Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–34

A Finite-Infinite Shared Atoms Nested Model for the Bayesian Analysis of Large Grouped Data Sets

Laura D'Angelo* and Francesco Denti†

Abstract. The use of hierarchical mixture priors with shared atoms has recently flourished in the Bayesian literature for partially exchangeable data. Leveraging on nested levels of mixtures, these models allow the estimation of a two-layered data partition: across groups and across observations. This paper discusses and compares the properties of such modeling strategies when the mixing weights are assigned either a finite-dimensional Dirichlet distribution or a Dirichlet process prior. Based on these considerations, we introduce a novel hierarchical nonparametric prior based on a finite set of shared atoms, a specification that enhances the flexibility of the induced random measures and the availability of fast posterior inference. To support these findings, we analytically derive the induced prior correlation structure and partially exchangeable partition probability function. Additionally, we develop a novel mean-field variational algorithm for posterior inference to boost the applicability of our nested model to large multivariate data. We then assess and compare the performance of the different shared-atom specifications via simulation. We also show that our variational proposal is highly scalable and that the accuracy of the posterior density estimate and the estimated partition is comparable with state-of-the-art Gibbs sampler algorithms. Finally, we apply our model to a real dataset of Spotify's song features, simultaneously segmenting artists and songs with similar characteristics.

Keywords: Dirichlet process, finite mixture, partially exchangeable data, multivariate data, variational Bayes.

1 Introduction

Inference with grouped data is a well-established statistical challenge and a common scenario in real-data analyses. Hierarchical models provide a standard framework for working with this type of data and are particularly well-suited for a Bayesian treatment. They allow having a large enough number of parameters to fit the data well without incurring the risk of overfitting; they enable borrowing information between samples while admitting the presence of heterogeneity. These strengths are fundamental in several application areas and lines of research. In multicenter clinical trials, for example, it is relevant to study the cross-group variability for assessing the performance of a specific treatment (Gray, 1994); in the early stages of an epidemic, integrating data from different outbreaks allows timely interventions (Lee et al., 2020); innovative cancer radiomics

*Department of Economics, Management and Statistics; University of Milan-Bicocca; Milan, Italy, laura.dangelo@unimib.it

†Department of Statistics; University of Padua, Italy, francesco.denti@unipd.it

 https://doi.org/10.1214/24-BA1458

techniques use computational approaches to improve diagnosis of tumors by borrowing information across subjects (Li et al., 2021); causal inference models can use data from randomized trials to study related observational scenarios (Wang and Rosner, 2019).

Meanwhile, nonparametric Bayesian methods have become increasingly popular thanks to their non-restrictive assumptions on the parametric form of the data and general algorithms for posterior inference. In the nonparametric context, the need to develop hierarchical modeling strategies has led to several formulations of layered prior distributions based on multi-level Dirichlet processes (DP) specifications. The hierarchical Dirichlet process (HDP) of Teh et al. (2006) provides a remarkable contribution in this direction. The HDP formulation relies on modeling groups of observations using distinct DPs with a common base measure, which, in turn, is itself a realization from another DP. Since the draws from this DP are almost surely discrete, all group-specific distributions are based on a shared set of atoms. In the framework of model-based clustering, this formulation leads to the nice property of allowing cross-group clusters of observations (observational clusters, OC), favoring the interpretation of the latent structure of the data. A different perspective on multicenter studies was considered by Rodríguez et al. (2008) with the nested Dirichlet process (nDP). Instead of focusing on partitioning observations, they considered the problem of investigating clusters of groups, that is, samples with similar distributional characteristics (distributional clusters, DC). This result is achieved by replacing the random atoms of a DP with random probability measures, which are themselves sampled from a DP, thus specifying a mixture over *distributional atoms*. The discreteness of this mixing measure at the distributional level leads to a positive probability of modeling two distinct samples with the same distribution. Additionally, the discreteness of the distributional atoms also results in the clustering of observations within the same distributional cluster. Models based on the nDP have been widely employed in various contexts: see, for example, Graziani et al. (2015); Rodríguez and Dunson (2014); Zuanetti et al. (2018). However, despite the attractiveness of this construction, Camerlenghi et al. (2019) revealed a critical drawback, which occurs whenever an observational atom is shared by two distinct groups. In this case, the nDP imposes the full exchangeability of the two samples, forcing their complete homogeneity (this behavior is also referred to as *degeneracy*). In response to this drawback, several authors have proposed alternatives to the nDP that admit the presence of cross-group observational clusters and avoid the degeneracy issue: notable examples are the semi-hierarchical DP of Beraha et al. (2021) and the hidden HDP (HHDP) of Lijoi et al. (2023a). In particular, the latter work investigates the theoretical properties of *admixture models* (see, e.g., Agrawal et al., 2013; Balocchi et al., 2022) where a nDP structure is placed over the realizations of an HDP, obtaining a discrete distribution over the space of random measures with shared atoms. Another formulation that conveys both the cross-DC observational clustering of the HDP and the distributional clustering of the nDP has been recently proposed by Denti et al. (2023) with the common atoms model (CAM). Their specification closely resembles that of the nDP, but has a crucial difference: the observational atoms are assumed to be the same across all distributional atoms. On the one hand, the shared atoms are essential for avoiding degeneracy while maintaining a clustering between observations in different groups. On the other hand, they pose a constraint on the random measures constituting the distributional atoms, which, by construction, are bound to have a correlation above 0.5.

Therefore, this approach could lead to biased posterior inference in scenarios where, for example, the distributional clusters are well-separated.

Somehow surprisingly, the literature on the parametric counterpart of nested mixture models is very limited. A finite version of the CAM (fCAM) was proposed by D'Angelo et al. (2023) for analyzing neuroimaging data. The proposed model made use of the telescoping sampler of Frühwirth-Schnatter et al. (2021) to define nested levels of finite mixtures with a random number of components. Such a specification exhibited promising results both on simulated data as well as in real settings, and indeed their work suggested even improved performances compared to the nonparametric formulation. However, despite the empirical results, they did not provide theoretical justifications on the rationale of this behavior. The modeling framework we introduce and analyze also provides a formal justification for their findings. For the interested reader, Section A.1 of the Supplementary Material (D'Angelo and Denti, 2024) contains a brief discussion in which we outline and compare the nDP, the CAM, and the fCAM.

In view of these considerations, we investigate nested priors based on a finite set of shared observational atoms. In particular, we explore a new class of models that we call the Shared Atoms Nested (SAN) priors. Our formulation adopts the flexible two-level structure of the CAM, but it departs from it for the use of a set of observational atoms of finite dimension. This modification preserves the simple structure of the CAM and, at the same time, has a considerably positive effect on the prior properties. The proposed structure can be combined with different specifications of the distributional weights to enhance its flexibility. Moreover, it allows the derivation of a coordinate ascent variational inference (CAVI) algorithm to improve scalability. Thanks to the availability of this fast and efficient algorithm for posterior inference, we can indeed develop this modeling framework in the context of large multivariate data to widen the applicability of nested Bayesian nonparametric models to datasets with thousands of observations and hundreds of groups.

The paper is organized as follows. In Section 2, we introduce the general setting of SAN models, present several prior properties, and compare them with other nested priors. Section 3 details the CAVI algorithm. In Section 4, we perform a simulation study to compare the adequacy of the proposed prior for clustering and density estimation with other state-of-the-art models. In particular, we analyze how the proposed framework scales with multivariate data and compare the performances and accuracy of the proposed variational algorithm against a standard Gibbs sampler approach. Finally, in Section 5, we apply our model to a dataset provided by Spotify that contains numerical features of thousands of songs authored by hundreds of artists. Our model is able to identify a sensible two-level clustering of similar artists and songs, which can be used to provide listening suggestions.

2 Shared atoms nested priors

Consider a nested design, where the data $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J)$ are divided into J different groups. Within each group $j = 1, \ldots, J$, N_j observations are measured: $\boldsymbol{y}_j = (\boldsymbol{y}_{1,j}, \ldots, \boldsymbol{y}_{N_j,j})$, $\boldsymbol{y}_{i,j} \in \mathcal{Y}$, with $\mathcal{Y}$ the common support of dimension $d \geq 1$. We as-

sume a partially exchangeable framework, in the sense that observations are assumed exchangeable *within* each sample, and we adopt the following mixture model specification:

$$\boldsymbol{y}_{1,j}, \ldots, \boldsymbol{y}_{N_j,j} \mid f_j \overset{ind.}{\sim} f_j, \qquad f_j(\cdot) = \int_{\Theta} p(\cdot \mid \theta) G_j(d\theta), \tag{1}$$

for $j = 1, \ldots, J$, where $p(\cdot \mid \theta)$ denotes a parametric kernel on $\mathcal{Y} \times \Theta$ and G_j is a group-specific mixing measure. Let Θ be the space of the mixing parameter θ, endowed with the respective Borel σ-field $\mathcal{X}$. We assume that the mixing measures G_j's are sampled from an almost surely discrete distribution Q, defined over the space of probability distributions on $\mathcal{X}$. In particular, to induce a two-layer clustering structure, we assume Q to have the following form, as originally proposed by Rodríguez et al. (2008) with the nDP:

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

Furthermore, we assume that each distributional atom G_k^* is given by

$$G_k^* = \sum_{l=1}^{L} \omega_{l,k} \delta_{\theta_l^*}, \tag{3}$$

with $\{\theta_l^*\}_{l=1}^L$ randomly sampled from a non-atomic base measure H defined on $(\Theta, \mathcal{X})$. In the following, we will refer to the sequence $\{\theta_l^*\}_{l=1}^L$ as *observational* atoms, and to $\{G_k^*\}_{k=1}^K$ as *distributional* atoms. The parameters K and L play a crucial role in the prior distributions we introduce. They indicate the number of mixture components, with $K \in \mathbb{N} \cup \{\infty\}$ and $L \in \mathbb{N} \cup \{\infty\}$. It is crucial to note that according to (3), the atoms θ_l^* are shared across all distributions G_k^*, following the idea introduced by Denti et al. (2023). This shared set of atoms is essential for allowing cross-DC observational clusters.

The distributional properties of such priors are intrinsically defined by the law of the observational weights $\boldsymbol{\omega}_k = \{\omega_{l,k}\}_{l=1}^L$ for $k = 1, \ldots, K$, and the distributional weights $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$. The CAM resorted to a stick-breaking construction for the random weights of Q and G_k^*'s. Specifically, they considered $K = L = \infty$, $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ and $\boldsymbol{\omega}_k \sim \text{GEM}(\beta)$ for $k \geq 1$, i.e., $\pi_k = v_k \prod_{j=1}^{k-1}(1 - v_j)$ and $\omega_{l,k} = u_{l,k} \prod_{r=1}^{l-1}(1 - u_{r,k})$ with $v_k \sim \text{Beta}(1, \alpha)$ and $u_{l,k} \sim \text{Beta}(1, \beta)$ for all $k \geq 1$, $l \geq 1$ (Sethuraman, 1994). This construction, however, leads to a well-known stochastically decreasing ordering of the weights: hence the atoms that appear earlier in the sequence will have larger associated mass, *for all distributions*. This feature stands at the basis of the rigid correlation structure of the CAM mentioned in the Introduction. We propose to modify the distribution of the mixing weights to "break" the correspondence between ordered weights and atoms. In this sense, we distinguish our proposals based on *shared* atoms, where no (implicit) order is assumed, from the model of Denti et al. (2023) based on *common* atoms, where the θ_l^*'s have stochastically decreasing weight in all distributional atoms. Specifically, we place a symmetric Dirichlet distribution on the observational weights instead of a Dirichlet process. Although it could appear as a simplification, this choice has a notable positive impact on the model's performance. Symmetric Dirichlet distributions indeed imply that, *a priori*, all atoms are equally likely to be resampled

across random measures. This characteristic grants a more flexible correlation structure, improving both prior and posterior properties.

The finite-infinite Shared Atoms Nested (fiSAN) prior is defined by Equations (2)-(3) and assumes $K = \infty$, a finite $L \in \mathbb{N}$, and that the mixing weights are assigned the following prior distributions

$$\begin{aligned} \boldsymbol{\pi} = \{\pi_k\}_{k=1}^{\infty} &\sim \text{GEM}(\alpha), \\ \boldsymbol{\omega}_k = (\omega_{1,k}, \ldots, \omega_{L,k}) &\sim \text{Dirichlet}_L(b_k, \ldots, b_k). \end{aligned} \tag{4}$$

For simplicity, from here on, we assume $b_k \equiv b$ for $k \geq 1$. This prior is based on a "hybrid" formulation, where a Dirichlet process drives the distributional partition, and finite Dirichlet distributions control the observational one. The crucial aspect of this prior, which will be extensively analyzed in the following sections, is that the finiteness of the Dirichlet distribution allows spreading the mass homogeneously over all the atoms θ_l^*, without favoring a small subset of them. Indeed, we will show that this prior has several interesting properties and that this mixed approach can have advantages over a purely nonparametric specification.

