HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

J Am Stat Assoc. 2023 ; 118(541): 405–416. doi:10.1080/01621459.2021.1933499.

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

Francesco Denti[*,a], **Federico Camerlenghi**[**,b], **Michele Guindani**[a], **Antonietta Mira**[c,d]

[a]Department of Statistics, University of California, Irvine, CA

[b]Department of Economics, Management and Statistics, University of Milano - Bicocca, Milan, Italy

[c]Università della Svizzera italiana, Lugano, Switzerland

[d]University of Insubria, Como, Italy

Abstract

The use of large datasets 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.

CONTACT Michele Guindani mguindan@uci.edu Department of Statistics, University of California, Irvine, CA.

[*]During the initial development of this article, Francesco Denti was also supported as a Ph.D. student by University of Milano - Bicocca, Milan, Italy and Universita della Svizzera italiana, Lugano, Switzerland.

[**]Also affiliated to Collegio Carlo Alberto, Piazza V. Arbarello 8, Torino and BIDSA, Bocconi University, Milano, Italy.

Supplementary material

Section A summarizes the terminology used throughout the main paper with a glossary. In addition, a diagram is presented that helps understanding the clustering structure induced by the Common Atom Model (CAM). Section B presents the proofs of the theoretical results in the main paper. Section C presents more details regarding the nested slice sampler algorithm. Section D contains additional plots that are related to the simulation studies and the microbiome application of the main paper. Section E illustrates how CAM performs the density estimation for every unit in the Scenario 1 - Case A of the main article. Section F presents a sensitivity study showing how different prior specifications affect the recovered partitions and estimated densities. Section G compares CAM with some competitor models in terms of distributional clustering performance. Section H compares and discusses the models and implementations in terms of efficiency, measured by simulation time. Section I reports the truncated Gibbs sampler that can be used in place of the nested slice sampler. Section J provides a theoretical evaluation of the errors arising when the truncated algorithm is adopted.

Keywords

Common atoms model; Microbiome abundance analysis; Nested dataset; Nested Dirichlet process; Partially exchangeable data

1. Introduction

The use of large datasets 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 United States 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 subgroups are typically unknown, they need to be estimated from the data.

The nested Dirichlet process (nDP, Rodríguez, Dunson, and Gelfand 2008) and its extensions have been widely employed to identify distributional subgroups in Bayesian nonparametric model-based approaches. For example, Rodriguez and Dunson (2014) proposed a generalization of the nDP for functional data analysis; Graziani, Guindani, and Thall (2015) 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) discussed a marginal nDP for clustering genes related to DNA mismatch repair via the distribution of gene–gene interactions. The nDP leads to a two-layered clustering: first, it allows grouping together similar units (distributional clustering-DC), and then, within each DC, it groups similar observations (observational clustering-OC). However, Camerlenghi et al. (2019a) 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) proposed 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 large datasets 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 article, 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 suitable modification of the nDP, where all the atoms are allowed to be shared across the subgroups, the CAM (i) does not suffer from the degeneracy issue of the nDP, and (ii) allows inference on large datasets with many units. Furthermore, in the nDP, observations can be clustered together only within units that are assigned to the same estimated subgroup. Thus, while the within-subgroup clustering still contributes to a compact representation of the data, observation level inference across subgroups is precluded. Instead, the proposed CAM framework naturally allows unit-level inference and clustering of observations across subgroups, since the structure of the common atoms allows mapping subgroup-specific distributional patterns to a shared support. Compared to 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, Griffin, and Walker 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, with high probability, the two subjects of Figure 1 to two different population subgroups.

The remainder of the article is as follows. In Section 2 we introduce our model for continuous measurements, and 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 units grow in different scenarios. Section 6 summarizes our contributions and discusses some future directions. We defer proofs, additional algorithms, and simulation studies to the supplementary material.

2. CAM 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. More in general, $y_{i,j}$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, may take values in a Polish space $\mathbb{X}$ endowed with the respective Borel σ-field $\mathscr{X}$. Similarly as in the nDP (Rodríguez, Dunson, and Gelfand 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 (DCs). However, Camerlenghi et al. (2019a) showed 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 subgroups. Here, we propose a 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 large datasets. More specifically, let G_j, as $j = 1, \ldots, J$, denote the distribution of the jth experimental unit, so that

$$y_{i,j} \mid G_1, \ldots, G_J \overset{\text{ind.}}{\sim} G_j, \tag{1}$$

independently across $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$. Then, similarly to 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 $\mathscr{X}$, namely

$$G_1, \ldots, G_J \mid Q \overset{\text{iid}}{\sim} Q, \qquad Q = \sum_{k \geq 1} \pi_k \delta_{G_k^*}. \tag{2}$$

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 $(\mathbb{X}, \mathscr{X})$. 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), that is, we consider $v_k \sim \text{Beta}(1, \alpha)$, $k \geq 1$, and then set $\pi_1 = v_1$, and $\pi_k = v_k \prod_{r=1}^{k-1}(1 - v_r)$, $k > 1$, indicated as $\boldsymbol{\pi} = \{\pi_k\}_{k \geq 1} \sim \text{GEM}(\alpha)$. Similarly, we define $\boldsymbol{\omega}_k = \{\omega_{l,k}\}_{l \geq 1} \sim \text{GEM}(\beta)$, where $\omega_{l,k} = r_{l,k} \prod_{s=1}^{l-1}(1 - r_{s,k})$, $l > 1$ and $r_{l,k} \sim \text{Beta}(1, \beta)$ for all $l, k \geq 1$. The weights $\boldsymbol{\pi}_k$'s govern the distributional clustering, whereas the $\omega_{l,k}$'s regulate the observational clustering. In Section A of the supplementary material, we report a diagram that illustrates the clustering structure of the model, along with a summary of the terms we adopted.

