right]H(A\cap B) + \left(1 - \sum_{l\geq 1}\mathbb{E}\left[\omega_{l,1}^2\right]\right)H(A)H(B) \\ =& \frac{1}{\beta+1}H(A\cap B) + \frac{\beta}{\beta+1}H(A)H(B),\end{aligned}$$

where we used the fact that H is the distribution of the atoms and

$$\mathbb{E}\left[\sum_{l\geq 1}\omega_{l,1}^2\right] = \frac{1}{1+\beta}.$$

The second expectation in (17) can be evaluated as follows:

$$\begin{aligned}\mathbb{E}\left[G_1^*(A)G_2^*(B)\right] =& \mathbb{E}\left[\sum_{r\geq 1}\omega_{r,1}\delta_{\theta_r}(A)\cdot\sum_{l\geq 1}\omega_{l,2}\delta_{\theta_l}(B)\right]\\ =& \mathbb{E}\left[\sum_{r\geq 1}\omega_{r,1}\omega_{r,2}\delta_{\theta_r}(A\cap B)\right]+\mathbb{E}\left[\sum_{r\neq l}\omega_{r,1}\omega_{l,2}\delta_{\theta_r}(A)\delta_{\theta_l}(B)\right]\\ =& \mathbb{E}\left[\sum_{r\geq 1}\omega_{r,1}\omega_{r,2}\right]H(A\cap B)+\mathbb{E}\left[\sum_{r\neq l}\omega_{r,1}\omega_{l,2}\right]H(A)H(B).\end{aligned}$$

We note that the previous equality holds true in particular when $A = B = \mathbb{X}$. In that case

$$1 = \mathbb{E}\left[G_1^*(\mathbb{X})\cdot G_2^*(\mathbb{X})\right] = \sum_{r\geq 1}\mathbb{E}\left[\omega_{r,1}\right]\mathbb{E}\left[\omega_{r,2}\right]H(\mathbb{X}) + \sum_{r\neq l}\mathbb{E}\left[\omega_{r,1}\right]\mathbb{E}\left[\omega_{l,2}\right]H(\mathbb{X})H(\mathbb{X})$$

which is tantamount to saying that

$$1-\sum_{r\geq 1}\mathbb{E}\left[\omega_{r,1}\omega_{r,2}\right] = \sum_{r\neq l}\mathbb{E}\left[\omega_{r,1}\right]\mathbb{E}\left[\omega_{l,2}\right]. \tag{18}$$

Coming back to the evaluation of $\mathbb{E}\left[G_1^*(A)\cdot G_2^*(B)\right]$, we have:

$$\begin{aligned}\mathbb{E}\left[G_1^*(A)\cdot G_2^*(B)\right] &= \sum_{r\geq 1}\mathbb{E}\left[\omega_{r,1}\right]\mathbb{E}\left[\omega_{r,2}\right]H(A\cap B)+\sum_{r\neq l}\mathbb{E}\left[\omega_{r,1}\right]\mathbb{E}\left[\omega_{l,2}\right]H(A)H(B)\\ &= \sum_{r\geq 1}\left\{\mathbb{E}\left[\omega_{r,1}\right]\right\}^2 H(A\cap B)+\left(1-\sum_{r\geq 1}\left\{\mathbb{E}\left[\omega_{r,1}\right]\right\}^2\right)H(A)H(B),\end{aligned} \tag{19}$$

where we used (18) and the fact that $\omega_{r,1}$ and $\omega_{r,2}$ are independent and identically distributed. It remains to evaluate the infinite series over $r \geq 1$ in (19), and this issue may be easily addressed, indeed:

$$\begin{aligned}\sum_{r\geq 1}\left\{\mathbb{E}\left[\omega_{r,1}\right]\right\}^2 &= \sum_{r\geq 1}\left\{\mathbb{E}\left[V_r\prod_{q=1}^{r-1}(1-V_q)\right]\right\}^2\\ &= \sum_{r\geq 1}\left[\frac{1}{(1+\beta)^2}\left(\frac{\beta}{1+\beta}\right)^{2(r-1)}\right] = \frac{1}{2\beta+1}.\end{aligned}$$

Substituting the previous expression in (19), we get:

$$\mathbb{E}\left[G_1^*(A)G_2^*(B)\right] = \frac{1}{1+2\beta}H(A\cap B)+\frac{2\beta}{1+2\beta}H(A)H(B).$$

Putting the expressions of $\mathbb{E}\left[G_1^*(A)G_2^*(B)\right]$ and $\mathbb{E}\left[G_1^*(A)G_1^*(B)\right]$ in (17), we obtain

$$\begin{aligned}&\mathbb{E}\left[G_1(A)G_2(B)\right]\\&\quad= H(A\cap B)\left(\frac{q_1}{1+\beta}+\frac{1-q_1}{1+2\beta}\right)+H(A)H(B)\left(q_1\frac{\beta}{1+\beta}+(1-q_1)\frac{2\beta}{1+2\beta}\right),\end{aligned}\tag{20}$$

where we recall that $q_1=\frac{1}{\alpha+1}$. We can use (20) to evaluate the covariance between $G_j(A)$ and $G_{j'}(A)$ for $j\neq j'$:

$$\begin{aligned}&Cov\left(G_j(A),G_{j'}(B)\right)\\&\quad= H(A\cap B)\left(\frac{q_1}{1+\beta}+\frac{1-q_1}{1+2\beta}\right)+H(A)H(B)\left(q_1\frac{\beta}{1+\beta}+(1-q_1)\frac{2\beta}{1+2\beta}-1\right)\\&\quad= H(A\cap B)\left(\frac{q_1}{1+\beta}+\frac{1-q_1}{1+2\beta}\right)+H(A)H(B)\left(-\frac{q_1}{1+\beta}-\frac{1-q_1}{1+2\beta}\right),\end{aligned}$$

hence (8) is now proved.

As for the determination of the correlation (9), we first specialize (8) when $A=B\in\mathcal{X}$, to get:

$$Cov\left(G_j(A),G_{j'}(A)\right)=\left(\frac{q_1}{1+\beta}+\frac{1-q_1}{1+2\beta}\right)H(A)(1-H(A)),\tag{21}$$

and then we divide $Cov\left(G_j(A),G_{j'}(A)\right)$ by the squared roots of the variances $Var(G_j(A)$ and $Var(G_{j'}(A)$. More precisely we have:

$$\begin{aligned}Corr\left(G_j(A),G_{j'}(A)\right)&=\frac{Cov\left(G_j(A),G_{j'}(A)\right)}{\sqrt{Var(G_j(A))\cdot Var(G_{j'}(A))}}\\&\overset{(21)}{=}\frac{H(A)(1-H(A))}{\sqrt{Var(G_j(A))\cdot Var(G_{j'}(A))}}\left(\frac{q_1}{1+\beta}+\frac{1-q_1}{1+2\beta}\right).\end{aligned}\tag{22}$$

where the variances in the denominator may be easily evaluated as follows

$$\begin{aligned}Var(G_j(A))=&\mathbb{E}\left[G_j(A)^2\right]-\mathbb{E}\left[G_j(A)\right]^2\\=&\mathbb{E}\left[\mathbb{E}\left[G_j(A)^2|Q\right]\right]-\mathbb{E}\left[G_j(A)\right]^2=\mathbb{E}\left[G_1^*(A)^2\right]-\mathbb{E}\left[G_1^*(A)\right]^2\\=&\frac{1}{\beta+1}H(A)+\frac{\beta}{1+\beta}H(A)^2-H(A)^2\\=&\frac{1}{\beta+1}H(A)\left(1-H(A)\right),\end{aligned}$$

for any $j = 1, \dots, J$. Putting the previous expression in (22) we get:

$$\begin{aligned}\rho_{j,j'} = Corr(G_j(A), G_{j'}(A)) &= \left(\frac{q_1}{\beta+1} + \frac{1-q_1}{2\beta+1}\right) \Big/ \frac{1}{\beta+1} \\ &= q_1 + \frac{\beta+1}{2\beta+1}(1-q_1) = 1 - \frac{\beta}{2\beta+1}(1-q_1) \\ &= 1 - \frac{\beta}{2\beta+1} \cdot \frac{\alpha}{1+\alpha},\end{aligned}$$

and (21) is now proved. From the last expression, we finally observe that $\rho_{j,j'}$ is always in between $1/2$ and 1.