Since many key properties of the model are driven by the prior on $\boldsymbol{\omega}_k$, one could specify different laws for the distributional weights. In particular, we will also study the fully finite case, i.e., $K \in \mathbb{N}$ and $L \in \mathbb{N}$ finite, and

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}_K(a, \ldots, a), \tag{5}$$

and we will call this specification a finite SAN (fSAN) prior.

The use of finite mixtures has long caused concerns about the choice of the dimension of the Dirichlet distribution. However, in the recent literature, several effective strategies have been proposed to overcome this issue. These works leverage the distinction between *components* and *clusters*, with the latter indicating the "filled" mixture components used to generate the data (Richardson and Green, 1997; Nobile, 2004). Rousseau and Mengersen (2011) demonstrated that in the context of sparse "overfitted" mixtures (i.e., mixtures where the number of components is deliberately set larger than the number of clusters present in the data), the posterior distribution asymptotically concentrates on the true mixture if appropriate concentration parameters are specified. In other words, the posterior will empty the extra components, allowing the model to automatically estimate the dimension of the partition of the observed sample. The practical effectiveness of this strategy has also been showcased by Malsiner-Walli et al. (2016). Another strategy that has received renewed interest is the specification of a prior over the dimension of the Dirichlet distribution: this leads to a mixture with a finite, random number of components. This class comprises several notable instances. Using a symmetric Dirichlet distribution with random dimension and a fixed parameter, as discussed in Miller and Harrison (2018), results in a special case of Gibbs-type prior (Gnedin and Pitman, 2006; De Blasi et al., 2013, 2015). An even more flexible approach was studied by Frühwirth-Schnatter et al. (2021), where they allowed the Dirichlet parameter to vary as a function of the dimension. In general, one could use the preferred approach to fix or estimate L and K in (4) and (5) without leading to restrictive assumptions.

The model of D'Angelo et al. (2023) is precisely an instance of this kind of approach, where both K and L are finite and random, following the mixture of finite mixtures framework of Frühwirth-Schnatter et al. (2021). However, the study of the properties induced by different prior specifications of $p(K)$ and $p(L)$ is beyond the scope of this paper. Indeed, in the following, we will study the properties of finite nested mixture models for fixed L and K (or conditionally on their values).

2.1 Correlation structure

The discreteness of the nested priors on the two levels, together with the commonality of the atoms, allows for the presence of observational and distributional ties. To compare the different priors, we analyze the dependence between pairs of distributions and observations. The following properties hold for fixed concentration parameters, all proofs are deferred to Section B of the Supplementary Material. The first important quantity to investigate is the probability of observational and distributional co-clustering. Consider two distributions G_j and $G_{j'}$ defined on $(\Theta, \mathcal{X})$, with $G_j, G_{j'} \mid Q \sim Q$ and Q defined as in (2). Moreover, consider two observations $\theta_{i,j} \mid G_j \sim G_j$ and $\theta_{i',j'} \mid G_{j'} \sim G_{j'}$.

Under the finite-infinite SAN prior, the co-clustering probabilities are

$$\mathbb{P}\left[G_j = G_{j'}\right] = \frac{1}{1+\alpha} \quad \text{and} \quad \mathbb{P}\left[\theta_{i,j} = \theta_{i',j'}\right] = \frac{L + \alpha + L(b + \alpha b)}{L(\alpha+1)(Lb+1)};$$

while for the finite SAN prior, it is immediate to show that

$$\mathbb{P}\left[G_j = G_{j'}\right] = \frac{1+a}{1+Ka} \quad \text{and} \quad \mathbb{P}\left[\theta_{i,j} = \theta_{i',j'}\right] = \frac{a(L+K-1) + L(b + Kab + 1)}{L(Ka+1)(Lb+1)}.$$

These expressions show that the probability of observational co-clustering between two observations belonging to different groups is non-null. Moreover, noticing that $\mathbb{P}\left[\theta_{i,j} = \theta_{i',j'} \mid G_j \neq G_{j'}\right] > 0$, we have a confirmation of the ability of these priors to convey a cross-DC observational clustering. Finally, the distributional co-clustering probability of the fiSAN is equivalent to that of the CAM and the nDP.

The following proposition provides the expression for the correlation between random measures induced by the SAN priors.

Proposition 2.1. *Consider two distributions $G_j, G_{j'} \mid Q \sim Q$ defined on $(\Theta, \mathcal{X})$, and $A \in \mathcal{X}$ a Borel set. Under the finite-infinite SAN prior, the correlation between two random measures evaluated on the same set A is:*

$$\rho_{j,j'}^{(fiSAN)} := Corr(G_j(A), G_{j'}(A)) = 1 - \frac{\alpha(L-1)}{L(\alpha+1)(b+1)}. \tag{6}$$

Under the finite SAN prior, the correlation is:

$$\rho_{j,j'}^{(fSAN)} := Corr(G_j(A), G_{j'}(A)) = 1 - \frac{a(K-1)(L-1)}{L(Ka+1)(b+1)}. \tag{7}$$

It is interesting to compare the prior correlation structure of SAN priors with the other nested nonparametric models; in particular, we consider the HHDP, the nDP, and

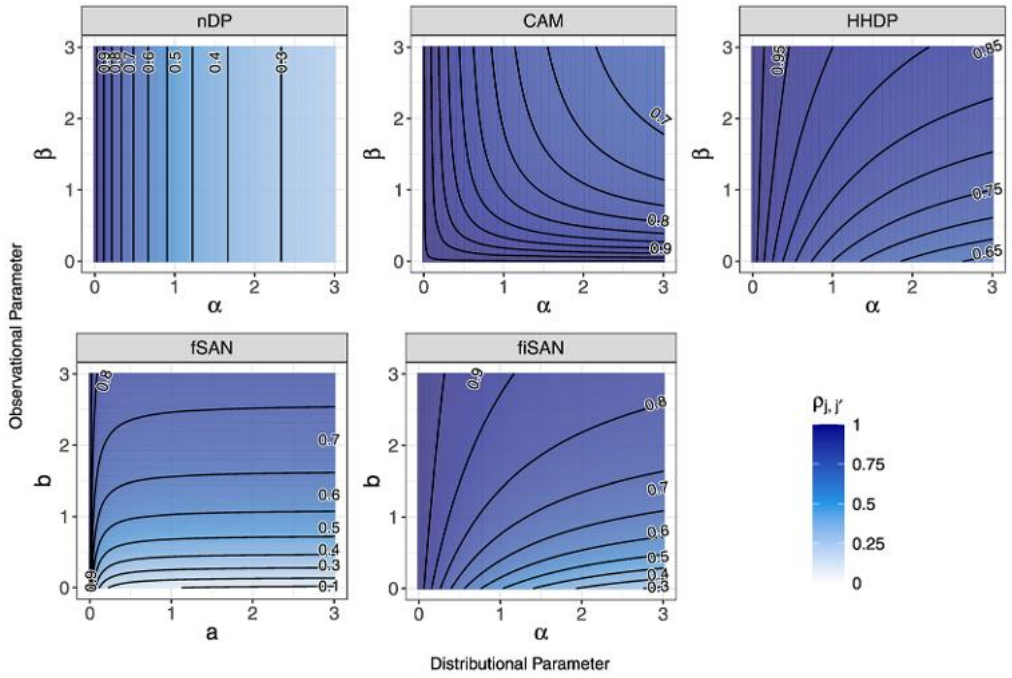

Figure 1: Correlation of different nested models for varying observational and distributional concentration parameters. For the fiSAN we fixed $L = 30$, and for the fSAN $L = K = 30$. For the HHDP, we set $\beta_0 = 1$.

the CAM (the correlation of these priors are derived in Lijoi et al., 2023a; Rodríguez et al., 2008; and Denti et al., 2023, respectively). For all these fully nonparametric models, we consider $\boldsymbol{\omega}_k \sim \text{GEM}(\beta)$ for all k at the observational level, and $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ at the distributional level, with both α and β fixed. Finally, under the HHDP, the G_k^*'s are realizations from an HDP, i.e., $G_k^* \sim DP(\beta, G_0)$ and $G_0 \sim DP(\beta_0, H)$. The expressions for the correlations of these models are

$$\begin{aligned} \rho_{j,j'}^{(nDP)} &= \frac{1}{1+\alpha}, \qquad \rho_{j,j'}^{(CAM)} = 1 - \frac{\alpha}{1+\alpha} \cdot \frac{\beta}{1+2\beta}, \\ \rho_{j,j'}^{(HHDP)} &= 1 - \frac{\alpha\beta_0}{(\alpha+1)\,(\beta+\beta_0+1)}. \end{aligned} \tag{8}$$

Figure 1 shows the correlation between two random probability measures obtained for varying values of the concentration parameters in the different nested models we are considering. Under the nDP formulation, the correlation does not depend on β since the model assumes independence between observational atoms of measures G_j and $G_{j'}$ assigned to different distributional clusters. On the contrary, for both the HHDP and the CAM, α and β have a joint impact on the induced correlation. There are several analogies between the two models: they are both based on nested levels of Dirichlet processes

combined with a shared set of atoms. This similarity of the fundamental structure is reflected in the implied correlation: indeed, $\rho^{(CAM)}_{j,j\prime}$ and $\rho^{(HHDP)}_{j,j\prime}$ are considerably high for standard default values of the concentration parameters (e.g., $\alpha = \beta = \beta_0 = 1$). The cause of this characteristic is rooted in the interaction between the weights and the (implicitly ordered) set of observational atoms. For the CAM, the stick-breaking weights explicitly assume the stochastic order of the atoms to be the same across the random measures. Indeed, it is not even possible to act on the parameters to relax this constrained dependence, and $\rho^{(CAM)}_{j,j\prime} > 0.5$ by construction. This implicit assumption is more subtle in the HHDP: their formulation involves a "resampling" of the weights through an additional layer, which, in principle, allows eluding the issue. However, the atoms are initially sampled from a DP, which already suggests a common importance across distributional atoms. To leverage this intermediate layer and relax the induced correlation, one should set β_0 to very large values; however, this aspect is often overlooked in practice. At the other extreme, it is remarkable that the limiting case of $\beta_0 \to \infty$ would make the HHDP revert to the nDP. Finally, we highlight how the limiting cases $\beta \to 0$ for the CAM and $\beta_0 \to 0$ for the HHDP are particularly troublesome, as they would force all the distributional atoms to place all the probability mass on the same observational atom (in Section A.2 of the Supplementary Material we report additional graphs that investigate this issue). We did not include the HDP in this discussion since it is not based on nested levels of DPs and therefore is not directly comparable.

SAN priors show instead a very flexible dependence structure. Moreover, unlike other models, there are no parameter combinations that lead to pathological behaviors. As we can see from Equations (6) and (7), both $\rho^{(\text{fiSAN})}_{j,j\prime}$ and $\rho^{(\text{fSAN})}_{j,j\prime}$ lie in $(0, 1)$, hence different combinations of the concentration parameters allow reaching a broad range of correlations. Although fSAN already attains increased flexibility, Figure 1 (first panel of the second row) shows that the correlation is adequately influenced by a only when b is very small. Indeed, its impact is almost negligible for most of its values, and all prior correlation is driven only by b. The fiSAN combines the best features of both the finite and infinite scenarios: both parameters sensibly affect the induced correlation without the need to push them close to the limit of their parameter space.

2.2 Partially exchangeable partition probability functions

Within the exchangeable framework, the probabilistic properties of the random partition can be characterized through the exchangeable partition probability function (EPPF, Pitman, 1995, 2006). This quantity is a fundamental tool for understanding the clustering structure induced by a process. When the data are organized into separate groups, an analogous notion to analyze the resulting partition is the *partially exchangeable partition probability function* (pEPPF, Lijoi et al., 2014a,b). Examples of pEPPFs in the Bayesian nonparametric literature for dependent models were provided by Lijoi et al. (2014a) for additive processes and by Camerlenghi et al. (2017) for hierarchical processes. In the nested framework, Camerlenghi et al. (2019) demonstrated the importance of the pEPPF for analyzing the clustering properties and assessing that no pathological behaviors arise, discovering, for example, the degeneracy issue of the nDP. Here, we provide explicit expressions for the pEFFPs of the fiSAN and fSAN

models, discussing their connections with the nDP. For notational convenience, we introduce the cluster allocation variables that express the cluster membership for each group and observation. In particular, we consider two sets of auxiliary categorical variables: $\boldsymbol{S} = \{S_j, j = 1, \dots, J\}$ with $S_j \in \{1, \dots, K\}$ indicating the DC allocation, and $\boldsymbol{M} = \{M_{i,j}, i = 1, \dots, N_j, j = 1, \dots, J\}$ with $M_{i,j} \in \{1, \dots, L\}$, indicating the OC allocation. Notice that the cluster allocation variables can take any value between 1 and the number of mixture components. However, some components could be empty (not used to generate the data), hence K and L do not coincide with the number of clusters, in general. Clearly, $p(S_j \mid \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \delta_k(\cdot)$, and $p(M_{i,j} \mid S_j, \boldsymbol{\omega}_{S_j}) = \sum_{l=1}^{L} \omega_{l,S_j} \delta_l(\cdot)$. The observations in each group can assume at most L distinct values. Thus, ties among the observations can appear, inducing a clustering configuration. We denote the frequency of cluster $l \in \{1, \dots, L\}$ in the j-th group as $n_{l,j} = \sum_{i=1}^{N_j} \mathbb{1}_{\{M_{i,j}=l\}}$.