Hatjispyros, Nicoleris, and Walker (2016) previously investigated the use of a common atoms structure to model pairwise-dependent Dirichlet processes across m *known* subgroups. 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

any clustering of distributions among the units. Our formulation preserves a two-layered clustering structure, across units (distributional clustering - DC) and between observations (observational clustering - OC). Thus, the proposed CAM is closer in spirit to recently developed hierarchical topic models, for example, the nested HDP by Paisley et al. (2015) and Tekumalla, Agrawal, and Bhattacharya (2015). Those formulations use the HDP as a base measure of an (outer) DP, in symbols $Q \sim \text{DP}(a, \text{DP}(\beta, G_0))$ and $G_0 \sim \text{DP}(\gamma, H)$ and are characterized by three concentration parameters. Their goal is 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, making use of only two concentration parameters. Finally, we mention an alternative semi-parametric model recently developed by Beraha, Guglielmi, and Quintana (2020) that also avoided the degeneracy issue of the nDP and allows for distributional clustering by extending the HDP of Teh et al. (2006). With respect to the work by Beraha, Guglielmi, and Quintana (2020), our proposal is fully nonparametric, yet computationally efficient, and it easily accommodates extensions aimed at clustering 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 subgroups. 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 J samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J$, which will be denoted by $y_1^*, \ldots, y_s^*$, with corresponding frequencies $\boldsymbol{n}_j = (n_{1,j}, \ldots, n_{s,j})$, where $n_{i,j}$ indicates the number of times that the ith 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 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

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

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, that is,

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

with $N = \sum_{j=1}^{J} \boldsymbol{n}_j$, where the expected value in (Equation 4) is taken with respect to the random probabilities G_j's, with distribution specified as in Equation (2). As a result, the probability of the observed allocation is driven by the distribution of the stick–breaking weights $\boldsymbol{\pi}_k$, $\omega_{l,k}$. We point out that the ith 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 follows:

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

where the expectation in (5) is taken with respect to the random probabilities $G_j^* = \sum_{l \geq 1} \omega_{l,j} \delta_{\theta_l}$, with $\{\omega_{l,j}\}_{l \geq 1} \sim \text{GEM}(\beta)$ and $\theta_l \overset{\text{iid}}{\sim} H$.

A closed-form expression of the pEPPF in Equation (5) is not available, due to the presence of the integral over $\mathbb{X}^s$ on the right-hand side. However, 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. Their proofs are reported in the supplementary material, where we also provide an explicit expression for the integral in Equation (5) (see

Equation B.9). The expression highlights how the integral depends on the distribution of the weights $\omega_{l,k}$'s and it is therefore positive.

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

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

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

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

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 derivation of Equations (6) and (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, $\mathcal{X}$ one has

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

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

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

See Section B of the supplementary material for the derivation of Equations (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 $(G_j(A), G_{j'}(B)) = (1+\alpha)^{-1} > 0$, 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) and (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

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

where $p(\cdot/\theta)$ denotes an appropriate parametric continuous kernel density, and $G_j \mid Q \overset{\text{iid}}{\sim} Q$ as in (2). In the rest of the article, we will adopt Gaussian kernels, that is, we assume $p(\cdot/\theta)$ to be Normal and $\theta = \mu, \sigma^2)$ 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, that is, $S_j = k$ and $M_{i,j} = l$ if the observation i in unit j is assigned to the lth observational cluster and the kth DC. Thus, we deal with the following model:

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

where we denoted with $\boldsymbol{\theta} = \{\theta_l\}_{l\geq 1}$. In the following, we consider a Normal-Inverse Gamma distribution for $\theta_l = (\mu_l, \sigma_l^2) \sim \text{NIG}(m_0, \kappa_0, a_0, \beta_0)$, that is, $\mu_l \mid \sigma_l^2 \sim N(m_0, \sigma_l^2/\kappa_0)$ and $\sigma_l^2 \sim IG(a_0, \beta_0)$. Finally, Gamma distributions are adopted for both the precision parameters: $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$ and $\beta \sim \text{Gamma}(a_\beta, b_\beta)$.