A.4. Proof of Proposition 1

Recalling the CAM model (1)-(2), we get

$$\begin{aligned}&\mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1) f_2(g_2) Q(dg_1) Q(dg_2)\right] \\ &\qquad = \mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1) f_2(g_2) \sum_{k_1 \geq 1} \pi_{k_1} \delta_{G^*_{k_1}}(dg_1) \sum_{k_2 \geq 1} \pi_{k_2} \delta_{G^*_{k_2}}(dg_2)\right] \\ &\qquad = \mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1) f_2(g_2) \sum_{k \geq 1} \pi_k^2 \delta_{G^*_k}(dg_1) \delta_{G^*_k}(dg_2)\right] \\ &\qquad\quad + \mathbb{E}\left[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(g_1) f_2(g_2) \sum_{k_1 \neq k_2} \pi_{k_1} \pi_{k_2} \delta_{G^*_{k_1}}(dg_1) \delta_{G^*_{k_2}}(dg_2)\right].\end{aligned}$$

Observe that the G^*_k's are all Dirichlet processes having the same law on the space $\mathsf{P}_{\mathbb{X}}$, which will be denoted by $\mathcal{P}$, depending on the total mass α and the base measure H. We also point out that the G^*_k's are not independent random elements for different values of k, indeed they share the same random atoms $(\theta_l)_{l \geq 1}$, nevertheless if $k_1 \neq k_2$, the distribution of $(G^*_{k_1}, G^*_{k_2})$ equals the distribution of (G^*_1, G^*_2), which will be denoted by

$\mathcal{P}_{[2]}$. Therefore, by applying the Tonelli–Fubini Theorem, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\int_{\mathsf{P}^2_{\mathbb{X}}} f_1(g_1)f_2(g_2)Q(dg_1)Q(dg_2)\right] \\
&\qquad = \sum_{k\geq 1}\mathbb{E}\pi_k^2\mathbb{E}\int_{\mathsf{P}^2_{\mathbb{X}}} f_1(g_1)f_2(g_2)\delta_{G_k^*}(dg_1)\delta_{G_k^*}(dg_2) \\
&\qquad\qquad + \sum_{k_1\neq k_2}\mathbb{E}\pi_{k_1}\pi_{k_2}\mathbb{E}\int_{\mathsf{P}^2_{\mathbb{X}}} f_1(g_1)f_2(g_2)\delta_{G_{k_1}^*}(dg_1)\delta_{G_{k_2}^*}(dg_2) \\
&\qquad = q_1\int_{\mathsf{P}^2_{\mathbb{X}}} f_1(g)f_2(g)\mathcal{P}(dg) + (1-q_1)\int_{\mathsf{P}^2_{\mathbb{X}}} f_1(g_1)f_2(g_2)\mathcal{P}_{[2]}(dg_1,dg_2),
\end{aligned}$$

and then the thesis follows.

A.5. Proof of Theorem 1

We first evaluate the expected value in the definition of pEPPF (4), for $J = 2$,

$$\begin{aligned}
\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*) &= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{2}\prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*)\Big|Q\right]\right] \\
&= \mathbb{E}\left[\int_{\mathsf{P}^2_{\mathbb{X}}}\prod_{j=1}^{2}\prod_{i=1}^{s} g_j^{n_{i,j}}(dy_i^*)Q(dg_1)Q(dg_2)\right].
\end{aligned}$$

Now we apply Equation (3) to the previous integral where the functions f_j, as $j = 1, 2$, are defined by

$$f_j(g_j) := \prod_{i=1}^{s} g_j^{n_{i,j}}(dy_i^*),$$

and then we get

$$\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s} G_j^{n_{i,j}}(dy_i^*) = q_1\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G_1^*)^{n_{i,j}}(dy_i^*) + (1-q_1)\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G_j^*)^{n_{i,j}}(dy_i^*). \tag{23}$$

We finally integrate over the space $\mathbb{X}^s$ to get the result, i.e. (5).

A.6. Proof of Proposition 2

Assume that the two samples $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ share $s_0 > 0$ distinct values denoted here as $y_{1,0}^*, \dots y_{s_0,0}^*$ with frequencies $(q_{1,j}, \dots, q_{s_0,j})$ in the j-th sample, as $j = 1, 2$. We further

suppose that the j-th sample contains exactly s_j distinct observations not shared with the other one, and denoted here by $y^*_{1,j}, \ldots, y^*_{s_j,j}$, for $j = 1, 2$; besides the vector of corresponding frequencies will be denoted as $(r_{1,j}, \ldots, r_{s_j,j})$. We obviously have that $s = s_0 + s_1 + s_2$.

Using the representation of the G^*_k's in the CAM model (1)–(2), we get

$$\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i) = \mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}\left(\sum_{l\geq 1}\omega_{l,j}\delta_{\theta_l}(dy^*_i)\right)^{n_{i,j}}.$$

Exploiting the partition of the data described at the beginning of the proof, we obtain

$$\begin{aligned}
&\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i) \\
&\quad= \mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s_j}\left(\sum_{l\geq 1}\omega_{l,j}^{r_{i,j}}\delta_{\theta_l}(dy^*_{i,j})\right)\prod_{i=1}^{s_0}\left(\sum_{l\geq 1}\omega_{l,1}^{q_{i,1}}\omega_{l,2}^{q_{i,2}}\delta_{\theta_l}(dy^*_{i,0})\right) + o\left(\prod_{j=0}^{2}\prod_{i=1}^{s_j}H(dy^*_{i,j})\right) \\
&\quad= \sum_{\neq}\mathbb{E}\left[\prod_{j=1}^{2}\prod_{i=1}^{s_j}\omega_{l_{i,j},j}^{r_{i,j}}\prod_{i=1}^{s_0}\omega_{l_{i,0},1}^{q_{i,1}}\omega_{l_{i,0},2}^{q_{i,2}}\right]\prod_{j=0}^{2}\prod_{i=1}^{s_j}H(dy^*_{i,j}) + o\left(\prod_{j=0}^{2}\prod_{i=1}^{s_j}H(dy^*_{i,j})\right).
\end{aligned}$$

where the sum $\sum_{\neq}$ is extended over all possible values of the distinct natural numbers $\{l_{i,j} : \ i = 1, \ldots, s_j, \ j = 0, 1, 2\}$. Integrating over $\mathbb{X}^s$ we get that

$$\int_{\mathbb{X}^s}\mathbb{E}\prod_{j=1}^{2}\prod_{i=1}^{s}(G^*_j)^{n_{i,j}}(dy^*_i) = \sum_{\neq}\mathbb{E}\left[\prod_{j=1}^{2}\prod_{i=1}^{s_j}\omega_{l_{i,j},j}^{r_{i,j}}\prod_{i=1}^{s_0}\omega_{l_{i,0},1}^{q_{i,1}}\omega_{l_{i,0},2}^{q_{i,2}}\right] \tag{24}$$

which is positive whenever $s_0 > 0$.

B. Truncated Blocked Gibbs Sampler for CAM

The posterior distribution is analytically intractable, which forces us to develop sampling algorithms to simulate from it. A Pólya Urn representation would be too expensive in computational cost. Instead, we provide two different algorithms: a Blocked Gibbs sampler (Ishwaran and James, 2001), mimicking the one proposed in (Rodríguez et al., 2008) and a nested slice sampler (Damien et al., 1999; Walker, 2007; Kalli et al., 2011).