For simplicity, in the rest of this section, we consider the setting where we only have two samples $\boldsymbol{\theta}_1 = \{\theta_{1,1}, \dots, \theta_{N_1,1}\}$ and $\boldsymbol{\theta}_2 = \{\theta_{1,2}, \dots, \theta_{N_2,2}\}$ of sizes N_1 and N_2, respectively. When only two groups are considered, the observed atoms can be either shared by the two samples or specific to only one group. We denote with s_0 the number of atoms that appear in both groups (i.e., those for which $n_{l,1} n_{l,2} > 0$); with s_1 the number of atoms specific to group 1 (i.e., $n_{l,1} > 0$ and $n_{l,2} = 0$); and, similarly, with s_2 the number of atoms specific to group 2. Hence, the number of empty clusters is $L - s$ with $s = s_0 + s_1 + s_2$. Our goal is to study the distribution of the partition of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Both the fiSAN and the fSAN model allow for a straightforward derivation and an analytical expression of the pEPPF, which is given in Theorem 2.1.

Theorem 2.1. *Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be two samples of sizes N_1 and N_2 from a SAN model. Let $s = s_0 + s_1 + s_2$ be the number of distinct values in $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and let $(\boldsymbol{n}_1, \boldsymbol{n}_2) = \big((n_{1,1}, \dots, n_{L,1}), (n_{1,2}, \dots, n_{L,2})\big)$ be the frequencies of each OC, with L the number of observational mixture components.*

The pEPPF of the finite-infinite SAN prior is expressed as

$$\Pi^{(fiSAN)}_{N_1,N_2,s}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{1}{\alpha+1} \Phi^{(D_L)}_{N_1+N_2,s}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \frac{\alpha}{\alpha+1} C^{L}_{s_0,s_1,s_2} \prod_{j=1}^{2} \Phi^{(D_L)}_{N_j,s_0+s_j}(\boldsymbol{n}_j), \quad (9)$$

where $\Phi^{(D_L)}_{N,s}(\boldsymbol{n})$ and $C^{L}_{s_0,s_1,s_2}$ denote the EPPF of a Dirichlet_L *distribution (Green and Richardson, 2001) and a correction constant, respectively, with expressions*

$$\Phi^{(D_L)}_{N,s}(\boldsymbol{n}) = \frac{L!\,\Gamma(Lb)\,\Gamma(b)^{-L}}{(L-s)!\,\Gamma(Lb+N)} \prod_{l=1}^{L} \Gamma(b+n_l), \quad C^{L}_{s_0,s_1,s_2} = \frac{(L-s_0-s_1)!(L-s_0-s_2)!}{L!(L-s_0-s_1-s_2)!}.$$

With the above definition of $\Phi^{(D_L)}_{N,s}(\boldsymbol{n})$, the pEPPF of the finite SAN is expressed as

$$\Pi^{(fSAN)}_{N_1,N_2,s}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{(1+a)}{(1+Ka)} \Phi^{(D_L)}_{N_1+N_2,s}(\boldsymbol{n}_1+\boldsymbol{n}_2) + \frac{(K-1)a}{(1+Ka)} C^{L}_{s_0,s_1,s_2} \prod_{j=1}^{2} \Phi^{(D_L)}_{N_j,s_0+s_j}(\boldsymbol{n}_j), \quad (10)$$

where K is the number of distributional mixture components.

The pEPPFs in (9) and (10) are convex combinations of two different scenarios, where the distributional cluster allocation probabilities, i.e., $\mathbb{P}[S_1 = S_2]$ and $\mathbb{P}[S_1 \neq S_2]$, play the role of mixing weights. The two scenarios correspond to the EPPFs of two extreme cases: the fully exchangeable case, represented by the EPPF of the pooled sample, and the unconditional independence case. Notice that the latter needs to be corrected by a constant $C^L_{s_0,s_1,s_2}$ to account for the presence of a finite set of shared atoms. A pure convex combination between the fully exchangeable case and the unconditional independence case – a finite-dimensional version of the results in Camerlenghi et al. (2019) – arises if, in our model, we replace $\theta^*_l \sim H$ with $\theta^*_{l,k} \sim H$, obtaining the finite-dimensional version of the nDP. However, this formulation would prevent the presence of any cross-DC observational cluster.

The pEPPFs derived in the previous paragraph are useful to understand the connection of SAN priors with other nonparametric models, in particular, the limiting behavior of the fSAN model when one sets $\alpha = a/K$ and studies $K \to \infty$, and that of the fiSAN model when one sets $b = \beta/L$ and $L \to \infty$.

Consider the pEPPF of SAN priors introduced before. Then, the following relationship holds across the different nested priors:

$$\lim_{L,K\to\infty} \Pi^{(\text{fSAN})}_{N_1,N_2,s}(\boldsymbol{n}_1,\boldsymbol{n}_2) = \lim_{L\to\infty} \Pi^{(\text{fiSAN})}_{N_1,N_2,s}(\boldsymbol{n}_1,\boldsymbol{n}_2) = \Pi^{(\text{nDP})}_{N_1,N_2,s}(\boldsymbol{n}_1,\boldsymbol{n}_2), \tag{11}$$

where

$$\Pi^{(\text{nDP})}_{N,s}(\boldsymbol{n}_1,\boldsymbol{n}_2) = \frac{1}{\alpha+1}\Phi^{(DP)}_{N_1+N_2,s}(\boldsymbol{n}_1+\boldsymbol{n}_2) + \frac{\alpha}{\alpha+1}\Phi^{(DP)}_{N_1,s}(\boldsymbol{n}_1)\Phi^{(DP)}_{N_2,s}(\boldsymbol{n}_2)\mathbb{1}_{\{s_0=0\}},$$

and $\Phi^{(DP)}_{N,s}(\boldsymbol{n})$ denotes the EPPF of a DP. More details are provided in Section B of the Supplementary Material.

The relations in (11) can be used to formalize some intuitions given in Section 2.1: sharing an infinite number of atoms that are, a priori, *all equally likely to be sampled* in each random measure G^*_k is not sufficient to prevent the model from collapsing to the fully exchangeable case, as indeed the probability of cross-DC ties approaches zero. Differently, both the CAM and the HHDP models impose, in terms of prior expectation, larger mass to the same sets of atoms across different groups: this characteristic is at the same time the solution of the degeneracy issue and the leading cause of the high correlation.

3 Posterior inference

In line with other nested mixture models, posterior inference for SAN models is not available in closed form, and we need to resort to computational approximations. The standard approach is to rely on Markov chain Monte Carlo (MCMC) techniques: in Section C.1 of the Supplementary Material, we outline a Gibbs sampler algorithm. However, it is well-known that MCMC methods have limited scalability when dealing with large datasets. As large amounts of data become ubiquitous, the computational burden of MCMC algorithms may constitute a problem and hinder the application of complex models.

3.1 Mean-field variational inference

Variational inference (VI) methods provide a viable solution to this problem, approaching posterior inference through optimization rather than simulation (Blei et al., 2017). These methods rely on finding, among the elements of a set of simple distributions, the one that better resembles the actual posterior in terms of their Kullback-Leibler (KL) divergence. The price of this greater scalability is a less precise posterior approximation, especially regarding the estimated variability. In the following, we extend mean-field VI strategies for mixtures models (Bishop, 2006) to the nested setting. In particular, we describe a coordinate ascent optimization algorithm to perform posterior inference under the fiSAN model (in line with the application in Section 5), while we defer to Section C.2 of the Supplementary Material for the corresponding derivation for the fSAN. To the best of our knowledge, no variational methods are yet available in the literature for nested Bayesian common atoms mixture models.

We consider d-variate observations $\boldsymbol{y}_{i,j}$, and we assume a mixture of multivariate Gaussian kernels $p(\cdot \mid \theta) = \phi_d(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, with $\boldsymbol{\mu}$ a d-dimensional mean vector and $\boldsymbol{\Lambda}$ a $d \times d$ precision matrix. Also, we assume a conjugate normal-Wishart prior distribution on the model parameters, $(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \text{NW}(\boldsymbol{\mu}_0, \kappa_0, \tau_0, \boldsymbol{\Gamma}_0)$. To derive the proposed algorithm, we exploit the model formulation based on the data augmentation scheme introduced in Section 2.2, which makes use of the cluster allocation variables $S_j \in \{1, 2, \dots\}$, $j = 1, \dots, J$, and $M_{i,j} \in \{1, \dots, L\}$, $i = 1, \dots, N_j$, $j = 1, \dots, J$. Thus, we can write the model as

$$p(\boldsymbol{y} \mid \boldsymbol{M}, \{\boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l\}_{l=1}^L) = \prod_{j=1}^{J} \prod_{i=1}^{N_j} \prod_{l=1}^{L} \phi_d(\boldsymbol{y}_{i,j} \mid \boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l^{-1})^{\mathbb{1}_{\{M_{i,j}=l\}}},$$

$$p(\boldsymbol{M} \mid \boldsymbol{S}, \boldsymbol{\omega}) = \prod_{j=1}^{J} \prod_{i=1}^{N_j} \prod_{k=1}^{\infty} \prod_{l=1}^{L} \omega_{l,k}^{\mathbb{1}_{\{M_{i,j}=l \cap S_j=k\}}}, \qquad p(\boldsymbol{S} \mid \boldsymbol{\pi}) = \prod_{j=1}^{J} \prod_{k=1}^{\infty} \pi_k^{\mathbb{1}_{\{S_j=k\}}}.$$

Finally, a gamma hyperprior on the concentration parameter of the distributional DP is assumed: $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$ (Escobar and West, 1995). In contrast, the Dirichlet parameter b is assumed to be known and set to a small value to ensure sparsity (i.e., $b < \zeta/2$, with ζ the dimension of the component-specific parameter, following Rousseau and Mengersen, 2011).

To set up a variational strategy for the fiSAN, we first need to define a suitable set of variational distributions. In the mean-field variational framework, this class is given by the set of densities where the latent variables are mutually independent. Following Blei and Jordan (2006), we use a truncated variational family to deal with the nonparametric mixture at the distributional level, where the truncation level is denoted with T. The fully factorized family of distributions that we assume can be written as

$$q(\boldsymbol{M}, \boldsymbol{S}, \{\boldsymbol{\omega}_k\}_{k=1}^T, \boldsymbol{v}, \{\boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l\}_{l=1}^L; \boldsymbol{\lambda}) = \prod_{j=1}^{J} q(S_j; \{\rho_{j,k}\}_{k=1}^T) \prod_{j=1}^{J} \prod_{i=1}^{N_j} q(M_{i,j}; \{\xi_{i,j,l}\}_{l=1}^L) \times$$
$$\times \prod_{k=1}^{T} q(v_k; \bar{a}_k, \bar{b}_k)\, q(\alpha; s_1, s_2) \prod_{k=1}^{T} q(\boldsymbol{\omega}_k; \{p_{l,k}\}_{l=1}^L) \prod_{l=1}^{L} q(\boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l; \boldsymbol{m}_l, t_l, c_l, \boldsymbol{D}_l),$$

where $q(S_j; \{\rho_{j,k}\}_{k=1}^T)$ and $q(M_{i,j}; \{\xi_{i,j,l}\}_{l=1}^L)$ are multinomial distributions; $q(v_k; \bar{a}_k, \bar{b}_k)$ are beta distributions, and they are such that $q(v_T = 1) = 1$ and $q(v_g = 0) = 1$ for $g > T$; $q(\alpha; s_1, s_2)$ is a gamma distribution; $q(\boldsymbol{\omega}_k; \{p_{l,k}\}_{l=1}^L)$ are Dirichlet distributions; and $q(\boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l; \boldsymbol{m}_l, t_l, c_l, \boldsymbol{D}_l)$ are normal-Wishart distributions. Under this representation, the set of latent variables is $\boldsymbol{\Theta} = \left(\boldsymbol{S}, \boldsymbol{M}, \boldsymbol{v}, \{\boldsymbol{\omega}_k\}_{k=1}^T, \{\boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l\}_{l=1}^L, \alpha\right)$ and the set of variational parameters is $\boldsymbol{\lambda} = \left(\boldsymbol{\rho}, \boldsymbol{\xi}, \bar{\boldsymbol{a}}, \bar{\boldsymbol{b}}, s_1, s_2, \boldsymbol{p}, \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{c}, \boldsymbol{D}\right)$. Optimization is then carried out by looking for the combination of variational parameters $\boldsymbol{\lambda}^\star$ that maximizes the evidence lower bound (ELBO). To this end, the most commonly used algorithm is the coordinate-ascent variational inference algorithm (CAVI – see, for example, Bishop, 2006). We report the CAVI updating rules for fiSAN in Algorithm 1, while we defer to Section C.3 in the Supplementary Material for additional details on the ELBO and its evaluation. We outline the algorithm for the specific case of multivariate Gaussian likelihood, however, this approach can be adapted in a straightforward manner whenever the data distribution is a member of the exponential family, as discussed in Blei and Jordan (2006).