2.3. CAM 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 $z_j = (z_{1,j}, \ldots, z_{n_j,j})$ will denote the observed microbiome abundance vector of individual j. We embed model (1) and (2) in the rounded mixture of Gaussian framework of Canale and Dunson (2011). 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

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

for a fixed sequence of thresholds $a_0 < a_1 < a_2 < \ldots < a_\infty$ and for some density function $g(\cdot)$, such that $\int_{a_0}^{a_{+\infty}} g(y)\,dy = 1$. 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. See also Bandyopadhyay and Canale (2016). Canale and Prünster (2017) showed that the rounded mixture framework provides a more flexible and robust specification for the distribution of count data than nonparametric mixtures of Poisson kernels. We propose a novel nested formulation, where $g(\cdot)$ is modeled as a CAM mixture (11). More specifically, conditionally on $y_{i,j}$, we set $z_{i,j} = q \in \mathbb{N}$ if $y_{i,j} \in [a_q, a_{q+1})$, where the distribution of $y_{i,j}$ is specified in Equation (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), that is, by choosing proper upper bounds for the infinite sums that appear in Equation (11). The model representation in Equation (11) can be used for such an algorithm, which we detail in Sections I and J of the supplementary material, where we also provide useful upper bounds to control the resulting truncation error. However, in this article, we develop and employ a novel nested version of the independent slice-efficient algorithm (Walker 2007; Kalli, Griffin, and Walker 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. In our experiments, the use of the slice sampler has also resulted in improved mixing of the chains. The proposed slice sampling scheme can be easily extended to the nDP, and is related to the sampling scheme in Banerjee, Murray, and Dunson (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 $p(y_{i,j}/\theta_l)$ denote a generic density function for the observation $y_{i,j}$, conditionally given θ_l, let $\boldsymbol{\pi} = \{\pi_k\}_{k\geq 1}$ and $\boldsymbol{\omega} = \{\omega_{l,k}\}_{k\geq 1}$ be the two sets of weights, one referred to the distributional clusters, the other one referred to the observational clusters. Then, we can write:

$$f(\mathbf{y}_j \mid \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\pi}) = \sum_{k \geq 1} \pi_k \prod_{i=1}^{n_j} \sum_{l \geq 1} \omega_{l,k} p(y_{i,j} \mid \theta_l).$$

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 follows:

$$f_{\xi^D,\xi^O}\left(\boldsymbol{y}_j, u_j^D, \boldsymbol{u}_j^O \mid \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\pi}\right) = \sum_{k \geq 1} \mathbb{1}_{\{u_j^D < \xi_k^D\}} \frac{\pi_k}{\xi_k^D} \prod_{i=1}^{n_j} \sum_{l \geq 1} \mathbb{1}_{\{u_{i,j}^O < \xi_{l,k}^O\}} \frac{\omega_{l,k}}{\xi_{l,k}^O} p\left(y_{i,j} \mid \theta_l\right). \tag{12}$$

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

$$f_{\xi^D,\xi^O}\left(\boldsymbol{y}, \boldsymbol{u}^D, \boldsymbol{u}^O, \boldsymbol{M}, \boldsymbol{S} \mid \boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\pi}\right) = \prod_{j=1}^{J} \mathbb{1}_{\{u_j^D < \xi_{S_j}^D\}} \frac{\pi_{S_j}}{\xi_{S_j}^D} \prod_{i=1}^{n} \mathbb{1}_{\{u_{i,j}^O < \xi_{M_{i,j},S_j}^O\}} \frac{\omega_{M_{i,j},S_j}}{\xi_{M_{i,j},S_j}^O} p\left(y_{i,j} \mid \theta_{M_{i,j}}\right). \tag{13}$$

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 $p(y_{i,j}/\theta_{M_{i,j}}) = \phi(y_{i,j}/_{M_{i,j}})$ we recover the CAM model in (10). Alternatively, to recover the DCAM model for discrete data z of Section 2.3, it is sufficient to adopt the mixing kernel $p(z_{i,j}/\theta_{M_{i,j}} = \Delta\Phi(a_{z_{i,j}}; \theta_{M_{i,j}}) = \Phi(a_{z+1}; \theta_{M_{i,j}}) - \Phi(a_z; \theta_{M_{i,j}})$, 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 Section H. 1 of the supplementary material). The computation of Steps 5–7 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 Section C of the supplementarymaterial.

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

1. Sample each u_j^D from a uniform distribution $\mathcal{U}\left(0, \xi^D_{S_j}\right)$.
2. Sample each $u_{i,j}^O$ from a uniform distribution $\mathcal{U}\left(0, \xi^O_{M_{i,j},S_j}\right)$.
3. Sample the distributional stick-breaking proportions $\boldsymbol{v}$ independently from $v_k \sim \text{Beta}\,(a_k, b_k)$, 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 observational stick–breaking proportions $\boldsymbol{r_k}$ independently from $r_{l,k} \sim \text{Beta}\left(a_l^k, b_l^k\right)$, 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, Murray, and Dunson (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.
8. Sample α following the procedure proposed in Escobar and West (1995). Finally, sample β from a Gamma distribution with parameters $a_\beta + \sum_{s=1}^{\bar{S}} \bar{M}_s$ and $b_\beta - \sum_{s=1}^{\bar{S}} \sum_{m=1}^{\tilde{M}_s} \log(1 - r_{m,s})$, where $\bar{S} = \max_j S_j$ and $\tilde{M}_s = \max_{i;j:S_j=s} M_{i,j}$.

end

Algorithm 1:
Nested slice-efficient sampler for the CAM

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*, that is, how the distribution of microbial units varies across subgroups of a population. Typically, summary statistics are used to capture characteristics of the species' distributions, for example, 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, Chen, and Ma 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, for example, through the use of 16S rRNA sequences: "practically identical" sequenced tags (≥ 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 ith OTU observed in the jth 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 3 in the supplementary material. The varying sequencing depths also affect the so-called library size, that is, 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, that is,

$$\begin{aligned} y_{i,j} \mid \boldsymbol{M}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 &\sim N\left(\gamma_j \cdot \mu_{M_{i,j}}, \gamma_j^2 \cdot \sigma^2_{M_{i,j}}\right) \\ \Leftrightarrow \frac{\gamma_{i,j}}{\gamma_j} \mid \boldsymbol{M}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 &\sim N\left(\mu_{M_{i,j}}, \sigma^2_{M_{i,j}}\right). \end{aligned} \tag{14}$$