Here we discuss the former one. The Truncated CAM model has the following form:

$$\begin{aligned} y_{i,j}|\boldsymbol{M},\boldsymbol{\theta} &\sim N\left(\cdot|\theta_{M_{i,j}}\right), & M_{i,j}|\boldsymbol{S},\boldsymbol{\omega} &\sim \sum_{l=1}^{L}\omega_{l,S_j}\delta_l(\cdot), \\ \boldsymbol{\omega}_k|\boldsymbol{S}=\boldsymbol{\omega}_k &\sim GEM(\alpha), & S_j|\boldsymbol{\pi} &\sim \sum_{k=1}^{K}\pi_k\delta_k(\cdot), \\ \boldsymbol{\pi} &\sim GEM(\beta), & \theta_l &\sim \pi(\theta_l). \end{aligned} \tag{25}$$

The Truncated version of CAM (TCAM) (25) can be extended to a Truncated version of DCAM (TDCAM) once the likelihood is modified according to (12), In the following we report the Gibbs Sampler for the TDCAM, the extension of the sampler to accomodate the presence of a covariate linearly introduced. Notice that some of the conditioning variables are collapsed (Liu, 1994), to enhance the speed of convergence and the mixing of the chains.

B.1. TDCAM: Gibbs Sampler

Denote with $\boldsymbol{V}$ the vector containing all the variables of model (25), and let $\boldsymbol{V}^{-\boldsymbol{s}}$ be the same vector $\boldsymbol{V}$ with the variable $\boldsymbol{s}$ removed.
The steps of the MCMC are the following:

1. The full conditional for each $y_{i,j}$ is Truncated Normal, with support $\left[a_{z_{i,j}}, a_{z_{i,j}+1}\right)$:

$$p\left(y_{i,j}|\boldsymbol{V}\right) \sim TN(\mu_{M_{i,j}}, \sigma^2_{M_{i,j}}; a_{z_{i,j}}, a_{z_{i,j}+1}).$$

   This can be easily done with the help of the R package `TruncatedNormal`, which relies on a recently improved algorithm exploiting minmax tilting (Botev, 2017).

2. The full conditional for the observational cluster labels $M_{i,j}$, once the latent variable $\boldsymbol{y}$ is integrated out, is a discrete distribution, given by

$$p\left(M_{i,j}=l|\boldsymbol{V}^{-\boldsymbol{y}}\right) \propto \omega_{l,S_j}\Delta\Phi\left(a_{z_{i,j}}; \mu_{M_{i,j}}, \sigma^2_{M_{i,j}}\right),$$

   for any $i=1,\ldots,n_j$, $j=1,\ldots,J$.

3. The full conditional for the distributional cluster labels S_j is given by:

$$p\left(S_j = k|\boldsymbol{V}^{-(\boldsymbol{y},\boldsymbol{M})}\right) \propto \pi_k \prod_{i=1}^{n_j} \left(\sum_{m=1}^{L} \omega_{m,k} \Delta\Phi\left(a_{y_{i,j}}; \mu_m, \sigma_m^2\right) \right),$$

for any $j = 1, \ldots, J$.

4. To sample the full conditional of the weights $\boldsymbol{\pi}$ at the distributional level, we first need to define m_k^* as the number of groups assigned to the same distributional cluster k, where $\sum_{k=1}^{K} m_k^* = J$ the total number of observed groups. Then,

$$p\left(\boldsymbol{\pi}|\boldsymbol{V}\right) \propto p\left(\boldsymbol{S}|\boldsymbol{\pi}\right) p\left(\boldsymbol{\pi}\right) \propto p\left(\boldsymbol{\pi}\right) \pi_1^{m_1^*} \cdots \pi_K^{m_K^*}.$$

Referring to the Stick Breaking representation, the full conditional of the different sticks v_k, as $k = 1, \ldots, K$, equals:

$$v_k \sim Beta\left(1 + m_k^*, \beta + \sum_{s=k+1}^{K} m_k^*\right).$$

5. The derivation of the full conditional for $\boldsymbol{\omega}$ is similar, even it requires more care. We have

$$p\left(\boldsymbol{\omega}|\boldsymbol{V}\right) \propto p\left(\boldsymbol{M}|\boldsymbol{S}, \boldsymbol{\omega}, \boldsymbol{\xi_0}\right) p\left(\boldsymbol{\omega}\right) \propto \prod_{k=1}^{K} p\left(\boldsymbol{\omega_k}\right) \prod_{j=1}^{J} \prod_{i=1}^{n_j} \left(\sum_{l=1}^{L} \omega_{l,S_j} \delta_l(\cdot) \right).$$

The previous formula can be decomposed into the product of K elements and we can focus only on the case $S_j = k$. Let us define $n_{l,k}$ as the total number of observations assigned to the distributional cluster k in the observational group l. The full conditional has this Stick-Breaking representation for $u_{l,k}, \forall k$:

$$u_{l,k} \sim Beta\left(1 + n_{l,k}, \alpha + \sum_{r=l+1}^{L} n_{r,k}\right), \quad l = 1, \ldots, L.$$

6. Let us define $n_{l,\cdot} = \sum_{k=1}^{K} n_{l,k}$ and

$$\bar{y}_{l,\cdot} := \frac{1}{n_{l,\cdot}} \sum_{i,j:M_{i,j}=l} y_{i,j}.$$

Exploiting the conjugacy property, we obtain the full conditional for $\theta_l = \left(\mu_l, \sigma_l^2\right)$:

$$\left(\mu_l, \sigma_l^2\right) | \boldsymbol{V} \sim NIG\left(m_0^*, \kappa_0^*, \alpha_0^*, \beta_0^*\right).$$

where

$$m_0^* = \frac{\kappa_0 m_0 + n_{l,\cdot}\bar{y}_{l,\cdot}}{\kappa_0 + n_{l,\cdot}} \quad \kappa_0^* = \kappa_0 + n_{l,\cdot} \quad \alpha_0^* = \alpha_0 + n_{l,\cdot}/2$$

and

$$\beta_0^* = \beta + 0.5\left(\sum_{ij:\, M_{i,j}=l} \left(y_{i,j} - \bar{y}_{l,\cdot}\right)^2 + \left(\frac{\kappa_0 n_{l,\cdot}}{\kappa_0 + n_{l,\cdot}}\right)\left(y_{l,k} - m_0\right)^2\right).$$

7. In case the precision parameters α and β of the two DPs are assumed stochastic, distributed as $Gamma\left(a_\alpha, b_\alpha\right)$ and $Gamma\left(a_\beta, b_\beta\right)$, we can still exploiting conjugacy. The full conditionals distributions are:

$$\alpha|\boldsymbol{V} \sim Gamma\left(a_\alpha + (K-1), b_\alpha - \sum_{k=1}^{K-1}\log(1-v_k)\right),$$
$$\beta|\boldsymbol{V} \sim Gamma\left(a_\beta + K\cdot(L-1), b_\beta - \sum_{l=1}^{L-1}\sum_{k=1}^{K}\log(1-u_{l,k})\right).$$

Notice that we naturally set $a_{y_{i,j}} = y_{i,j}$. As suggested in (Rodríguez et al., 2008), each step of this algorithm can be parallelized, in order to gain computational speed.