4 Simulation studies

We illustrate two simulation studies to assess the performances of the proposed priors and computational approach on synthetic data. The first study aims to compare the fSAN and fiSAN models introduced in Section 2 with two methods based on the same common atoms structure, to ease the comparison (specifically, the CAM and fCAM). For this simulation study, we consider univariate Gaussian kernels. This setting also provides a first assessment of the efficacy of the variational inference approach outlined in Section 3 in a simple univariate setting. The second study focuses instead on the evaluation of how the proposed model and CAVI algorithm behave in a multivariate framework. Here, we especially focus on the clustering accuracy and scalability of MCMC and VI as the dimensionality and sample size increase.

4.1 Univariate case

This simulation study aims to provide empirical evidence for the prior properties presented in the first part of the paper and, specifically, on the accuracy of different prior specifications. We compare both the ability to recover the true two-level partition of the data and to estimate the posterior densities of the different DCs. Additionally, we test the accuracy and efficiency of variational inference compared to the well-established MCMC procedure. This way, we are able to simultaneously evaluate:

i. if the proposed fSAN and fiSAN models are competitive with state-of-the-art models (based on "standard" MCMC methods);

ii. if the proposed CAVI algorithm for estimating SAN priors has accuracy comparable with a Gibbs sampler approach in a simple, univariate setting.

We compare the performances of the fSAN and fiSAN models based on overfitted finite mixtures with the CAM of Denti et al. (2023) and the fCAM of D'Angelo et al. (2023).

Algorithm 1: CAVI updates for the fiSAN model.

Input: $t \leftarrow 0$. Randomly initialize $\boldsymbol{\lambda}^{(0)}$. Define the threshold ϵ and randomly set $\Delta > \epsilon$.
while $\Delta(t-1,t) > \varepsilon$ **do**
Set $t = t+1$; Let $\boldsymbol{\lambda}^{(t-1)} = \boldsymbol{\lambda}^{(t)}$;
Update the variational parameters according to the following CAVI steps:

1. For $j = 1, \dots, J$, $q^\star(S_j)$ is a T-dimensional multinomial, with $q^\star(S_j = k) = \rho_{j,k}$ for $k = 1, \dots, T$, and
$$\log \rho_{j,k} = g(\bar{a}_k, \bar{b}_k) + \sum_{r=1}^{k-1} g(\bar{b}_r, \bar{a}_r) + \sum_{l=1}^{L} \left(\sum_{i=1}^{N_j} \xi_{i,j,l} \right) h_l(\boldsymbol{p}_k),$$
where $g(x,y) = \psi(x) - \psi(x+y)$ and $h_k(\boldsymbol{x}) = \psi(x_k) - \psi(\sum_{k=1}^K x_k)$, with ψ denoting the digamma function.
2. For $j = 1, \dots, J$ and $i = 1, \dots, N_j$, $q^\star(M_{i,j})$ is a L-dimensional multinomial, with $q^\star(M_{i,j} = l) = \xi_{i,j,l}$ for $l = 1, \dots, L$, and
$$\log \xi_{i,j,l} = \frac{1}{2}\ell_l^{(1)} + \frac{1}{2}\ell_{i,j,l}^{(2)} + \sum_{k=1}^{T} \rho_{j,k} h_l(\boldsymbol{p}_k).$$
where $\ell_l^{(1)} = \sum_{x=1}^{d} \psi\left((c_l - x + 1)/2\right) + d\log 2 + \log|\boldsymbol{D}_l|$ and $\ell_{i,j,l}^{(2)} = -d/t_l - c_l(\boldsymbol{y}_{i,j} - \boldsymbol{m}_l)^\top \boldsymbol{D}_l(\boldsymbol{y}_{i,j} - \boldsymbol{m}_l)$.
3. For $k = 1, \dots, T$, $q^\star(\boldsymbol{\omega}_k)$ is $\text{Dirichlet}_L(\boldsymbol{p}_k)$ with $p_{l,k} = b + \sum_{j=1}^{J}\sum_{i=1}^{N_j} \xi_{i,j,l}\rho_{j,k}$.
4. For $k = 1, \dots, T$, $q^\star(v_k)$ is a $\text{Beta}(\bar{a}_k, \bar{b}_k)$ distribution with
$$\bar{a}_k = 1 + \sum_{j=1}^{J} \rho_{j,k}, \quad \bar{b}_k = s_1/s_2 + \sum_{j=1}^{J}\sum_{q=k+1}^{T-1} \rho_{j,q}.$$
5. For $l = 1, \dots, L$ $q^\star(\theta_l)$ is a $\text{NW}(\boldsymbol{m}_l, t_l, c_l, \boldsymbol{D}_l)$ distribution with parameters
$$\boldsymbol{m}_l = t_l^{-1}(\kappa_0\,\boldsymbol{\mu}_0 + N_{\cdot l}\bar{\boldsymbol{y}}_l), \quad t_l = \kappa_0 + N_{\cdot l}, \quad c_l = \tau_0 + N_{\cdot l},$$
$$\boldsymbol{D}_l^{-1} = \boldsymbol{\Gamma}_0^{-1} + \frac{\kappa_0 N_{\cdot l}}{\kappa_0 + N_{\cdot l}} (\bar{\boldsymbol{y}}_l - \boldsymbol{\mu}_0)(\bar{\boldsymbol{y}}_l - \boldsymbol{\mu}_0)^\top + \mathcal{S}_{\cdot l},$$
where
$$N_{\cdot l} = \sum_{j=1}^{J}\sum_{i=1}^{N_j} \xi_{i,j,l}, \qquad \bar{\boldsymbol{y}}_l = N_{\cdot l}^{-1}\Big(\sum_{j=1}^{J}\sum_{i=1}^{N_j} \xi_{i,j,l}\boldsymbol{y}_{i,j}\Big),$$
$$\mathcal{S}_{\cdot l} = \sum_{j=1}^{J}\sum_{i=1}^{N_j} \xi_{i,j,l}\,(\boldsymbol{y}_{i,j} - \bar{\boldsymbol{y}}_l)(\boldsymbol{y}_{i,j} - \bar{\boldsymbol{y}}_l)^\top.$$
6. $q^\star(\alpha)$ is a $\text{Gamma}(s_1, s_2)$ distribution with parameters
$$s_1 = a_\alpha + T - 1, \quad s_2 = b_\alpha - \sum_{k=1}^{T-1} g(\bar{b}_k, \bar{a}_k).$$

Store the updated parameters in $\boldsymbol{\lambda}$ and let $\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}$;
Compute $\Delta(t-1,t) = \text{ELBO}(\boldsymbol{\lambda}^{(t)}) - \text{ELBO}(\boldsymbol{\lambda}^{(t-1)})$.
return $\boldsymbol{\lambda}^\star$, *containing the optimized variational parameters.*

In particular, under the latter model, both L and K are random and estimated using the telescoping sampler of Frühwirth-Schnatter et al. (2021). While we have examined the similarities and differences of the correlation structure with other nested priors from a theoretical perspective in Section 2, the computational complexity of those models hinders a practical comparison in settings with many groups and large sample sizes.

For this experiment, we considered a nested dataset where each group-specific distribution is a mixture of univariate Gaussian kernels, denoted as $\phi(y \mid \mu, \sigma^2)$. Specifically, the data-generating process is a nested mixture made of three distributional clusters with homogeneous probabilities $(\pi_1, \pi_2, \pi_3) = (1/3, 1/3, 1/3)$, where the density $f_k(y)$ characterizing the k-th cluster, $k = 1, 2, 3$, is given by

$$f_1(y) = 0.5\, \phi(y \mid -5, 0.6^2) + 0.5\, \phi(y \mid -2, 0.6^2),$$
$$f_2(y) = 0.5\, \phi(y \mid 2, 0.6^2) + 0.5\, \phi(y \mid 5, 0.6^2),$$
$$f_3(y) = \phi(y \mid 0, 0.6^2).$$

We independently extracted $J = 6$ groups with equal sample sizes from this distribution, obtaining two samples for each distributional cluster. We considered four configurations corresponding to varying sample sizes of each group: $N_j \in \{10, 50, 500, 2500\}$, for $j = 1, \ldots, 6$. Therefore, the total sample size N ranges from 60 to 15,000. Considering small samples allows us to investigate if there are problematic situations when the information conveyed by the data is limited, and the posterior estimates are heavily influenced by the prior. On the contrary, the large-sample scenarios allow us to evaluate the computational burden of the algorithms. For each configuration, we replicated the experiment over 50 independently simulated datasets.

All priors comprise some relevant parameters that affect their distributional properties, as discussed in Section 2.1. Following Denti et al. (2023), for the CAM the concentration parameters α and β are now assigned Gamma$(1, 1)$ hyperpriors. Since the parameters are random, we do not have an analytical expression of the prior correlation between pairs of random measures. Instead, we computed a Monte Carlo estimate, which resulted in a mean correlation of 0.8914. Similarly, for the fCAM, the Dirichlet parameters a and b are assigned Gamma$(10, 10)$ distributions, while the parameters L and K are assigned beta-negative-binomial distributions BNB$(1, 4, 3)$, following the indications of Frühwirth-Schnatter et al. (2021). Interestingly, this specification leads to a Monte Carlo estimate of the correlation equal to the one of the CAM.

For models based on finite overfitted mixtures, choosing appropriate Dirichlet parameters (dimension and concentration parameter) is key for obtaining good posterior estimates. The Dirichlet concentration parameter should satisfy the condition derived in Rousseau and Mengersen (2011) to induce sparsity and empty superfluous components. The Dirichlet dimension should be set large enough to guarantee that it exceeds the number of clusters expected in the data. However, it is interesting that, as long as these conditions are satisfied, the posterior estimates are quite robust to the specific values of such parameters (a sensitivity analysis is reported in Section D.3 of the Supplementary Material). Specifically, for the fiSAN, the parameter α of the DP at the top level is assigned a Gamma$(1, 1)$ hyperprior; at the observational level, the parameters

are instead set to $L = 25$ and $b = 0.05$. This combination leads to a correlation of 0.6309. For the fSAN, all parameters are fixed, and, specifically, $a = b = 0.05$, $K = 20$, and $L = 25$. The resulting prior correlation is equal to 0.5657. Notice that increasing the Dirichlet dimensions K and L requires allocating larger matrices, and this is particularly true for MCMC approaches. Therefore, especially with large data sets, the choice of these parameters should be made in a reasoned way, taking into account the computational cost. Finally, given the univariate nature of the problem, we adopted a conjugate normal-inverse gamma base measure $(\mu_l, \sigma_l^2) \sim \text{NIG}(\mu_0, \kappa_0, \tau_0, \Gamma_0)$, whose hyperparameters were fixed to $(\mu_0, \kappa_0, \tau_0, \Gamma_0) = (0, 0.01, 3, 2)$.

Turning now to the parameters of the variational density, in the sensitivity analysis we evaluated the impact of the truncation parameter T used in the variational version of the fiSAN (which uses a DP at the distributional level). For this parameter, a similar reasoning to that of L and K applies: as long as T is fixed large enough, the algorithm can freely explore the space of reasonable partitions, and the estimates are unaffected by the specific truncation value. Specifically, we used $T = 20$. Finally, the variational distribution of the univariate kernel parameters is a normal-inverse gamma, denoted as $q(\mu_l, \sigma_l^2; m_l, t_l, c_l, d_l)$.

In the following, we discuss the relevant aspects of the posterior inference. The algorithms used for this simulation study are written in efficient Rcpp language, and they are available in the R packages `SANple` (D'Angelo and Denti, 2023) and `SANvi` (Denti and D'Angelo, 2023), both downloadable from the Comprehensive R Archive Network; the scripts to replicate the analyses are available at the GitHub repository `Fradenti/SAN4ba`. All analyses were performed on a Linux server running an AMD EPYC-Rome Processor with 405 GB of RAM.