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_0, κ_0, α_0, β_0, a_β, b_β). Following an empirical Bayes rationale, we set m_0 and κ_0 to be equal to the grand mean and the inverse of the overall sample variance. According to Rodríguez, Dunson, and Gelfand (2008), we then set $\beta_0 = 1$ and $\alpha_0 = a_\alpha = b_\alpha = a_\beta = b_\beta = 3$. A MCMC sample of 25,000 iterations was collected after a burn in period of the same length.

DCs.

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 proposed to rely on the minimization of the Variation of Information loss function developed by Meilúa (2007). The results are reported in Figure 2, where we also summarize the main characteristics of these DCs in terms of cardinality, nationality, and gender. It is remarkable how the different subgroups 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, that is, the richness (number) and evenness (the frequencies' similarity) of the different OTUs observed in a sample. Conditionally on the optimal configuration, we compute nine summary statistics for each subject. The DCs capture the different levels of α-diversity of the microbiome subgroups. Indeed, the Shannon Index and the Simpson Index vary substantially across the subgroups. 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 than DC-1. See Figure 7 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 ith highest frequence among the observed OTUs in subject j. We define the cumulative relative frequency (CRF) for subject j as $\mathrm{CRF}_j(i) = \sum_{l=1}^{i} z_{(l),j} / \sum_{i=1}^{n} z_{i,j}$. The left panel of 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 subgroups. 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 almost the 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 9 OCs, representing different intensities of the latent process underlying the counts. For a visual comparison, we report in the right panel of Figure 3 the boxplots of the taxa counts grouped by OC, with the value of the median superimposed. For simplicity, we group the 9 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. 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 $PCM_k(l, g) = \sum_{h=1}^{n_k} \mathbb{1}_{\{AC(g) = AC(l)\}} / n_k$, that is, 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 4. 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. 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* and *Oscillospira guillermondii*. The latter is also highly expressed in DC-2. 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 in protein and fat—typical of western diets—lead to a predominance of *Bacteroides spp*. (Graf et al. 2015; Preda et al. 2019).

5. Simulation Study

We show the performances of the proposed methodology for continuous (CAM) and discrete measurements (DCAM) within a simulation study composed of three scenarios. For every scenario, we generate the units containing the observations from highly overlapping mixture densities. We first 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. The model hyperparameters are set as in the case study, except for the fact that we set $\alpha = \beta = 1$ to facilitate comparisons. We estimate the best partitions by minimizing the Variation of Information given the MCMC output. Each scenario articulates into six configurations, that we now explain:

Scenario 1—CAM. We define six different distributions of the simulated data $\mathbf{Y}_h$, as

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

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, $nj = n_B \cdot j$ for $j = 1, \ldots, 6$ and $n_B \in \{5, 10, 20\}$.

Scenario 2—CAM. Four highly overlapping mixtures are considered

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

The true number of DCs is 4 and there are five OCs, corresponding to the five 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 r of units obtained from each distribution, considering six cases. We denote the number of sampled units in each case by $J_r = 4 \cdot r$, with $r = 1, \ldots, 6$. In other words, the total number of units in each experiment ranges from $J_1 = 4$ (one unit sampled from each distribution) to $J_6 = 24$ (six units sampled from each distribution). 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 $\omega_g = n_g / \sum_{l=1}^{3} n_l$, $g = 1, 2, 3$ denoting the mixture weights. We set $\omega_1 = \omega_2$ by generating $n_1 = 50$ 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, that is, $n_3 \in \{10, 15, 25, 50, 75, 100\}$. We design this simulation study to test how DCAM performs 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 four OCs, where the first set contains "low-expressed" observations (i.e., constituted of zeros and ones). The remaining three groups are obtained partitioning the support into abundance classes corresponding to the intervals [2, 10], (10, 50] and (50, 100].

For each scenario and configuration, we run our model on 30 simulated datasets. In Figures 5 and 6, we assess the goodness of the estimated optimal partition by comparing the adjusted Rand index (ARI - Hubert and Arabie 1985) between the estimated optimal partition and the ground truth. Moreover, we also compare the normalized Frobenius distance (Horn et al. 2013) between the estimated posterior pairwise co-clustering matrices and the true co-clustering 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 $\text{NFD}(A, B) = \sum_{i,j=1}^{p} (a_{ij} - b_{ij})^2 / p^2$.

From the pattern of the boxplots, we appreciate how the model can recover the ground truth, even for small sample sizes. In particular, the NFD between the DC 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 when enough data points are provided. We see how CAM misassigns 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. Overall, the NFD computed for the OC is satisfactorily small across all the configurations of Scenario 2.

Furthermore, we have performed additional numerical studies for comparing the CAM with the nDP and the nested hierarchical DP (Paisley et al. 2015; Tekumalla, Agrawal, and Bhattacharya 2015). The CAM achieves better performances than its competitors in terms of distributional clustering recovery and efficiency. The reader can find the details of those comparisons in Section G and H of the supplementary material. Additionally, we have also investigated the sensitivity to different hyperprior specifications (Section F) and the accuracy of the within-unit density estimation (Section E).