B.2. Linearly incorporating a covariate in the Likelihood

If we want to linearly add regressor to the mean, we update model (25) simply assuming:

$$z_{i,j}|y_{i,j} \sim \sum_{g=0}^{+\infty}\delta_g(\cdot)\mathbf{1}_{[a_g, a_{g+1})}\left(y_{i,j}\right) \qquad y_{i,j}|\boldsymbol{M}, \boldsymbol{\mu}, \boldsymbol{\sigma^2} \sim N\left(\mu_{M_{i,j}} + \beta X_j, \sigma^2_{M_{i,j}}\right). \tag{26}$$

We espouse such representation because of its interpretability: the latent continuous random variable $y_{i,j}$ can be decomposed as $y_{i,j} = \mu_{M_{i,j}} + \beta X_j + \varepsilon_{i,j}$, where $\varepsilon \sim N\left(0, \sigma^2_{M_{i,j}}\right)$. In other words, we model the every single latent value as the sum of an effect specific for each observational cluster, an effect due to the regressor value of each individual multiplied by a overall coefficient and a completely random effect, whose entity still depends

on the observational cluster. This choice does not complicate the algorithm presented in the previous section: the full conditionals 1-3 are preserved if the mean is modified accordingly, switching from $\mu_{M_{i,j}}$ to $\mu_{M_{i,j}} + \beta X_j$. Step 6 remains the same once we substitute $y_{i,j}$ with $d_{i,j} = y_{i,j} - \beta X_j$. Steps 4, 5 and 7 are not affected by this change. Finally, if we assume $\beta \sim N\left(m_\beta, \frac{1}{\kappa_\beta}\right)$, we can perform inference on the introduced coefficient. Define $R^1 = \sum_{i,j} \frac{X_j^2}{\sigma^2_{M_{i,j}}}$ and $R^2 = \sum_{i,j} \frac{d_{i,j} \cdot X_j}{\sigma^2_{M_{i,j}}}$. The full conditional for β is:

$$\beta | \boldsymbol{V} \sim N\left(\frac{m_\beta \kappa_\beta + R^2}{\kappa_\beta + R^1}, \frac{1}{\kappa_\beta + R^1}\right).$$

This framework can be easily extended to accommodate for the presence of multiple covariates.

C. Error bounds in total variation distance

In Section B we have depicted a truncated blocked Gibbs sampler, we now evaluate the truncation error arising from these algorithms for the CAM model (Section C.1) and the CAMM (Section C.2). The errors between the random distribution and its truncated version will be evaluated using the total variation distance. For the reader's convenience we recall that if $P, Q \in \mathsf{P}_{\mathbb{X}}$ are probability measures defined on $(\mathbb{X}, \mathcal{X})$, the distance in total variation between P and Q is defined as

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{X}} |P(A) - Q(A)|.$$

If P, Q are absolutely continuous w.r.t. a measure μ then it can be expressed as

$$d_{TV}(P, Q) = \frac{1}{2} \int_{\mathcal{X}} \left| \frac{dP}{d\mu} - \frac{dQ}{d\mu} \right| d\mu.$$

Moreover, if $\mathcal{X}$ is a discrete space or if P and Q are concentrated on a countable set $\Omega \subset \mathcal{X}$ then

$$d_{TV}(P, Q) = \frac{1}{2} \sum_{x \in \Omega} \Big| P(x) - Q(x) \Big|.$$

C.1. Truncation error in CAM

In this section we quantify the error committed when we replace the random probability measures G_j with the corresponding truncated versions. We recall that $G_1, \ldots, G_J | Q \overset{i.i.d.}{\sim} Q$, where Q has been defined in (2):

$$Q = \sum_{k \geq 1} \pi_k \, \delta_{G_k^*}, \quad G_k^* = \sum_{l \geq 1} \omega_{l,k} \delta_{\theta_l}.$$

In order to formally define the truncated versions of $G_1, \ldots, G_J$, we exploit the latent random variable $\xi_j | Q \overset{iid}{\sim} \sum_{k=1}^{+\infty} \pi_k \delta_k$, as $j = 1, \ldots, J$, which identifies the mixture component from which G_j is generated, conditionally on Q. Thus, conditionally on the value $\xi_j = k$, the truncated random probability measures associated to each G_j are formally defined as follows

$$G_j^{(K,L)} = \begin{cases} \sum_{l=1}^{L} \omega_{l,k}^{(K,L)} \delta_{\theta_l} & \text{if } \xi_j \leq K \\ \sum_{l=1}^{L} \omega_{l,K}^{(K,L)} \delta_{\theta_l} & \text{if } \xi_j > K \end{cases} \tag{27}$$

and

$$\begin{aligned} \omega_{l,k}^{(K,L)} &= \omega_{l,k} \qquad \text{if } l \leq L-1, \quad \text{and} \quad \omega_{L,k}^{(K,L)} = 1 - \omega_{1,k} - \ldots - \omega_{L-1,k} \\ \pi_k^{(K,L)} &= \pi_k \qquad \text{if } k \leq K-1, \quad \text{and} \quad \pi_K^{(K,L)} = 1 - \pi_1 - \ldots - \pi_{K-1} \end{aligned}$$

where $K, L > 0$ define the truncation levels for the different random probability measures.

Proposition 3. *Let $G_j | Q \sim Q$ and $G_j^{(K,L)}$ the truncation of G_j defined in (27), then the expected value of the distance in total variation between them can be estimated as follows:*

$$\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)} \right) \right] \leq \left(1 - \left(\frac{\alpha}{1+\alpha} \right)^K \right) \left(\frac{\beta}{1+\beta} \right)^L + \left(\frac{\alpha}{1+\alpha} \right)^K, \tag{28}$$

for any $j = 1, \ldots, J$.

Proof. First of all observe that, conditioning on $\xi_j = k$, we recognize two distinct situations to upper bound the total variation distance between G_j and its truncated counterpart as described below.

1. If $\xi_j = k \leq K$, then we have

$$
\begin{aligned}
d_{TV}(G_j, G_j^{(K,L)}) &= \frac{1}{2}\left(\sum_{l=1}^{L} |\tilde{\omega}_{lk} - \tilde{\omega}_{lk}^{(K,L)}| + |\tilde{\omega}_{lk} - \tilde{\omega}_{lk}^{(K,L)}| + \sum_{l \geq L+1} |\tilde{\omega}_{lk} - 0|\right) \\
&= \frac{1}{2}\left(|\tilde{\omega}_{Lk} - 1 + \tilde{\omega}_{1k} + \ldots + \tilde{\omega}_{L-1k}| + \sum_{l \geq L+1} \tilde{\omega}_{lk}\right) \\
&= \frac{1}{2}\left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk} + \sum_{l \geq L+1} \tilde{\omega}_{lk}\right) = \left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right).
\end{aligned}
$$

2. If $\xi_j > K$, we use the following trivial upper bound $d_{TV}(G_j, G_j^{(K,L)}) \leq 1$.

In light of the previous considerations, we are now ready to compute

$$
\begin{aligned}
\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)|\xi_j, Q\right]\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K} \pi_k \mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)|\xi_j = k, Q\right]\right] \\
&\qquad + \mathbb{E}\left[\sum_{k=K+1}^{+\infty} \pi_k \mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)|\xi_j = k, Q\right]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^{K} \pi_k \left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right) + \sum_{k=K+1}^{+\infty} \pi_k\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{K} \pi_k\right] \cdot \mathbb{E}\left[\left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right)\right] + \mathbb{E}\left[\sum_{k=K+1}^{+\infty} \pi_k\right] \\
&\leq \mathbb{E}\left[\left(1 - \sum_{l=1}^{L} \tilde{\omega}_{lk}\right)\right] + \mathbb{E}\left[1 - \sum_{k=1}^{K} \pi_l\right] \\
&= \left(1 - \left(\frac{\alpha}{1+\alpha}\right)^K\right)\left(\frac{\beta}{1+\beta}\right)^L + \left(\frac{\alpha}{1+\alpha}\right)^K
\end{aligned}
$$

where the last equality follows by straightforward calculations based on the stick–breaking representation of the weights.