Comparison between shared atoms nested priors

We start by analyzing the accuracy of the different nested priors in estimating the two-level partition. Depending on the computational strategy, the posterior point estimates of the clusters were obtained using different procedures. When dealing with the MCMC output, the chains of the cluster allocation variables S_j and $M_{i,j}$ were used to compute the corresponding posterior similarity matrices. Then, the optimal partitions were estimated by minimizing the variation of information loss (Wade and Ghahramani, 2018), employing the algorithm developed by Dahl et al. (2022). When dealing with the VI approach, the algorithm returns instead the optimized variational parameters corresponding to the cluster assignment probabilities, i.e., $\hat{\rho}_{j,k} = q^*(S_j = k)$ and $\hat{\xi}_{i,j,l} = q^*(M_{i,j} = l)$. The former indicates the posterior point estimate of the probability of assigning the j-th group to the k-th distributional cluster; similarly, the latter represents the posterior estimate of the probability of assigning the i-th observation of the j-th group to the l-th observational cluster. Hence, in this case, the partitions were estimated as

$$\hat{S}_j = \underset{k=1,\dots,T}{\arg\max}\, \hat{\rho}_{j,k} \quad \text{and} \quad \hat{M}_{i,j} = \underset{l=1,\dots,L}{\arg\max}\, \hat{\xi}_{i,j,l}$$

for $j = 1, \dots, J$ and $i = 1, \dots, N_j$. Then, to measure the accuracy of the estimated partitions, we compared the adjusted Rand index (ARI, Rand, 1971; Hubert and Arabie,

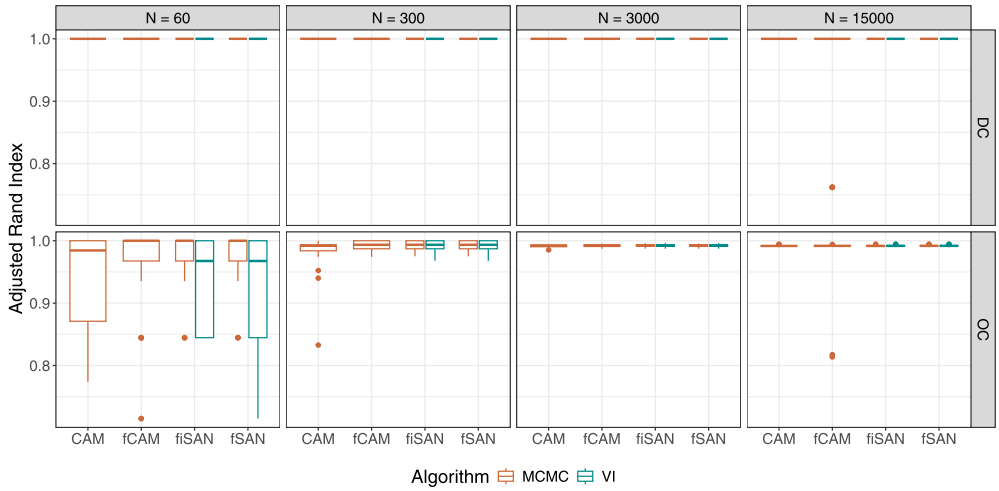

Figure 2: Accuracy of the estimated distributional (top panel) and observational (bottom panel) clustering for the CAM, fCAM, fiSAN, and fSAN. Each panel shows the distribution of the ARI obtained across the 50 replications, for each configuration. For the fiSAN and fSAN, colors correspond to the algorithm.

1985) between the posterior point estimate and the true partition. An ARI equal to zero corresponds to random labeling, while an ARI equal to one indicates that the clusterings are identical. Figure 2 shows the distribution of the ARIs (over the 50 replications) obtained by the different models for each configuration. All models adequately estimate the distributional partition under each scenario. At the observational level, all models have remarkable performances, with an ARI over 0.8 for almost all replications; moreover, the posterior point estimates of the clustering consistently improve with increasing sample size. For the fSAN and fiSAN, the variational approach produces slightly worse posterior estimates than the MCMC. Still, in general, the difference is small and is due to a very limited number of misclassified observations.

We now inspect the ability of the different models to estimate the density of the data. We first needed to obtain a posterior point estimate of the density of each group. When using an MCMC approach, we computed the pointwise mixture density on a grid of points at each iteration by substituting the current values of the chains into the theoretical data density; then, we obtained the posterior estimate by averaging them. This strategy was necessary for dealing with the varying partition across iterations and the additional complications caused by the label-switching. We used instead a slightly different procedure when using a CAVI algorithm since this approach does not produce a sample of replications but only a single estimated model. Having a unique point estimate of the model hyperparameters allows for overcoming the problem of label-switching; moreover, representing only one of the several modes (given by permutations of the indices) is sufficient to adequately perform posterior inference (Blei et al., 2017). Here, we substituted the estimated variational posterior expected values $\hat{\mu}_l$, $\hat{\sigma}_l^2$ and $\hat{\omega}_{l,k}$

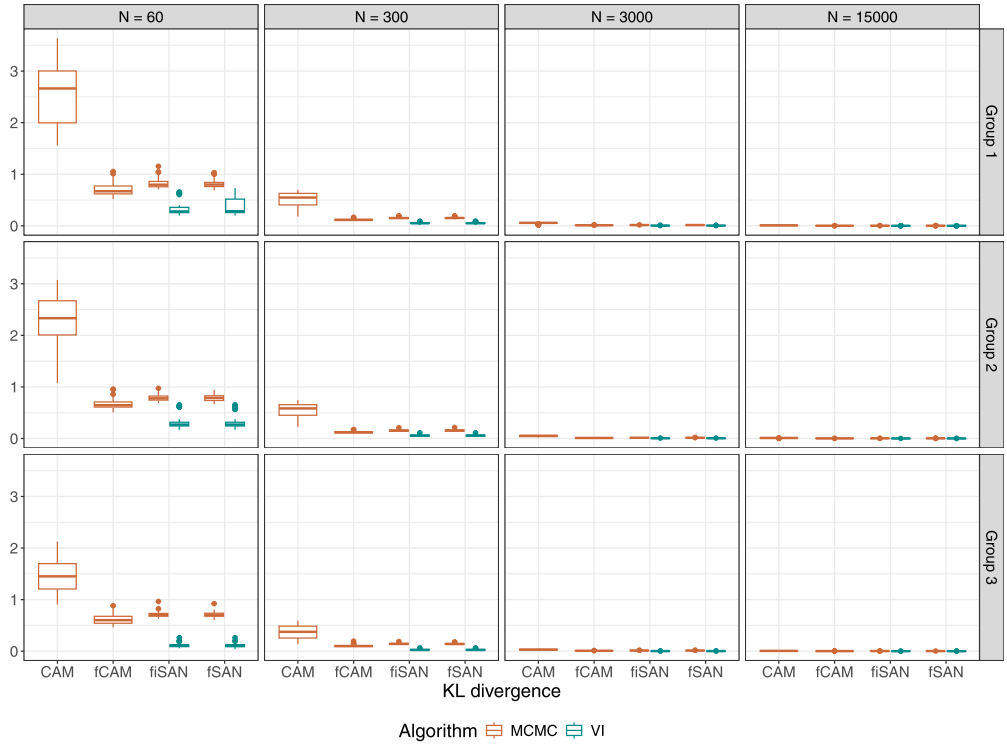

Figure 3: Accuracy of the density estimates for groups 1, 2, and 3, for the considered models. Each panel shows the distribution of the KL divergence (over the 50 replications), for each configuration. For the fiSAN and fSAN, colors correspond to the algorithm.

into the parametric mixture density. In this context, $\hat{\mu}_l = m_l^\star$, $\hat{\sigma}_l^2 = d_l^\star/(c_l^\star - 1)$, while the weights $\hat{\omega}_{l,k}$'s are the posterior means of the corresponding Dirichlet distributions. With this estimate, we computed the density on the same grid of points. Figure 3 shows the Kullback-Leibler divergence distribution between the true and the estimated mean density. In line with the previous analysis, we show the CAM, fCAM, fSAN, and fiSAN results, where the latter were estimated via Gibbs sampler and CAVI. The plot clearly shows that the CAM has difficulties estimating the density when the sample size is small, as highlighted by the large KL divergence in the first configuration. Increasing the sample size alleviates this problem, as the data convey enough information to overcome the prior distribution. On the contrary, all models based on finite observational mixtures have good posterior estimates. This result is particularly interesting if one considers that the correlations of the CAM and fCAM were equal, *a priori*. Hence, the improved posterior density estimates of the fCAM are purely a consequence of the Dirichlet distributions of the weights.

To gain more insight into the reason that led to this behavior, Figure 4 shows the posterior point estimate of the density of the first three groups computed via Gibbs sampler under the first configuration, corresponding to $N = 60$. Figure S.6 in the Supplementary Material shows the posterior density estimate obtained via CAVI for the fSAN and fiSAN. We considered three groups whose distributions are representative of the three DCs. The first column shows the density estimates for CAM and highlights the pitfalls of using this model when the distributional atoms do not share observational atoms. The CAM indeed expects the different distributional clusters to have some atoms in common, and even if the data do not support this assumption, the model forces a similarity in the estimated density. In particular, the forced positive correlation between distributions causes the presence of "ghost" modes inherited from the other DCs. Conversely, models based on symmetric Dirichlet distributions can avoid this spurious borrowing of information, even for small sample sizes.

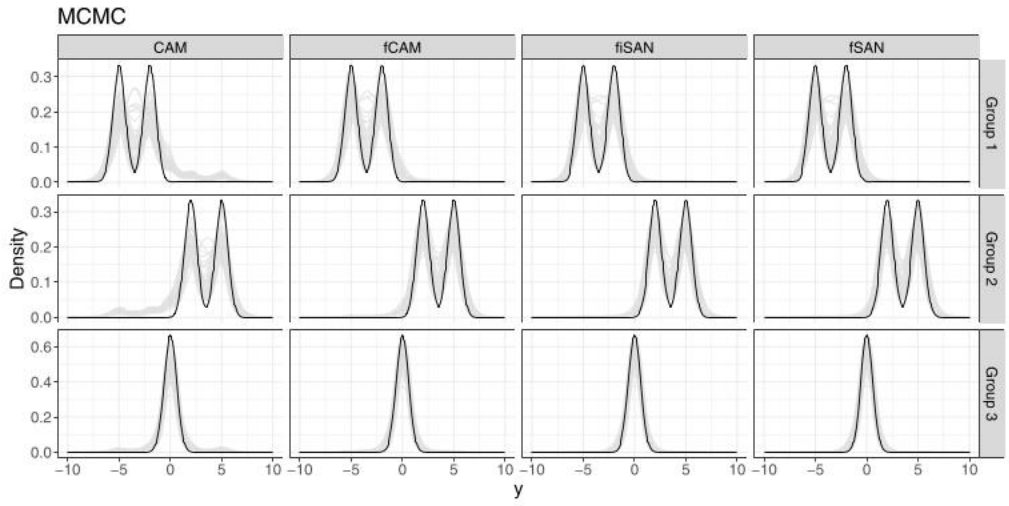

Figure 4: Posterior density estimate for groups 1, 2, and 3, for the CAM, fCAM, fiSAN, and fSAN. Each panel shows the true density (black line) and the posterior density estimates (grey lines) obtained over the 50 replications, for configuration 1 ($N = 60$).

Comparison between computational approaches

With the previous setting, we have already provided evidence that the proposed variational inference algorithm is a valid tool for estimating the data partition and density. That said, it is well-known that CAVI can severely underestimate the variability of the posterior distributions of the parameters. In this section, we further investigate and compare the performances of the Gibbs sampler and CAVI algorithms. Here, again, the goal is twofold: assessing the accuracy of the posterior distributions, and investigating the computational burden of the two approaches. Indeed, the actual need for a VI procedure for this type of model has not yet been discussed or corroborated by empirical evidence.

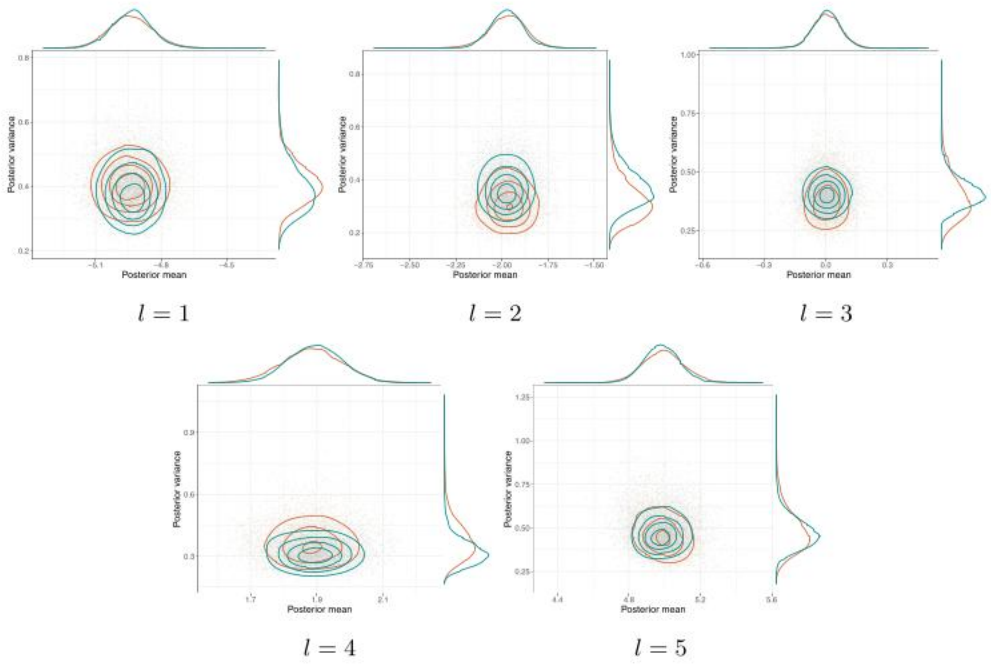

Figure 5: fiSAN prior - Configuration 2: posterior density estimate of (μ_l, σ_l^2), for $l = 1, 2, 3, 4, 5$, obtained using a Gibbs sampler (orange line) and a CAVI algorithm (green line). Each panel shows the contour plot of the joint density, together with the two marginal densities.