6. Discussion

We have introduced a nested nonparametric model that allows investigating distributional heterogeneity in nested data. The proposed CAM allows a two-layered clustering at the distributional and observational level, similarly to the nDP of Rodríguez, Dunson, and Gelfand (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 CAM specification is appealing and convenient for a variety of reasons: it is simple, allows a more refined description of DCs, 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 dataset is limited by the type and number of clinical and demographic covariates that are available. If additional covariates were available, then they could be used to define more complex dependencies, for example, by constructing dependent random measures with covariate-dependent weights as in MacEachern (2000) (see, also Barrientos, Jara, and Quintana 2012) or to build risk-prediction models. Another interesting extension considers the incorporation of a

time dimension and the study of how DCs vary across time. Finally, to handle datasets with hundreds of thousands of observations, it will be important to explore the use of approximate inference techniques, for example, via Mean Field Variational Bayes algorithms. We leave those directions to future investigation. The code employed for this article is openly available at https://github.com/Fradenti/CommonAtomModel

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

F. Denti was partially funded as a postdoctoral scholar by the NIH grant R01MH115697. Federico Camerlenghi received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Federico Camerlenghi gratefully acknowledges also the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022. Michele Guindani was partially funded by the US National Science Foundation Award SES-1659921. Antonietta Mira was partially funded by the Swiss National Science Foundation grant 163196.

References

Bandyopadhyay D, and Canale A (2016), "Non-Parametric Spatial Models for Clustered Ordered Periodontal Data," Journal of the Royal Statistical Society, Series C, 65, 619–640.

Banerjee A, Murray J, and Dunson DB (2013), "Bayesian Learning of Joint Distributions of Objects," Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, 1–9.

Barrientos AF, Jara A, and Quintana FA (2012), "On the Support of MacEachern's Dependent Dirichlet Processes and Extensions," Bayesian Analysis, 7, 277–310.

Beraha M, Guglielmi A, and Quintana FA (2020), "The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions," arXiv no. arXiv:2005.10287

Camerlenghi F, Dunson DB, Lijoi A, Prünster I, and Rodríguez A (2019a), "Latent Nested Nonparametric Priors" (with discussion), Bayesian Analysis, 14, 1303–1356. [PubMed: 35978607]

Camerlenghi F, Lijoi A, Orbanz P, and Prünster I (2019b), "Distribution Theory for Hierarchical Processes," The Annals of Statistics, 47, 67–92.

Canale A, and Dunson DB (2011), "Bayesian Kernel Mixtures for Counts," Journal of the American Statistical Association, 106, 1529–1539.

Canale A, and Prünster I (2017), "Robustifying Bayesian Nonparametric Mixtures for Count Data," Biometrics, 73, 174–184. [PubMed: 27124115]

Escobar MD, and West M (1995), "Bayesian Density Estimation and Inference Using Mixtures," Journal of the American Statistical Association, 90, 577–588.

Ewens WJ (1972), "The Sampling Theory of Selectively Neutral Alleles," Theoretical Population Biology, 3, 87–112. [PubMed: 4667078]

Ferguson TS (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," Recent Advances in Statistics, 24, 287–303.

Graf D, Di Cagno R, Fåk F, Flint HJ, Nyman M, Saarela M, and Watzl B (2015), "Contribution of Diet to the Composition of the Human Gut Microbiota," Microbial Ecology in Health & Disease, 26, 26164. [PubMed: 25656825]

Graziani R, Guindani M, and Thall PF (2015), "Bayesian Nonparametric Estimation of Targeted Agent Effects on Biomarker Change to Predict Clinical Outcome," Biometrics, 71, 188–197. [PubMed: 25319212]

Hatjispyros SJ, Nicoleris T, and Walker SG (2016), "Random Density Functions With Common Atoms and Pairwise Dependence," Computational Statistics and Data Analysis, 101, 236–249.

Horn RA, Johnson CR, Horn RA, and Johnson CR (2013), "Norms for Vectors and Matrices," in Matrix Analysis, 313–386.

Hubert L, and Arabie P (1985), "Comparing Partitions," Journal of Classification, 2, 193–218.

Ishwaran H, and James LF (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," Journal of the American Statistical Association, 96, 161–173.

Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, and Wong GK-S (2016), "Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics," Frontiers in Microbiology, 7, 459. [PubMed: 27148170]

Kalli M, Griffin JE, and Walker SG (2011), "Slice Sampling Mixture Models," Statistics and Computing, 21, 93–105.

Kaul A, Mandal S, Davidov O, and Peddada SD (2017), "Analysis of Microbiome Data in the Presence of Excess Zeros," Frontiers in Microbiology, 8.

Lo AY (1984), "On a Class of Bayesian Nonparametric Estimates: I.Density Estimates," The Annals of Statistics, 12, 351–357.

MacEachern SN (2000), "Dependent Dirichlet Processes," Technical Report, Department of Statistics, The Ohio State University.

Mao J, Chen Y, and Ma L (2020). Bayesian Graphical Compositional Regression for Microbiome Data," Journal of the American Statistical Association, 115, 610–624.