□

C.2. Approximation Error in Mixture Models (CAMM)

Consider J groups, each of them containing n_j observations, $j = 1, \ldots, J$. Denote by $\boldsymbol{y}_j = \left(y_{1,j}, \ldots, y_{n_j,j}\right)$ for $j = 1, \ldots, J$ the observations from the j-th component of the

mixture model $y_{i,j}|\theta_{i,j} \sim f(\cdot|\theta_{i,j})$ with $\theta_{i,j}|G_1, \ldots, G_J \sim G_j$ where the G_j's are generated according to a CAM. We suppose that $\theta_{i,j} \in \Theta$, where Θ is a Polish space equipped with its corresponding Borel σ–field $\mathcal{T}$. We further denote by $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J)$ the vector containing all the observations. We would like to upper bound the distance in total variation between the law of the data $\boldsymbol{y}$

$$\pi(\boldsymbol{y}) = \mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{\Theta} f(y_{i,j}|\theta_{i,j})\, G_j(d\theta_{i,j})\right],$$

and the law of the data $\pi^{(K,L)}$ when the random probability measures G_j's are replaced with the corresponding truncated version $G_j^{(K,L)}$ defined in (27), i.e.

$$\pi^{(K,L)}(\boldsymbol{y}) = \mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{\Theta} f(y_{i,j}|\theta_{i,j})\, G_j^{(K,L)}(d\theta_{i,j})\right].$$

Proposition 4. *The distance in total variation between π and $\pi^{(K,L)}$ satisfies*

$$d_{TV}\left(\pi, \pi^{(K,L)}\right) \leq N\left[\left(\frac{\beta}{1+\beta}\right)^{L} + \left(\frac{\alpha}{1+\alpha}\right)^{K}\right], \tag{29}$$

where $N = n_1 + \cdots + n_J$.

Proof. The distance $d_{TV}\left(\pi, \pi^{(K,L)}\right)$ can be evaluated as follows:

$$\begin{aligned}
&d_{TV}\left(\pi, \pi^{(K,L)}\right) \\
&\quad= \frac{1}{2}\int_{\mathbb{X}^N}\left|\frac{d\pi}{d\boldsymbol{y}} - \frac{d\pi^{K,L}}{d\boldsymbol{y}}\right| d\boldsymbol{y} \\
&\quad= \frac{1}{2}\int_{\mathbb{X}^N}\left|\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{\Theta} f(y_{i,j}|\theta_{i,j})\, G_j(d\theta_{i,j}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{\Theta} f(y_{i,j}|\theta_{i,j})\, G_j^{(K,L)}(d\theta_{i,j})\right]\right| d\boldsymbol{y} \\
&\quad= \frac{1}{2}\int_{\mathbb{X}^N}\left|\mathbb{E}\left[\int_{\Theta^N}\prod_{j=1}^{J}\prod_{i=1}^{n_j} f(y_{i,j}|\theta_{i,j})\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{i,j})\right.\right. \\
&\qquad\qquad\left.\left. - \int_{\Theta^N}\prod_{j=1}^{J}\prod_{i=1}^{n_j} f(y_{i,j}|\theta_{i,j})\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{i,j})\right]\right| d\boldsymbol{y}.
\end{aligned}$$

By an application of the Tonelli-Fubini theorem and the Jensen inequality, we obtain

$$
\begin{aligned}
&d_{TV}\left(\pi, \pi^{(K,L)}\right) \\
&= \frac{1}{2}\int_{\mathbb{X}^N}\left|\int_{\Theta^N}\prod_{j=1}^{J}\prod_{i=1}^{n_j} f\left(y_{i,j}|\theta_{i,j}\right)\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{i,j}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{i,j})\right]\right| d\boldsymbol{y} \\
&\leq \frac{1}{2}\int_{\mathbb{X}^N}\int_{\Theta^N}\prod_{j=1}^{J}\prod_{i=1}^{n_j} f\left(y_{i,j}|\theta_{i,j}\right) d\boldsymbol{y}\left|\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{i,j}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{i,j})\right]\right| \\
&= \frac{1}{2}\int_{\Theta^N}\underbrace{\int_{\mathbb{X}^N}\prod_{j=1}^{J}\prod_{i=1}^{n_j} f\left(y_{i,j}|\theta_{i,j}\right) d\boldsymbol{y}}_{=1}\left|\underbrace{\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{i,j})\right]}_{=:m} - \underbrace{\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{i,j})\right]}_{=:m^{(K,L)}}\right| \\
&= \frac{1}{2}\int_{\Theta^N}\left|\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(d\theta_{i,j}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(d\theta_{i,j})\right]\right| = d_{TV}\left(m, m^{(K,L)}\right) \\
&= \sup_{A_{i,j}\in\mathcal{T}}\left|\mathbb{E}\left[\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(A_{i,j}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(A_{i,j})\right]\right|.
\end{aligned}
$$

We can exchange the expected value with the supremum to get:

$$
\begin{aligned}
d_{TV}\left(\pi, \pi^{(K,L)}\right) &\leq \sup_{A_{i,j}\in\mathcal{T}}\mathbb{E}\left[\left|\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(A_{i,j}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(A_{i,j})\right|\right] \\
&\leq \mathbb{E}\left[\sup_{A_{i,j}\in\mathcal{T}}\left|\prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j(A_{i,j}) - \prod_{j=1}^{J}\prod_{i=1}^{n_j} G_j^{(K,L)}(A_{i,j})\right|\right].
\end{aligned}
$$

We now apply (Billingsley, 1995, Lemma 1, pg. 358) to obtain

$$
\begin{aligned}
d_{TV}\left(\pi, \pi^{(K,L)}\right) &\leq \mathbb{E}\left[\sup_{A_{i,j}\in\mathcal{T}}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left|G_j(A_{i,j}) - G_j^{(K,L)}(A_{i,j})\right|\right] \\
&\leq \mathbb{E}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\sup_{A_{i,j}\in\mathcal{T}}\left|G_j(A_{i,j}) - G_j^{(K,L)}(A_{i,j})\right|\right]
\end{aligned}
$$

where we recognize that $\sup_{A_{i,j}\in\mathcal{T}}\left|G_j(A_{i,j}) - G_j^{(K,L)}(A_{i,j})\right| = d_{TV}\left(G_j, G_j^{(K,L)}\right)$. As a

consequence, by an application of Proposition 3, we get

$$d_{TV}\left(\pi, \pi^{(K,L)}\right) \leq \mathbb{E}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j} d_{TV}\left(G_j, G_j^{(K,L)}\right)\right] = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbb{E}\left[d_{TV}\left(G_j, G_j^{(K,L)}\right)\right]$$
$$\leq N\left[\left(\frac{\beta}{1+\beta}\right)^{L} + \left(\frac{\alpha}{1+\alpha}\right)^{K}\right]$$

and the result follows. □

D. Additional Details about the Nested Slice Sampler

As we mentioned, at each iteration we sample among K^* possible distributional cluster labels and $L^{**} = \max\{L_1^*, \ldots, L_{K^*}^*\}$ possible observational labels. If $\xi_k^D = \pi_k$ and $\xi_{l,k}^O = \omega_{l,k}$, the values are the lowest integers that ensure, respectively, that

$$\sum_{k=1}^{K^*}\pi_k \geq 1 - \min_{j\in\{1,\ldots,J\}} u_j^D \quad \text{and} \quad \sum_{l=1}^{L_k^*}\omega_{l,k} \geq 1 - \min_{i\in\{1,\ldots,n_j\}} u_{i,j}^O \quad \forall k = 1, \ldots, K^*. \tag{30}$$