Figure 5 shows the estimated posterior density of the cluster-specific parameters (μ_l, σ_l^2), for $l = 1, 2, 3, 4, 5$ (i.e., the atoms actually used to generate the data), under the fiSAN prior estimated via MCMC and VI. Each panel shows the contour plot of the joint density, together with the marginal densities. The results refer to the second configuration, corresponding to $N = 300$. In Section D.2 of the Supplementary Material, we provide additional graphs corresponding to different sample sizes and to the estimates under the fSAN prior. Inference on the parameters of the observational atoms is not immediate when employing an MCMC approach, as chains may be affected by label-switching, a common problem when dealing with mixture models (Stephens, 2000). Moreover, a comparison between the two algorithms requires matching the cluster-specific parameters arising from two fundamentally different approaches. To solve the issue, we post-processed the chains using the relabelling Equivalence Classes Representatives algorithm (Rodríguez and Walker, 2014), as implemented in the R package `label.switching` (Papastamoulis, 2016). Sections D.1 and D.2 of the Supplementary Material detail the post-processing to treat the label-switching and the procedure to derive these plots.

Figure 5, as well as Figures S7, S8, and S9, do not highlight any systematic troubling behavior of the VI approach compared to the MCMC. We acknowledge that, in some

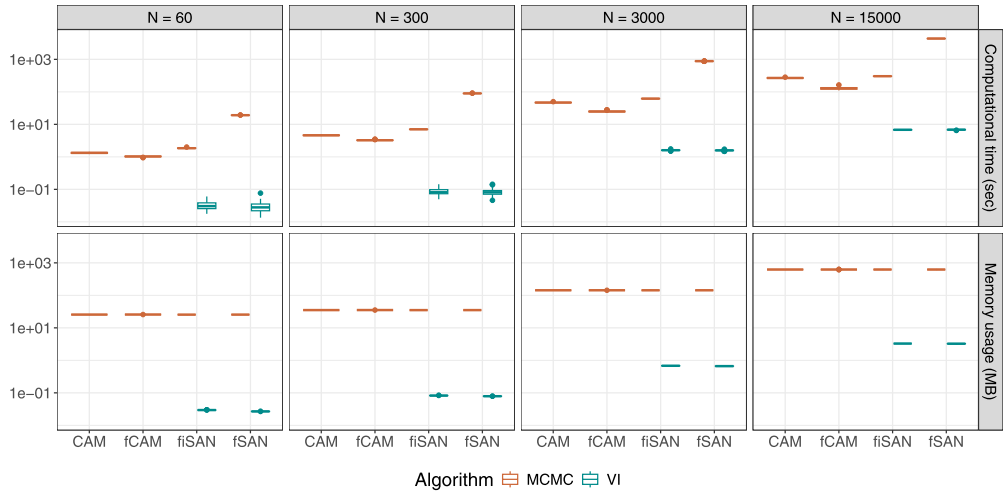

Figure 6: Top row: distributions of the computational time (in seconds) over the 50 replications for the two algorithms, for each configuration. Bottom row: distributions of the memory usage (in MB) over the 50 replications for the two algorithms for each configuration. Values are displayed on $\log_{10}$ scale.

cases, the modes of the distributions do not strictly coincide, and the VI estimates have a larger bias compared to the MCMC. Moreover, the marginal variances of the posterior distributions are sometimes underestimated using the VI approach. However, such discrepancies are not severe, and the CAVI estimates are overall satisfactory. Moreover, the computational advantages of VI largely compensate for the approximation, as displayed in the subsequent simulations, and, especially, in the multivariate setting discussed in Section 4.2.

Turning now to the rationale for the need for a variational approach, we compare the computational cost of the two algorithms: here, we consider both the memory usage and the computational time. Indeed, for the MCMC methods, the need to store the entire Markov chains (at most, excluding the burn-in) can raise issues with memory allocation when dealing with big data. On the contrary, VI methods only require storing the optimized parameters. The panels in the top row of Figure 6 clearly show the advantage of a VI approach over the MCMC regarding memory usage for all sample sizes, even in this univariate case.

Obtaining a fair efficiency comparison is more complex since the two algorithms are fundamentally different. However, to guarantee convergence and adopt a procedure that would be reasonable in a real-data application, we proceeded as follows. The computational time of the Gibbs sampler algorithm was computed as the total run time to generate 10,000 iterations. As for the VI approach, we considered $\Delta(t-1,t) < 10^{-4}$ as a stopping rule to define the convergence of the ELBO. Although the discrepancy between the variational distribution and the target posterior is reduced at each iteration, there is

no guarantee that the CAVI algorithm will lead to a global optimum. On the contrary, depending on the initial configuration, it will likely obtain a local solution. Hence, we executed 50 distinct runs of the algorithm with different starting points, keeping the one with the highest ELBO to draw the inference. The ultimate advantage of VI is that these optimizations are easily parallelizable; thus, we report the maximum individual run time obtained over the 50 runs for each dataset, which can be seen as an indicator of the computational cost of the CAVI. The panels in the top row of Figure 6, displaying the elapsed seconds on the log scale, confirm that the VI approach is at least one order of magnitude faster than the MCMC, and this gap increases with increasing sample size.

4.2 Multivariate case

The second simulation study is devised to investigate the performance of the fiSAN model when dealing with large multivariate data. We focus solely on this model since the two competitors, CAM and fCAM, were only developed for univariate data. Moreover, from the findings in the previous paragraph, we see that all the formulations relying on a finite set of observational atoms appear to have overall the best performances. Since there is no clear evidence of superiority of a particular specification over the others, we decided to focus on the finite-infinite one. Let $\boldsymbol{y}_{i,j}$ represent a vector of dimension $d \geq 1$, $\boldsymbol{y}_{i,j} = (y_{i,j,1}, \ldots, y_{i,j,d})^\top$ for $i = 1, \ldots, N_j$, $j = 1, \ldots, J$. Here, we study and compare the MCMC and VI approaches in terms of classification accuracy and scalability in this multivariate framework.

The data-generating process is now a nested mixture of multivariate Gaussian kernels. Specifically, it is an extension of the data-generating process of Section 4.1 to dimensions $d \in \{2, 5, 10\}$. Again, we have $J = 6$ groups extracted from a nested mixture of three distributional atoms $f_k(\boldsymbol{y})$, $k = 1, 2, 3$, with homogeneous probabilities $1/3$, where each atom is a mixture of multivariate Gaussian kernels with different mean vectors and covariance matrices. The densities $f_k(\boldsymbol{y})$ are defined as

$$
\begin{aligned}
f_1(\boldsymbol{y}) &= 0.5\, \phi_d(\boldsymbol{y} \mid -5 \cdot \mathbf{1}_d, 0.2 \cdot \mathrm{I}_d) + 0.5\, \phi_d(\boldsymbol{y} \mid -2 \cdot \mathbf{1}_d, 0.2 \cdot \mathrm{I}_d \cdot \mathrm{R}_1) \\
f_2(\boldsymbol{y}) &= 0.5\, \phi_d(\boldsymbol{y} \mid 2 \cdot \mathbf{1}_d, 0.2 \cdot \mathrm{I}_d) + 0.5\, \phi_d(\boldsymbol{y} \mid 5 \cdot \mathbf{1}_d, 0.2 \cdot \mathrm{I}_d \cdot \mathrm{R}_2) \\
f_3(\boldsymbol{y}) &= \phi_d(\boldsymbol{y} \mid 0 \cdot \mathbf{1}_d, 0.2 \cdot \mathrm{I}_d \cdot \mathrm{R}_3).
\end{aligned}
$$

where $\mathbf{1}_d$ denotes a d-dimensional vector of ones and I_d a $d \times d$ identity matrix. The matrices R_k, $k = 1, 2, 3$ are correlation matrices that induce different types of dependence across variables. The correlation matrix R_1 is a band matrix, with entries equal to 0.25 for $|h_1 - h_2| < 2$ and 0 otherwise ($h_1, h_2 = 1, \ldots, d$); the matrix R_2 assumes a correlation equal to 0.5 between each pair of variables; finally, R_3 assumes a correlation equal to 0.85 between each pair of variables. Similarly to the previous section, we considered homogeneous group sample sizes N_j and studied the performances for varying $N_j \in \{50, 500, 1000\}$. Hence, the total sample size ranges from 300 to 6000. We replicated the experiment over 50 independently simulated datasets. The Dirichlet parameters are equal to the ones in Section 4.1. The hyperparameters on the multivariate normal-Wishart base measures, $(\boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l) \sim \mathrm{NW}(\boldsymbol{\mu}_0, \kappa_0, \tau_0, \boldsymbol{\Gamma}_0)$, are set to $(\boldsymbol{\mu}_0, \kappa_0, \tau_0, \boldsymbol{\Gamma}_0) = (\mathbf{0}_d, 0.01, d + 5, \mathrm{I}_d)$.

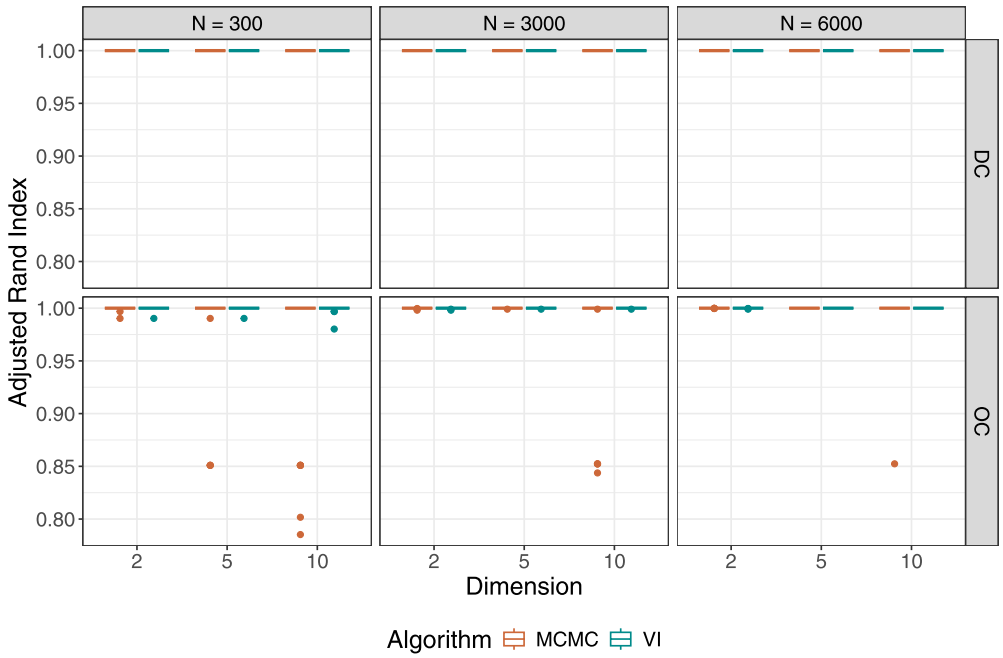

Figure 7: Distributions of the distributional (top row) and observational (bottom row) ARIs over the 50 replications for the fiSAN model estimated via VI and MCMC for each data dimension and sample size.

Accuracy of posterior inference

We start by analyzing the accuracy of the fiSAN model in recovering the observational and distributional data partition. Figure 7 shows the observational and distributional ARI obtained in each scenario using both an MCMC and a VI approach. Overall, the model performs well in all cases, with larger sample sizes leading to better posterior point estimates. Moreover, the two algorithms have comparable performances at clustering groups and observations, for all sample sizes and dimensions.