McMurdie PJ, and Holmes S (2014), "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible," PLoS Computational Biology, 10, e1003531. [PubMed: 24699258]

Meilúa M (2007), "Comparing Clusterings — An Information Based Distance," Journal of Multivariate Analysis, 98, 873–895.

O'Keefe SJ, Li JV, Lahti L, Ou J, Carbonero F, Mohammed K, Posma JM, Kinross J, Wahl E, Ruder E, Vipperla K, Naidoo V, Mtshali L, Tims S, Puylaert PG, Delany J, Krasinskas A, Benefiel AC, Kaseb HO, Newton K, Nicholson JK, De Vos WM, Gaskins HR, and Zoetendal EG (2015), "Fat, Fibre and Cancer Risk in African Americans and Rural Africans," Nature Communications, 6.

Paisley J, Wang C, Blei DM, and Jordan MI (2015), "Nested Hierarchical Dirichlet Processes," IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 256–270. [PubMed: 26353240]

Pitman J (1995), "Exchangeable and Partially Exchangeable Random Partitions," Probability Theory and Related Fields, 102, 145–158.

Porteous I, Ihler A, Smyth P, and Welling M (2006), "Gibbs Sampling for (Coupled) Infinite Mixture Models in the Stick-Breaking Representation," Proceedings of UAI, 22, 385–392.

Preda M, Popa MI, Mihai MM, Oţelea TC, and Holban AM (2019), "Effects of Coffee on Intestinal Microbiota, Immunity, and Disease," in Caffeinated and Cocoa Based Beverages, eds. Alexandru Mihai Grumezescu and Alina Maria Holban, Woodhead Publishing, pp. 391–421. 10.1016/B978-0-12-815864-7.00012-X.

Rodriguez A, and Dunson DB (2014), "Functional Clustering in Nested Designs: Modeling Variability in Reproductive Epidemiology Studies," Annals of Applied Statistics, 8, 1416–1442.

Rodríguez A, Dunson DB, and Gelfand AE (2008), "The Nested Dirichlet Process," Journal of the American Statistical Association, 103, 1131–1144.

Sethuraman AJ (1994), "A Constructive Definition of Dirichlet Priors," Statistica Sinica, 4, 639–650.

Shannon CE (1948), "A Mathematical Theory of Communication," The Bell System Technical Journal, 27, 379–423.

Teh YW, Jordan MI, Beal MJ, and Blei DM (2006), "Hierarchical Dirichlet Processes," Journal of the American Statistical Association, 101, 1566–1581.

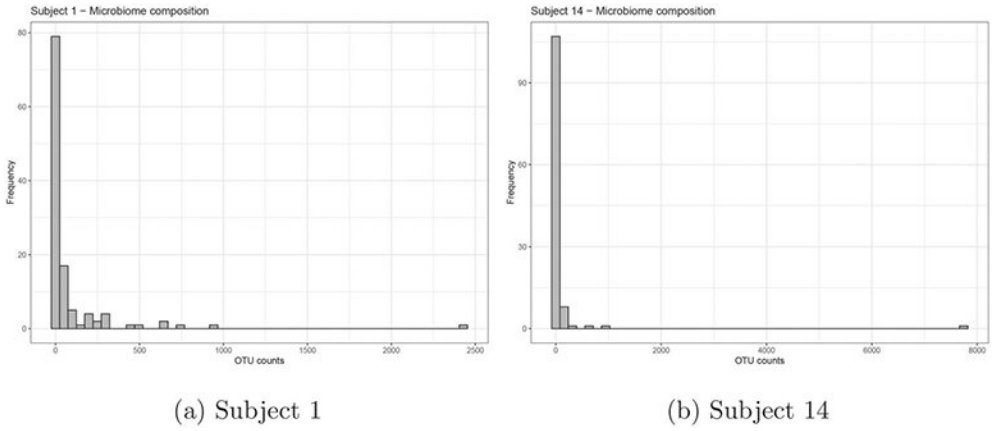
Tekumalla LS, Agrawal P, and Bhattacharya I (2015), "Nested Hierarchical Dirichlet Processes for Multi-Level Non-Parametric Admixture Modeling. Arxiv no. arXiv no. 1508.06446

Wade S, and Ghahramani Z (2018), "Bayesian Cluster Analysis: Point Estimation and Credible Balls" (with discussion), Bayesian Analysis, 13, 559–626.

Walker SG (2007), "Sampling the Dirichlet Mixture Model With Slices. Communications in Statistics: Simulation and Computation, 36, 45–54.

Whittaker RH (2006). Vegetation ofthe Siskiyou Mountains, Oregon and California. Ecological Monographs, 30, 279–338.

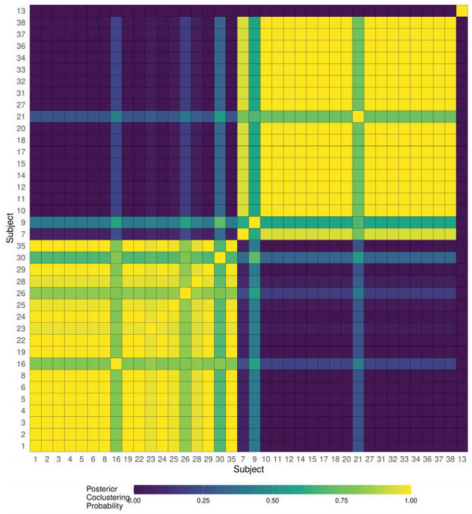
Zuanetti DA, Muller P, Zhu Y, Yang S, and Ji Y (2018), "Clustering Distributions with the Marginalized Nested Dirichlet Process," Biometrics, 74, 584–594. [PubMed: 28960246]



(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.