Instead of relying on the efficient-dependent version, according to Kalli et al. (2011); Hong and Martin (2017), we adopt the following geometric deterministic sequences: $\xi_k^D = (1-\kappa_D)\,\kappa_D^{k-1}$, and $\xi_{l,k}^O = \xi_l^O = (1-\kappa_O)\,\kappa_O^{l-1}$. In this case, it is sufficient to focus only on one observational deterministic sequence, being $\boldsymbol{\xi}_k^O$ the same for every k. Thus, putting $u_{min}^D = \min_j u_j^D$ and $u_{min}^O = \min_{i,j} u_{i,j}^O$, we can compute the two thresholds at each MCMC sweep:

$$K^* = \left\lfloor \frac{\log\left(u_{min}^D\right) - \log\left(1-\kappa_D\right)}{\log\left(\kappa_D\right)} \right\rfloor, \qquad L^* = \left\lfloor \frac{\log\left(u_{min}^O\right) - \log\left(1-\kappa_O\right)}{\log\left(\kappa_O\right)} \right\rfloor.$$

If the precision parameters α and β of the two DPs are assumed stochastic and conjugate Gamma distributions are adopted, the full conditionals can be sampled following the procedure proposed in Walker (2007); Escobar and West (1995): denote with c^* the number of unique values sampled and with n the number of observations ($n = J$ when the Outer DP is considered, otherwise $n = \sum_{j=1}^{J} n_j$). Then the precision parameter of the DP γ ($\gamma = \alpha$ when Outer DP, $\gamma = \beta$ otherwise), for both the DPs, can be sampled

in two stage, introducing another latent variable η: (a) sample $\eta|\gamma, c^* \sim Beta\,(\gamma+1, n)$ and (b) sample a new γ from the mixture

$$\gamma \sim \pi_\eta G(a+k, b-\log(\eta)) + (1-\pi_\eta)\, G\,(a+k-1, b-\log(\eta))$$

where $\pi_\eta = \pi_\eta/\,(1-\pi_\eta) = (a+k-1)/\{n(b-\log(\eta)\}$.

The exploration of the space of cluster membership labels is a delicate task. Differently from the marginal specification, where simulation methods are devised in a way that the resulting Markov Chain explores the space of the partitions as equivalence classes over cluster values, a conditional/stick-breaking specification operates on the space of the explicit cluster labels (Porteous et al., 2006). In this second scenario, it could happen that the chain exploring the cluster membership shows poor mixing, being stuck in one of the local maxima of the posterior. To overcome this issue, the label switching moves described in (Papaspiliopoulos and Roberts, 2008; Hastie et al., 2015) can be added to our setup to improve the mixing.

E. Additional Plots

E.1. Densities of the three scenarios considered in the simulation study

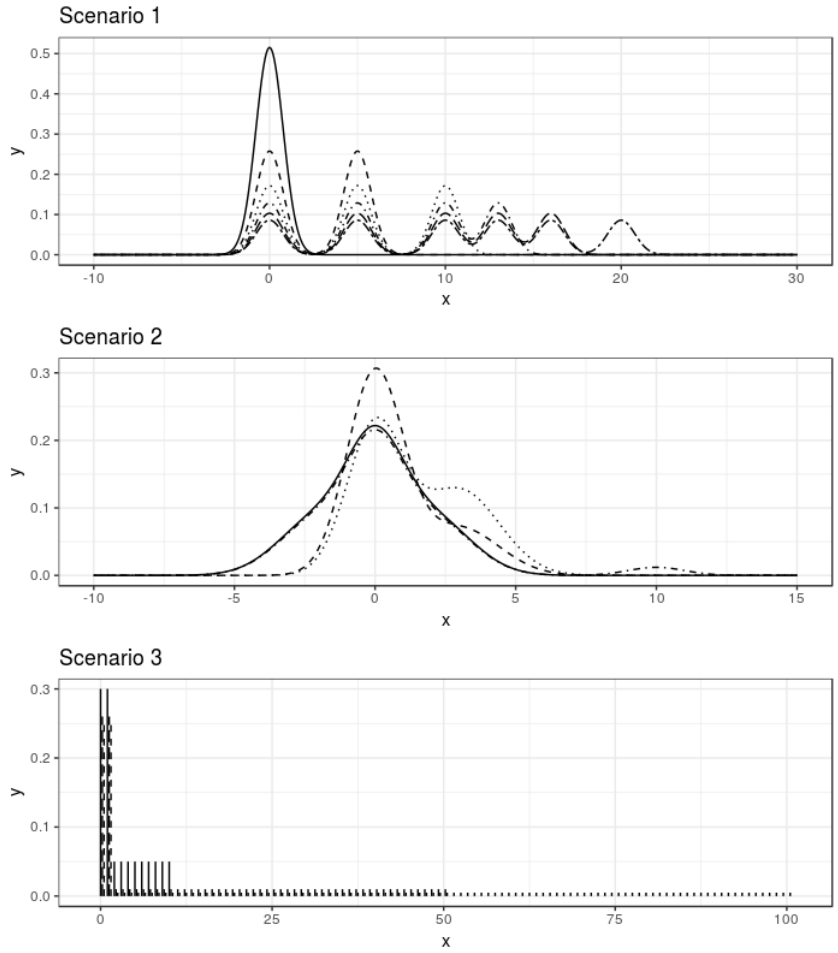


Figure 6: The densities distributions of each unit in three scenarios considered.

E.2. Additional plots for the microbiome application

Visual description of the dataset Percentage of abundance classes per OTU

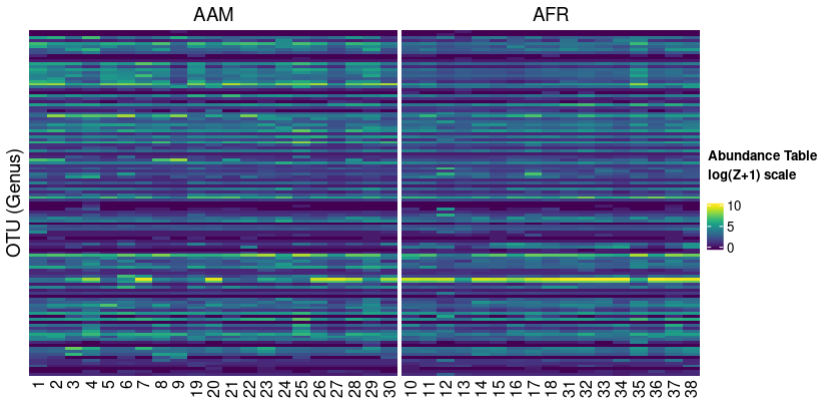


Figure 7: Heatmap of the considered abundance table. The OTUs (at the Genus level) are reported by row, while the columns indicate the subjects, divided by nationality. The count data are transformed as $\log(\boldsymbol{Z}+1)$.