Computational aspects

The computational advantages of variational inference algorithms compared to standard MCMC approaches are presumably the main reason for turning to these approximate methods. We already highlighted these advantages in the univariate case, but they become particularly evident in this multivariate framework. The bottom panels of Figure 8 show the distribution of the memory usage in MB for the two types of algorithm in each scenario. The top panels show, instead, the total computational time. The computational time was obtained using the same definition as the previous simulation study. Both sets of plots showcase the need for a variational approach when the data are multivariate,

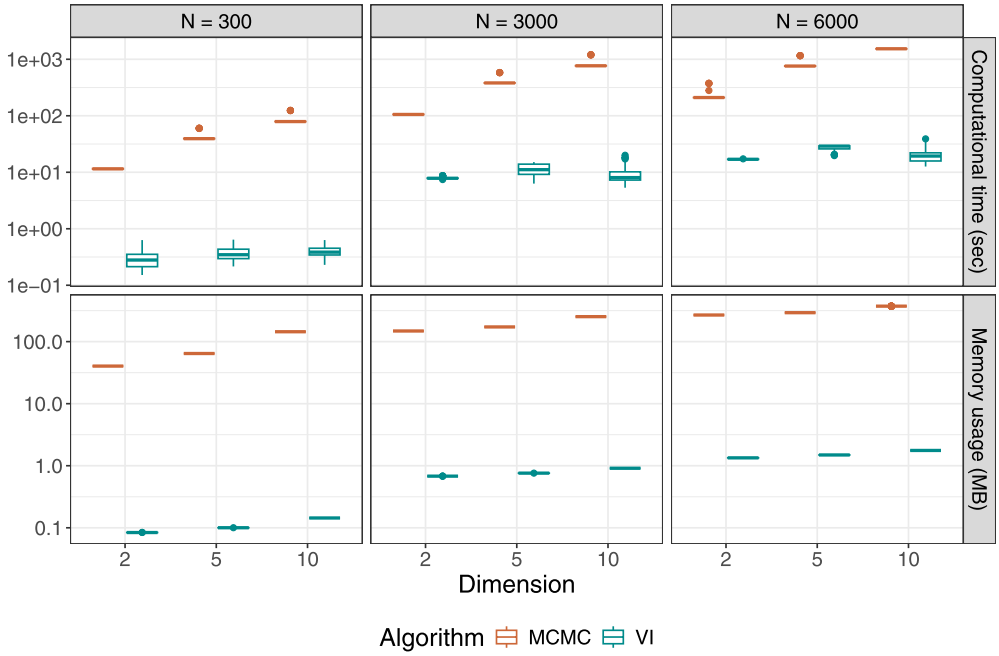

Figure 8: Distributions of the computational time (top row) and memory usage (bottom row) over the 50 replications for the fiSAN model estimated via VI and MCMC for each data dimension and sample size. Values are displayed on $\log_{10}$ scale.

and the sample size is moderately large. The memory usage and computational time of MCMC methods appear to grow with an exponential trend as d and N increase: both aspects can indeed hinder the application of complex Bayesian models to large multivariate data. Ultimately, variational inference appears to be a good compromise, balancing good posterior inference with a remarkably efficient implementation.

5 Spotify data analysis for the discovery of music profiles

Our study considers an open-source Spotify dataset available from the Kaggle platform.[1] Spotify is one of the largest music streaming service providers. Part of its success is due to a personalized recommendation system, which analyzes the user's listening history and suggests new, potentially relevant tracks. Because of the success of these algorithms, increasing interest has gone into understanding what makes two songs "similar" from the user's enjoyment point of view.

To this end, Spotify developed several scores to summarize various features of a song.

[1] https://www.kaggle.com/ektanegi/spotifydata-19212020

These features provide a description of a song's mood (e.g., *danceability*, *energy*), audio characteristics (e.g., *loudness*, *speechiness*, *instrumentalness*), and context (e.g., *liveness*, *acousticness*). One can find more details about these features in the documentation available on the *Spotify for developer* webpage.[2] The original dataset contains ten scores for over 160,000 songs released between 1921 and 2020, authored by over 1500 artists. We performed a preprocessing phase meant to discard outliers (e.g., tracks containing entire concerts, thus having exceptional duration; silent tracks, with extremely low energy, or, conversely, pure applause in live tracks, with extremely high energy levels). We then proceeded to select a large subset of artists for our analysis. We kept artists that authored more than 100 songs to ease the detection of DCs, and less than 200 songs, obtaining a dataset with 19,315 songs partitioned into 154 artists. Additional details on the preprocessing phase are available as Supplementary Material (Section E.1).

The goal of our analysis is to identify clusters of similar artists and songs based on their characteristics. This way, the system could rely on songs and artists within the same observational or distributional cluster, respectively, for the creation of playlists and listening suggestions. Specifically, we focus on three meaningful indicators: the *duration* (D), the *energy* (E), and the *speechiness* (S) of each song. In this sense, our problem can be formalized as a two-level multivariate clustering, where the songs (i.e., the observations) are exchangeable data points "within" each artist (i.e., the groups). For example, during a workout, users could find it more enjoyable to listen to a brief, energetic rock song rather than a piano sonata, even if, in principle, they might like both. Hence, a segmentation driven by multiple features could convey sensible suggestions that go beyond simple genre similarities and personal taste.

All the variables were properly transformed to fit a mixture of multivariate normal distributions: first, the three variables were marginally normalized in the $(0, 1)$ interval; then, they were mapped onto the real line using a probit transformation. Figure 9 displays the three-dimensional data with the help of three pairwise scatterplots. Each point represents a song, and we highlighted the songs authored by four different artists, i.e., 2Pac, AC/DC, D. Carnegie, and S. Rachmaninoff. The four artists belong to fundamentally different genres: the scores of their tracks indeed show different distributional characteristics, although there is an overlap in some of the features. However, an adequate model should be able to recognize their differences and assign them to distinct distributional clusters.

We fit the proposed fiSAN model using the CAVI algorithm outlined in Section 3. Indeed, in Section 4, we have seen how MCMC methods already pose major computational issues when the sample size is $N = 6000$. Because of the fast increase of memory allocation and computational time of MCMC methods as N increases, here, with almost 20,000 observations, a Gibbs sampler approach would require considerable computational resources. Similarly to the procedure adopted in the simulation study, we ran our CAVI algorithm 1000 times using independent random initialization, and we kept the iteration with the highest ELBO to draw inference. In line with the reasoning outlined in Section 4, we fixed the Dirichlet parameters $L = 35$ and $b = 0.05$; the truncation

[2] https://developer.spotify.com/discover/

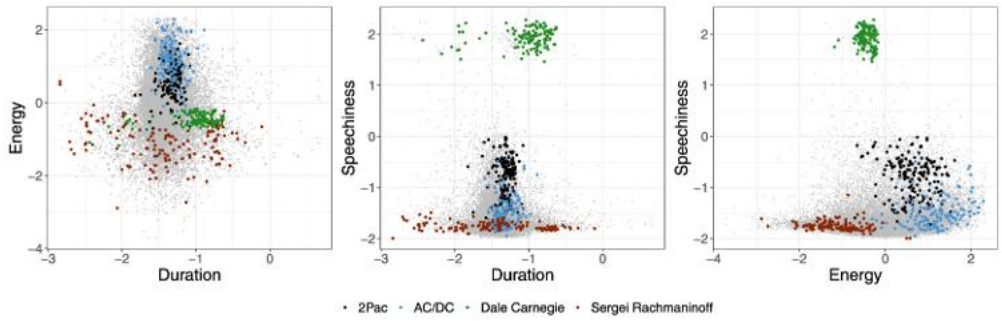

Figure 9: Distribution of the songs (points) across the three features. Colored points correspond to songs by four artists: 2Pac (black), AC/DC (blue), D. Carnegie (green), and S. Rachmaninoff (red).

DC	Artists
`Audio lectures`	D. Carnegie – E. H. Gombrich
`Classical`	A. Copland – A. Scriabin – B. Evans – B. Evans Trio – C. Mingus – C. Baker E. Satie – F. Mendelssohn – F. J. Haydn – F. Liszt – F. Schubert – G. F. Handel G. Mahler – M. Ravel – P. I. Tchaikovsky – R. Strauss – S. Rachmaninoff – S. Getz
`Hard rock`	AC/DC – Aerosmith – blink-182 – Bob Seger – BTS – Def Leppard – Green Day Iron Maiden – Journey – Judas Priest – KISS – Linkin Park – Nirvana – Ramones Rush – The Smiths – Van Halen
`Rap`	2Pac – Beastie Boys – Beyoncé – JAY-Z – Kanye West – Lil Uzi Vert – Lil Wayne Mac Miller – Sublime – The Notorious B.I.G.

Table 1: Artists in the four analyzed distributional clusters.

parameter T of the Dirichlet process at the distributional level was set equal to 30. As for the remaining hyperparameters, they were set as in the simulation studies. On this dataset, the fiSAN estimates 21 OCs and 20 DCs. In what follows, we discuss how we can exploit the estimated two-layer partition to obtain interesting insights.

Analysis of the clusters of artists (DC). Our algorithm groups the 154 artists into 20 clusters. In Table S.2 of the Supplementary Material, we report the complete segmentation of the artists. Here, we analyze four notable DCs, whose members are reported in Table 1. Looking at the members of these clusters, we are able to characterize them according to distinctive "genres", broadly interpreted as "audio lectures", "classical", "hard rock", and "rap". Figure 10 shows the distribution of the three features in these clusters. In terms of duration, all DCs are quite similar. However, tracks in the classical music cluster have a more variable duration than the other groups, and audio lectures last longer, on average. The energy and speechiness features are the ones with the greater heterogeneity across clusters. Hard rock songs have the highest energy, followed by rap songs. However, rap songs have a higher speechiness score. Classical music has low energy and low speechiness, while the audio lectures are, as expected, the most verbose.

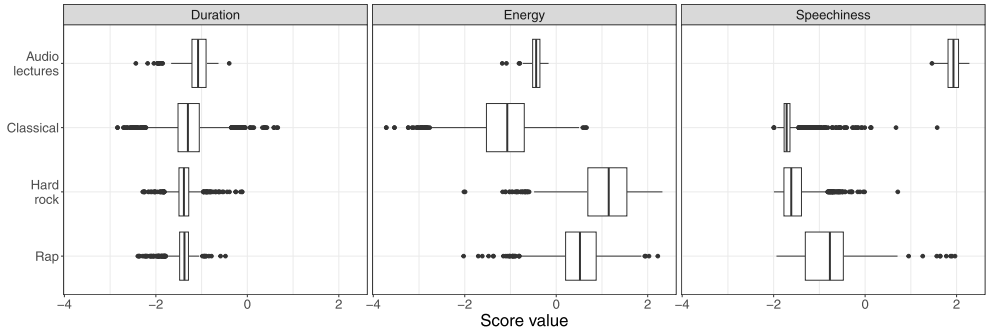

Figure 10: Distribution of the three features in the four analyzed distributional clusters.

Our model-based clustering approach, however, allows for much deeper insights into what drives the segmentation. According to our model, the observational atoms are shared across all DCs; however, in each DC, some of these atoms may be assigned a negligible posterior probability. It is then interesting to examine the active components that ultimately characterize each DC. In Figure 11, we show again the pairwise scatterplots of the three song features. Here, we highlighted the points belonging to the four analyzed DCs. Moreover, for every DC, we also represented all the "active" observational atoms, associated with non-empty observational clusters. Specifically, all observational atoms are represented by their estimated mean $\hat{\boldsymbol{\mu}}_l = (\hat{\mu}_{\mathrm{D}}, \hat{\mu}_{\mathrm{E}}, \hat{\mu}_{\mathrm{S}})_l^\top$ with crosses (2-dimensional subvectors) and covariances $\hat{\boldsymbol{\Sigma}}_l[r, s]$ $(r, s \in \{\mathrm{D}, \mathrm{E}, \mathrm{S}\})$ with ellipses. To favor the interpretation of each DC, we highlighted which atoms are the most relevant by drawing the intensity of the line color as proportional to the posterior weight $\hat{\omega}_{l,\hat{S}_j}$. Regarding the "audio lecture" cluster, we can appreciate that it is defined by a simple distribution made of two mixture components with very heterogeneous posterior weights. The leading OC is characterized by average duration and energy but very high speechiness, while the second one accounts for a few low-duration tracks. The "rap" DC comprises a larger number of active OCs; however, its distribution is fundamentally characterized by two predominant components. Looking at the posterior means of these two observational atoms, we see that they strongly overlap in the energy and duration variables (all rap songs have high energy and average duration). However, the speechiness feature distinguishes them and allows differentiating between more and less verbose tracks. Finally, the "classical" and "hard rock" DCs are more complex and nuanced. In particular, the latter distribution can be described as a mixture of two predominant normal kernels that capture different traits in all three considered dimensions.