Cluster	DC-1	DC-2	DC-3
Cardinality	18	19	1
Africans	2	14	1
Americans	16	5	0
Female	11	6	0
Male	7	13	1

Figure 2.

Left: pairwise posterior probability matrix of co-clustering 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.

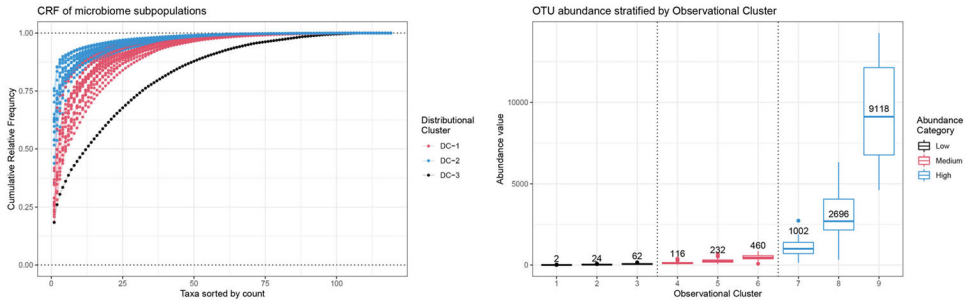


Figure 3.

Left panel: 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. Right panel: 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.

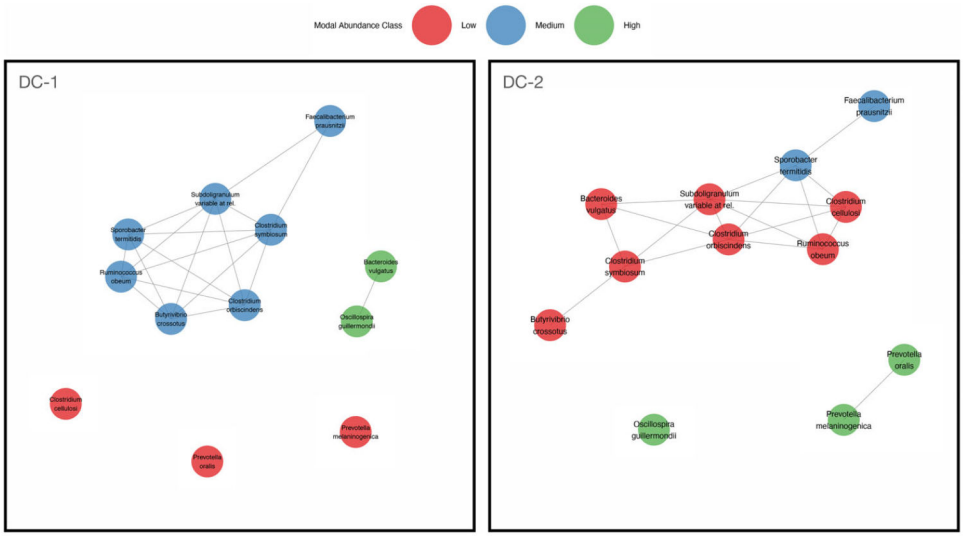


Figure 4.