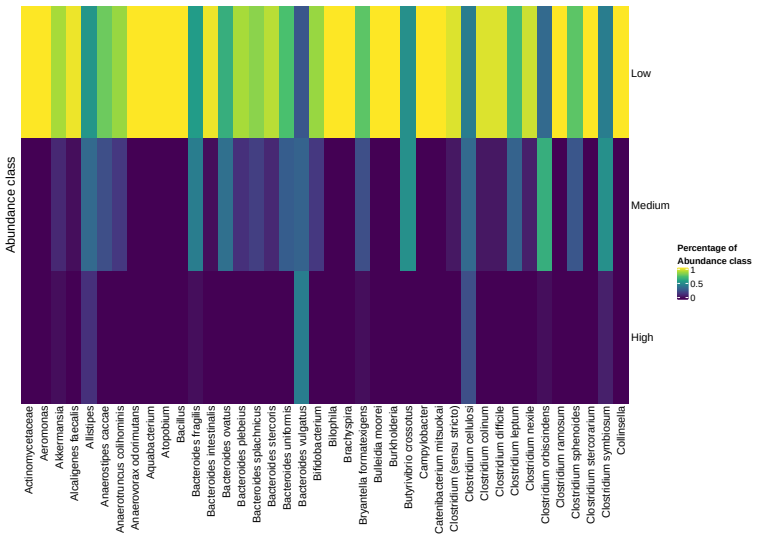


Figure 8: Distribution of the three estimated abundance classes - Part I

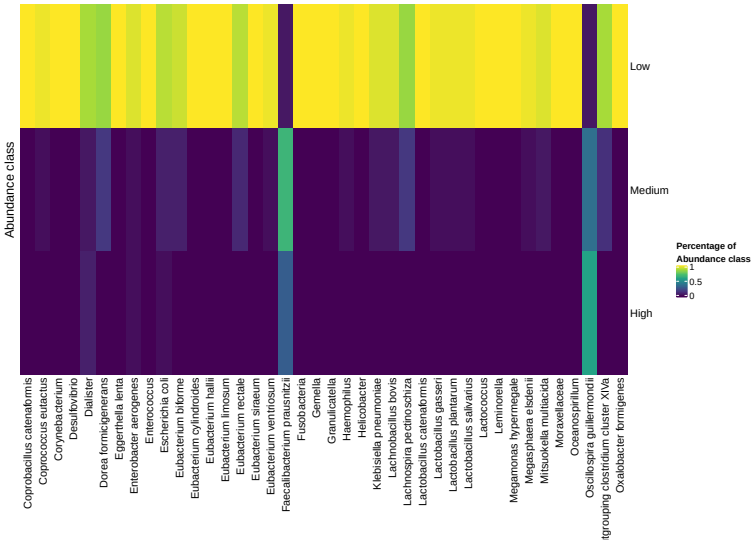


Figure 9: Distribution of the three estimated abundance classes - Part II

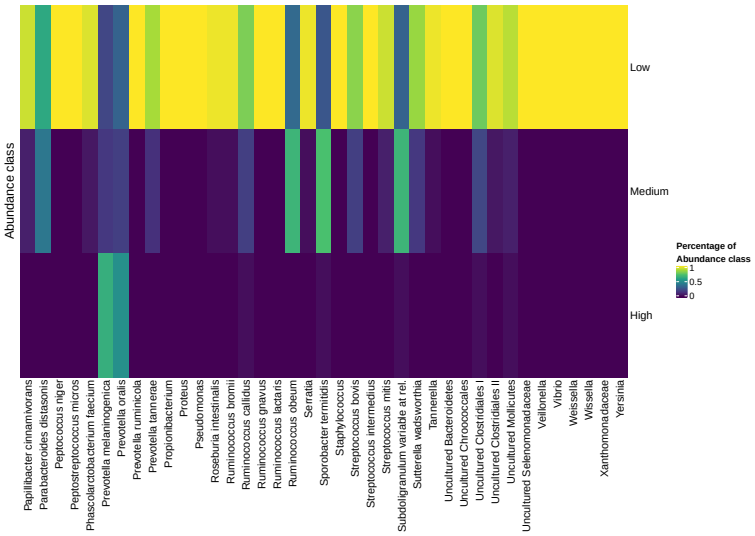


Figure 10: Distribution of the three estimated abundance classes - Part III

Boxplots of the distributional characteristics across DCs

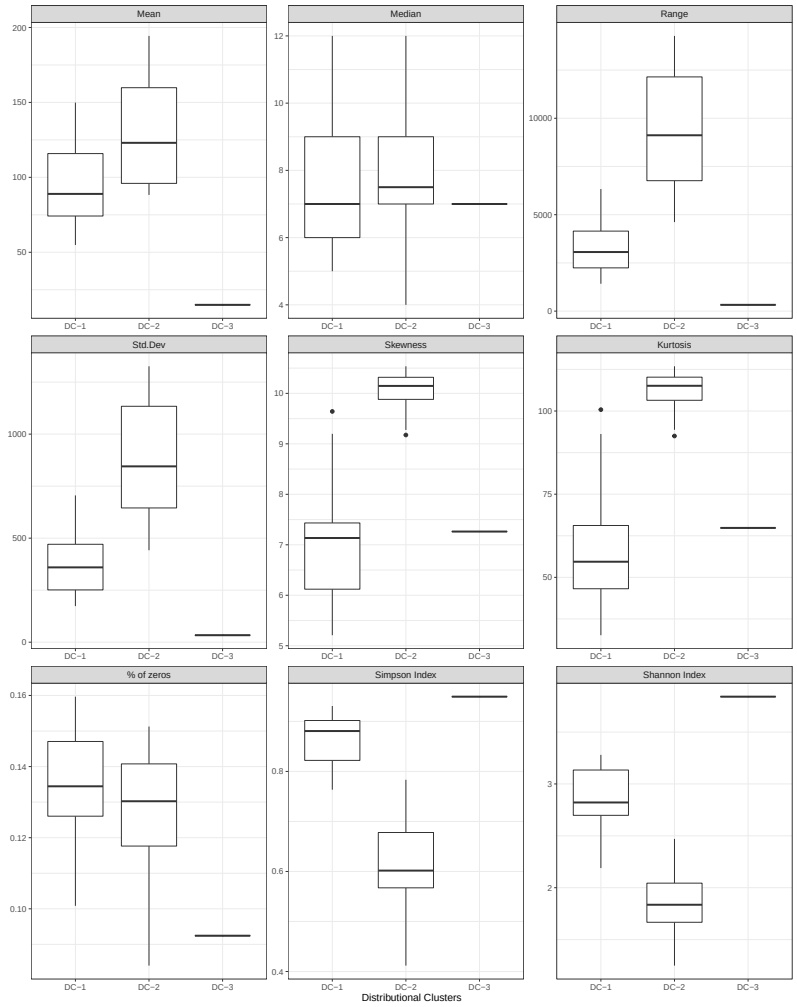


Figure 11: Boxplots representing how mean, median, range, standard deviation, skewness, kurtosis, % of zeros, Shannon index, and Simpson index of each microbiome are distributed across the DCs. The plots highlight different distributional differences among the three DCs.

References

Dipankar Bandyopadhyay and Antonio Canale. Non-parametric spatial models for clustered ordered periodontal data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 65(4):619–640, 2016. ISSN 14679876. doi: 10.1111/rssc.12150.

Anjishnu Banerjee, Jared Murray, and David B Dunson. Bayesian Learning of Joint Distributions of Objects. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 31(Mdm):1–9, 2013. ISSN 15337928. doi: 10.1121/1.4785361.

Andrés F Barrientos, Alejandro Jara, and Fernando A Quintana. On the Support of MacEachern's Dependent Dirichlet Processes and Extensions. *Bayesian Analysis*, 7 (2):277–310, 2012. doi: 10.1214/12-BA709.

Vladimir Batagelj, Nataša Kejžar, and Simona Korenjak-Černe. Clustering of Modal Valued Symbolic Data. *Arxiv Preprint*, 2015. URL http://arxiv.org/abs/1507.06683.

Mario Beraha, Alessandra Guglielmi, and Fernando A. Quintana. The semi-hierarchical dirichlet process and its application to clustering homogeneous distributions. *Arxiv Preprint*, 2020.

Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, third edition, 1995. ISBN 0-471-00710-2. A Wiley-Interscience Publication.

Z I Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79(1):125–148, 2017. ISSN 14679868. doi: 10.1111/rssb.12162.

Federico Camerlenghi, David B. Dunson, Antonio Lijoi, Igor Prünster, and Abel Rodríguez. Latent nested nonparametric priors (with Discussion). *Bayesian Anal-*

ysis, 14:1303–1356, 2019a. doi: 10.1214/19-BA1169. URL http://arxiv.org/abs/1801.05048.

Federico Camerlenghi, Antonio Lijoi, Peter Orbanz, and Igor Prünster. Distribution theory for hierarchical processes. *The Annals of Statistics*, 47:67–92, 2019b.