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

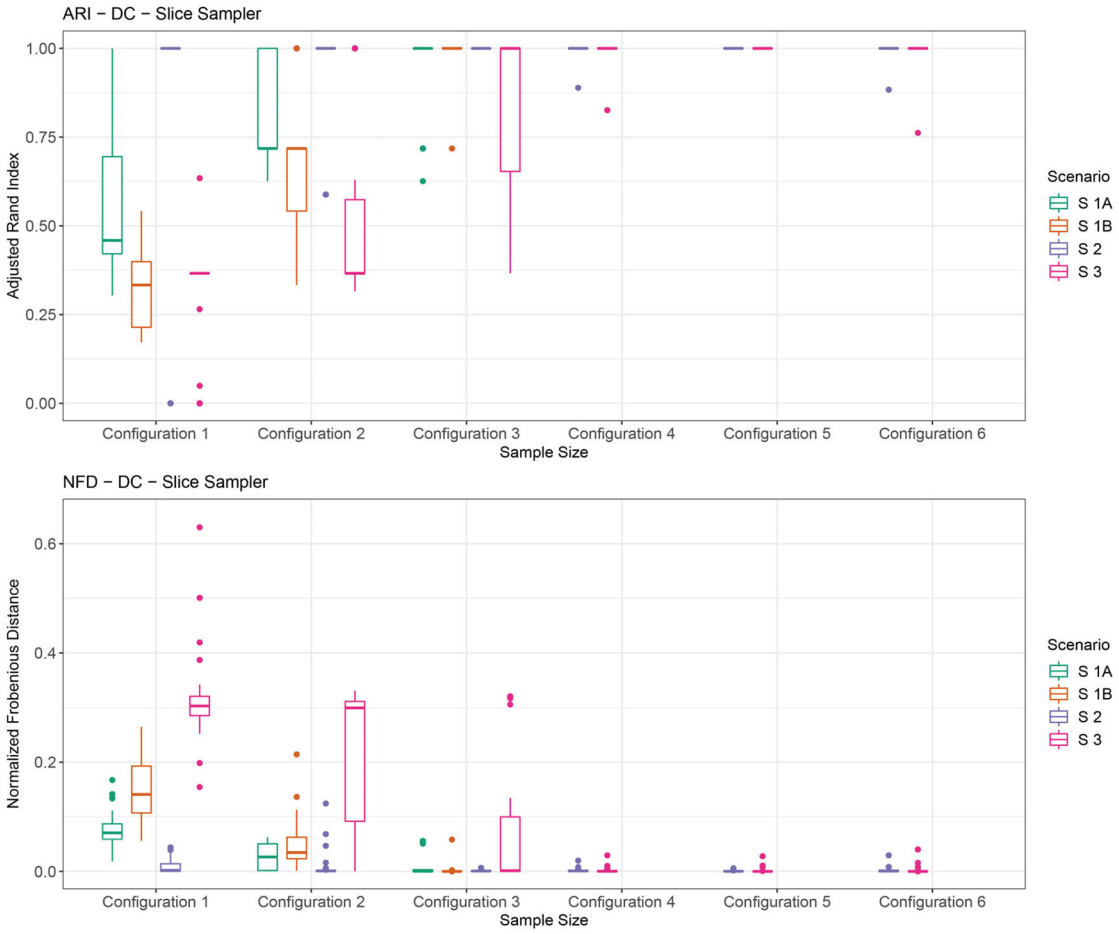


Figure 5.

Distributional clustering (DC) performance for CAM and DCAM across 30 simulations, evaluated according the number of the adjusted Rand index (ARI) and the normalized Frobenius distance (NFD) between posterior pairwise co-clustering matrices.

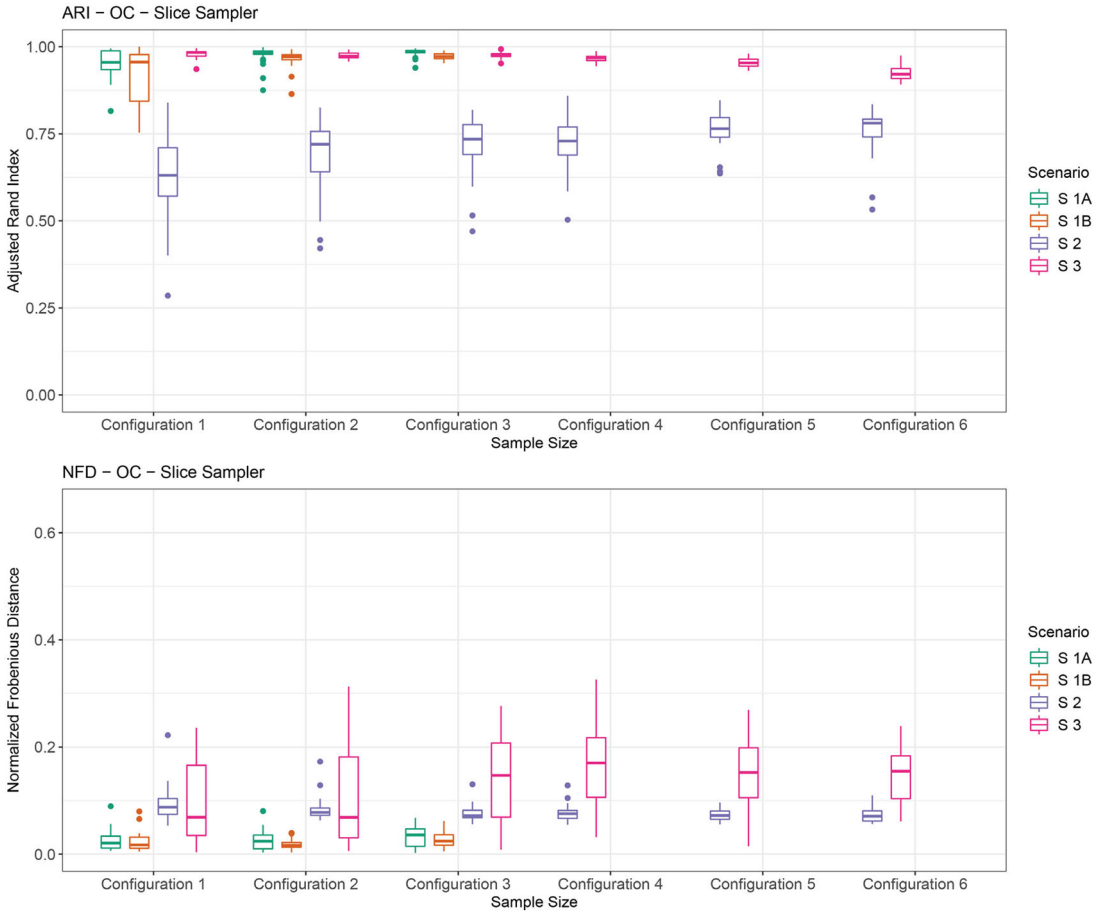


Figure 6.

Observational clustering (OC) performance for CAM and DCAM across 30 simulations, evaluated according the number of the adjusted Rand index (ARI) and the normalized Frobenius distance (NFD) between posterior pairwise co-clustering matrices.