Antonio Canale and David B Dunson. Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1529–1539, 2011. ISSN 01621459. doi: 10.1198/jasa.2011.tm10552.

Antonio Canale and Igor Prünster. Robustifying Bayesian nonparametric mixtures for count data. *Biometrics*, 73(1):174–184, 2017. ISSN 15410420. doi: 10.1111/biom.12538.

Paul Damien, Jon Wakefield, and Stephen G Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999. doi: 10.1111/1467-9868.00179.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995. ISSN 1537274X. doi: 10.1080/01621459.1995.10476550.

W J Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, 3:87–112, 1972.

T S Ferguson. Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics*, pages 287–303, 1983.

Daniela Graf, Raffaella Di Cagno, Frida Fåk, Harry J. Flint, Margareta Nyman, Maria Saarela, and Bernhard Watzl. Contribution of diet to the composition of the human gut microbiota. *Microbial Ecology in Health & Disease*, 26(0), 2015. ISSN 0891-060X. doi: 10.3402/mehd.v26.26164.

Rebecca Graziani, Michele Guindani, and Peter F Thall. Bayesian nonparametric estimation of targeted agent effects on biomarker change to predict clinical outcome. *Biometrics*, 71(1):188–197, 2015. ISSN 15410420. doi: 10.1111/biom.12250.

David I Hastie, Silvia Liverani, and Sylvia Richardson. Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, 25(5):1023–1037, 2015. ISSN 15731375. doi: 10.1007/s11222-014-9471-3.

Spyridon J. Hatjispyros, Theodoros Nicoleris, and Stephen G. Walker. Random density functions with common atoms and pairwise dependence. *Computational Statistics and Data Analysis*, 101:236–249, 2016. ISSN 01679473. doi: 10.1016/j.csda.2016.03.008.

Liang Hong and Ryan Martin. A Flexible Bayesian Nonparametric Model for Predicting Future Insurance Claims. *North American Actuarial Journal*, 21(2):228–241, 2017. ISSN 10920277. doi: 10.1080/10920277.2016.1247720.

Roger A. Horn, Charles R. Johnson, Roger A. Horn, and Charles R. Johnson. Norms for Vectors and Matrices. *Matrix Analysis*, pages 313–386, 2013. doi: 10.1017/cbo9781139020411.008.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. ISSN 01764268. doi: 10.1007/BF01908075.

Antonio Irpino and Rosanna Verde. Basic statistics for distributional symbolic variables: a new metric-based approach. *Advances in Data Analysis and Classification*, 9(2):143–175, 2015. ISSN 18625355. doi: 10.1007/s11634-014-0176-4.

Hemant Ishwaran and Lancelot F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001. ISSN 1537274X. doi: 10.1198/016214501750332758.

Juan Jovel, Jordan Patterson, Weiwei Wang, Naomi Hotte, Sandra O'Keefe, Troy Mitchel, Troy Perry, Dina Kao, Andrew L Mason, Karen L Madsen, and Gane K.-S.

Wong. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. *Frontiers in Microbiology*, 7:459, 2016. ISSN 1664-302X. doi: 10.3389/fmicb.2016.00459.

Maria Kalli, Jim E Griffin, and Stephen G Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011. ISSN 09603174. doi: 10.1007/s11222-009-9150-y.

Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D. Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in Microbiology*, 8(NOV), 2017. ISSN 1664302X. doi: 10.3389/fmicb.2017.02114.

Jun S Liu. The collapsed gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994. ISSN 1537274X. doi: 10.1080/01621459.1994.10476829.

Albert Y. Lo. On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12(1):351–357, 1984. ISSN 0090-5364. doi: 10.1214/aos/1176346412.

Steven N. MacEachern. Dependent dirichlet processes. *Manuscript*, 2000.

Jialiang Mao, Yuhan Chen, and Li Ma. Bayesian Graphical Compositional Regression for Microbiome Data. *Journal of the American Statistical Association*, 115(530):610–624, 2020. ISSN 1537274X. doi: 10.1080/01621459.2019.1647212. URL http://arxiv.org/abs/1712.04723.

Paul J. McMurdie and Susan Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*, 10(4), 2014. ISSN 15537358. doi: 10.1371/journal.pcbi.1003531.

M Meilúa. Comparing clusterings — an information based distance. *Journal of Multivariate Analysis*, 98:873–895, 2007.

Stephen J.D. O'Keefe, Jia V. Li, Leo Lahti, Junhai Ou, Franck Carbonero, Khaled Mohammed, Joram M. Posma, James Kinross, Elaine Wahl, Elizabeth Ruder, Kishore Vipperla, Vasudevan Naidoo, Lungile Mtshali, Sebastian Tims, Philippe G.B. Puylaert, James Delany, Alyssa Krasinskas, Ann C. Benefiel, Hatem O. Kaseb, Keith Newton, Jeremy K. Nicholson, Willem M. De Vos, H. Rex Gaskins, and Erwin G. Zoetendal. Fat, fibre and cancer risk in African Americans and rural Africans. *Nature Communications*, 6, 2015. ISSN 20411723. doi: 10.1038/ncomms7342.

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2015. ISSN 01628828. doi: 10.1109/TPAMI.2014.2318728.

Omiros Papaspiliopoulos and Gareth O Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008. ISSN 00063444. doi: 10.1093/biomet/asm086.

Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995. ISSN 01788051. doi: 10.1007/BF01213386.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006.

Ian Porteous, Alex Ihler, Padhraic Smyth, and Max Welling. Gibbs sampling for (coupled) infinite mixture models in the stick-breaking representation. *Proceedings of UAI*, 22(4):385–392, 2006. ISSN 15280020. doi: 10.1182/blood-2007-08-106153.

Mdlina Preda, Mircea Ioan Popa, Mara Mdlina Mihai, Teodora Cristiana Oelea, and Alina Maria Holban. Effects of coffee on intestinal microbiota, immunity, and disease. *Caffeinated and Cocoa Based Beverages: Volume 8. The Science of Beverages*, pages 391–421, 2019. doi: 10.1016/B978-0-12-815864-7.00012-X.

Abel Rodriguez and David B. Dunson. Functional clustering in nested designs: Modeling

variability in reproductive epidemiology studies. *Annals of Applied Statistics*, 8(3): 1416–1442, 2014. ISSN 19417330. doi: 10.1214/14-AOAS751.

Abel Rodríguez, David B Dunson, and Alan E Gelfand. The nested dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1144, 2008. ISSN 01621459. doi: 10.1198/016214508000000553.

A J Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4: 639–650, 1994.

C E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. ISSN 01621459. doi: 10.1198/016214506000000302.

Lavanya Sita Tekumalla, Priyanka Agrawal, and Indrajit Bhattacharya. Nested Hierarchical Dirichlet Processes for Multi-Level Non-Parametric Admixture Modeling. *Arxiv Preprint*, 2015. URL http://arxiv.org/abs/1508.06446.

Sara Wade and Zoubin Ghahramani. Bayesian Cluster Analysis: Point estimation and credible balls (with Discussion). *Bayesian Analysis*, 13(2):559–626, 2018. ISSN 19316690. doi: 10.1214/17-BA1073. URL http://arxiv.org/abs/1505.03339.

Stephen G Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation*, 36(1):45–54, 2007. ISSN 03610918. doi: 10.1080/03610910601096262.

R. H. Whittaker. Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, 30(3):279–338, 2006. ISSN 00293970. doi: 10.2307/1943563.

Daiane Aparecida Zuanetti, Peter Müller, Yitan Zhu, Shengjie Yang, and Yuan Ji. Clustering distributions with the marginalized nested Dirichlet process. *Biometrics*, 74(2): 584–594, 2018. ISSN 15410420. doi: 10.1111/biom.12778.