Analysis of the clusters of songs (OC). Until now, we have only discussed the distributional clusters. Nonetheless, observational clusters allow for a refined analysis of the similarities between songs. Here, we show how we can take advantage of the shared-atom structure of our model to find songs that have similar characteristics, despite belonging to artists in different DCs. We report an example in the scatterplots in Figure 12. We highlighted songs assigned to the same OC: this cluster contains 2057 songs authored by 63 artists (belonging to 6 different DCs). The red points identify four

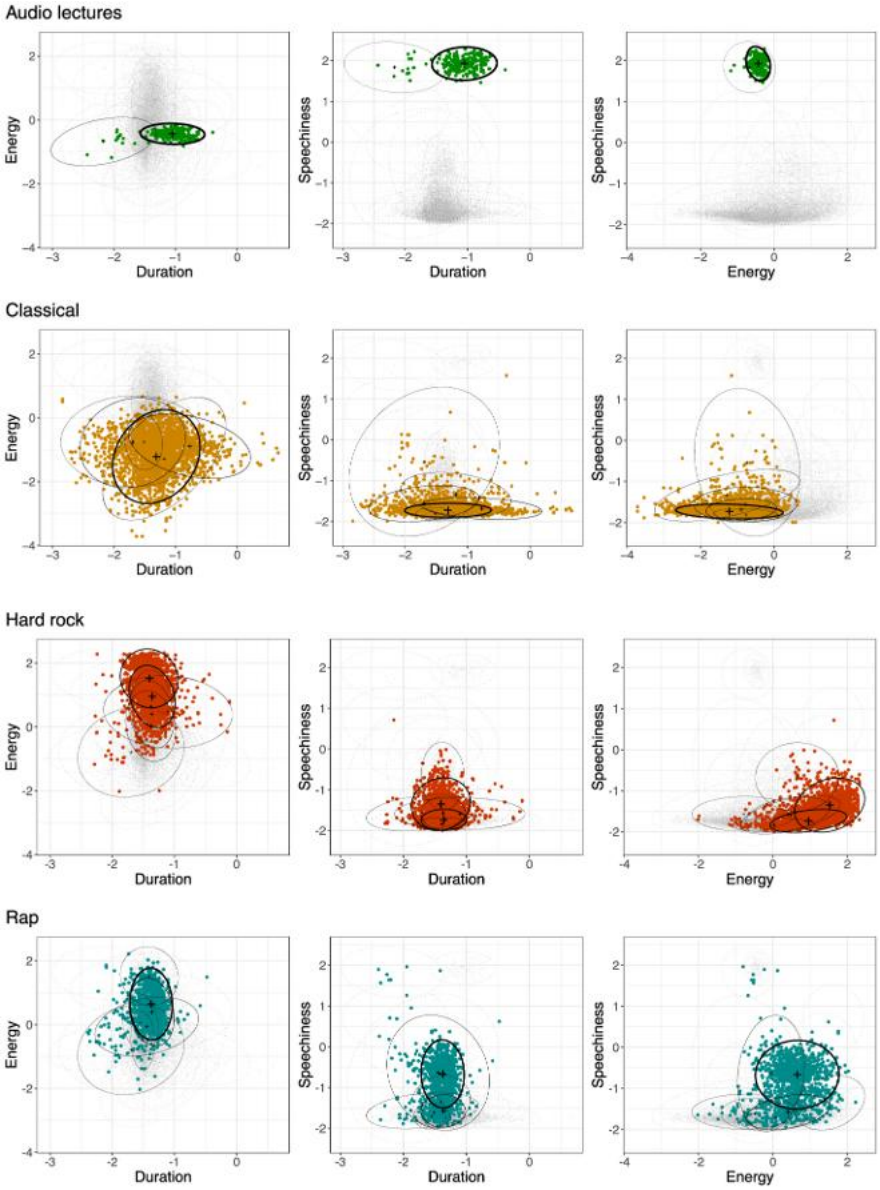

Figure 11: Pairwise scatterplots of songs (points). Each row corresponds to a different DC ("genre"), colored points indicate songs belonging to that DC. In each plot, crosses and ellipses correspond to the active observational atoms, and the intensity of the color is proportional to their posterior weight.

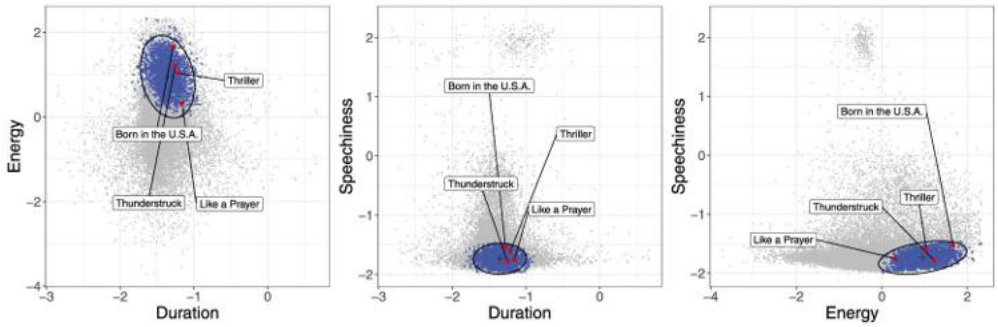

Figure 12: Pairwise scatterplots of song features. Colored (blue) points correspond to songs in one observational cluster, and red points highlight four famous tracks. The black crosses and ellipses indicate the estimated mean and variance/covariance.

famous songs: "Like a Prayer" by Madonna, "Thriller" by Micheal Jackson, "Thunderstruck" by AC/DC, and "Born in the U.S.A." by Bruce Springsteen. Despite belonging to different musical genres (a trait captured by the DC), the scores of these songs are similar, indicating that all these pieces have common characteristics. Specifically, they all have high energy, average duration, and low speechiness.

6 Discussion

In this paper, we developed a novel Bayesian nonparametric model for density estimation and clustering with grouped data, which conveys a two-level partition of groups and observations. The advantages of the proposed modeling framework are threefold. First, it allows shared observational clusters across different random measures; second, it provides a more flexible correlation structure, which induces improved distributional clusters' characterization. Finally, its simple formulation allows the derivation of efficient estimation algorithms to scale up its applicability to large multivariate datasets with thousands of observations and hundreds of groups.

The proposed work stimulates several questions that are worth investigating. The new insights into the interaction between common atoms and stick-breaking weights pave the way for additional research on other nested models with dependent DP (Quintana et al., 2022). In particular, it brings attention back to the possibility of directly exploiting the stochastic ordering of the atoms for modeling purposes, as done, for example, by Griffin and Steel (2006) with the order-based dependent DP. This reasoning is not limited to nested mixtures: nonparametric priors based on common atoms have been used, for example, by Chandra et al. (2023b) in the context of regression in clinical trials.

The advantages of using finite mixtures within the proposed nested setting could be further enhanced by employing more flexible distributions than the Dirichlet. Argiento and de Iorio (2022), for example, introduced a class of priors based on the normalization of a point process which includes the Dirichlet mixture model as a particular case, and

that was recently extended to the hierarchical setting (Colombi et al., 2023). Other finite-dimensional nonparametric priors were proposed by Lijoi et al. (2020) and Lijoi et al. (2023b), which introduced flexible and robust alternatives to the Dirichlet that reduce the influence of the mass parameter and allow for more control over the expected number of clusters and resulting partition.

Another crucial aspect of the proposed framework is its applicability to data with large sample sizes and high dimensionality. We addressed this issue from a computational perspective, implementing an efficient posterior inference algorithm. Additionally, we focused on prior properties that are not affected by the particular choice of the likelihood and base measure, and thus apply also to the multivariate case. However, it could be interesting to study how the particular structure and covariance of the data affect the clustering results in such multivariate settings. Moreover, instead of relying on fast computation, another possible approach could be to leverage models developed for high-dimensional frameworks and extend them to our nested setting. For example, one could induce a sparser partition of the data via repulsive mixtures (see, e.g., Beraha et al., 2022; Ghilotti et al., 2023), or explicitly assume the existence of a set of low-dimensional latent variables that drive the clustering (Chandra et al., 2023a).

Finally, the proposed nested variational algorithm – developed for partially exchangeable data – can be easily extended to other similar frameworks. Stochastic variational inference (Hoffman et al., 2013) solutions might be investigated to grant the immediate applicability of these models to the original Spotify dataset, including hundreds of thousands of observations. Additionally, one could devise a similar finite-infinite model to account for separable exchangeability. For example, our hybrid mixture weight specification, along with a VI approach, could be applied to the common atoms model proposed by Rebaudo et al. (2021), granting an efficient and powerful nonparametric method for matrix biclustering.

Acknowledgments

The authors would like to thank the Editor, the Associate Editor, and the anonymous Reviewers for their suggestions and comments, which significantly improved the exposition and the content of the paper. We would also like to express our gratitude to Prof. Michele Guindani and Prof. Antonio Canale for their constructive comments on an early version of this work. Finally, the authors acknowledge Università Cattolica del Sacro Cuore, where F. D. held the position of Assistant Professor while initially developing this research, for providing the computational resources to conduct part of the research reported in this article.

Supplementary Material

Supplementary Material for "A Finite-Infinite Shared Atoms Nested Model for the Bayesian Analysis of Large Grouped Data Sets" (DOI: 10.1214/24-BA1458SUPP; .pdf). The Supplementary Material provides all proofs of the prior properties; additional details on the relationships between the considered prior distributions; derivation of the algorithms for posterior inference; additional details on the simulation studies and on the data analysis.

References

Agrawal, P., Tekumalla, L. S., and Bhattacharya, I. (2013). "Nested Hierarchical Dirichlet Process for Nonparametric Entity-topic Analysis." *Lecture Notes in Computer Science*, LNAI, volume 8189: 564–579. 2

Argiento, R. and de Iorio, M. (2022). "Is Infinity That Far? A Bayesian Nonparametric Perspective of Finite Mixture Models." *Annals of Statistics*, 50(5): 2641–2663. MR4505373. doi: https://doi.org/10.1214/22-aos2201. 28

Balocchi, C., George, E. I., and Jensen, S. T. (2022). "Clustering Areal Units at Multiple Levels of Resolution to Model Crime in Philadelphia." *arXiv preprint arXiv: 2112.02059v2*, 1–17. MR4595458. doi: https://doi.org/10.1080/01621459.2022.2156348. 2

Beraha, M., Argiento, R., Møller, J., and Guglielmi, A. (2022). "MCMC Computations for Bayesian Mixture Models Using Repulsive Point Processes." *Journal of Computational and Graphical Statistics*, 31(2): 422–435. MR4425075. doi: https://doi.org/10.1080/10618600.2021.2000424. 29

Beraha, M., Guglielmi, A., and Quintana, F. A. (2021). "The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions." *Bayesian Analysis*, 16(4): 1187–1219. MR4381132. doi: https://doi.org/10.1214/21-BA1278. 2

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, volume 4 of *Information Science and Statistics*. Springer, New York. MR2247587. doi: https://doi.org/10.1007/978-0-387-45528-0. 11, 12

Blei, D. M. and Jordan, M. I. (2006). "Variational Inference for Dirichlet Process Mixtures." *Bayesian Analysis*, 1(1): 121–144. MR2227367. doi: https://doi.org/10.1214/06-BA104. 11, 12

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association*, 112(518): 859–877. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 11, 16

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019). "Latent Nested Nonparametric Priors (With Discussion)." *Bayesian Analysis*, 14(4): 1303–1356. MR4044854. doi: https://doi.org/10.1214/19-BA1169. 2, 8, 10

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2017). "Distribution Theory for Hierarchical Processes." *Annals of Statistics*, 47(1): 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 8

Chandra, N. K., Canale, A., and Dunson, D. B. (2023a). "Escaping the Curse of Dimensionality in Bayesian Model-based Clustering." *Journal of machine learning research*, 24(144): 1–42. MR4596091. doi: https://doi.org/10.4995/agt.2023.18320. 29

Chandra, N. K., Sarkar, A., de Groot, J. F., Yuan, Y., and Müller, P. (2023b). "Bayesian Nonparametric Common Atoms Regression for Generating Synthetic Controls in Clin-

ical Trials." *Journal of the American Statistical Association*, 118(544): 2301–2314. MR4681584. doi: https://doi.org/10.1080/01621459.2023.2231581. 28

Colombi, A., Argiento, R., Camerlenghi, F., and Paci, L. (2023). "Mixture Modeling Via Vectors of Normalized Independent Finite Point Processes." *arXiv preprint arXiv: 2310.20376*, 1 – 53. 29

Dahl, D. B., Johnson, D. J., and Müller, P. (2022). "Search Algorithms and Loss Functions for Bayesian Clustering." *Journal of Computational and Graphical Statistics*, 31: 1189–1201. MR4513380. doi: https://doi.org/10.1080/10618600.2022.2069779. 15

D'Angelo, L., Canale, A., Yu, Z., and Guindani, M. (2023). "Bayesian nonparametric analysis for the detection of spikes in noisy calcium imaging data." *Biometrics*, 79(2): 1370–1382. MR4606359. doi: https://doi.org/10.1111/biom.13626. 3, 6, 12

D'Angelo, L. and Denti, F. (2023). *SANple: Fitting Shared Atoms Nested Models via Markov Chains Monte Carlo*. R package, version 0.1.0. URL https://CRAN.R-project.org/package=SANple 15

D'Angelo, L. and Denti, F. (2024). "Supplementary Material for "A Finite-Infinite Shared Atoms Nested Model for the Bayesian Analysis of Large Grouped Data Sets"." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1458SUPP. 3

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 212–229. 5

De Blasi, P., Lijoi, A., and Prünster, I. (2013). "An Asymptotic Analysis of a Class of Discrete Nonparametric Priors." *Statistica Sinica*, 23(3): 1299–1321. MR3114715. 5

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023). "A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data." *Journal of the American Statistical Association*, 118(541): 405–416. MR4571130. doi: https://doi.org/10.1080/01621459.2021.1933499. 2, 4, 7, 12, 14

Denti, F. and D'Angelo, L. (2023). *SANvi: Fitting Shared Atoms Nested Models via Variational Bayes*. R package, version 0.1.0. URL https://CRAN.R-project.org/package=SANvi 15

Escobar, M. D. and West, M. (1995). "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 11

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Generalized Mixtures of Finite Mixtures and Telescoping Sampling." *Bayesian Analysis*, 16(4): 1279–1307. MR4381135. doi: https://doi.org/10.1214/21-BA1294. 3, 5, 6, 14

Ghilotti, L., Beraha, M., and Guglielmi, A. (2023). "Bayesian Clustering of High-dimensional Data via Latent Repulsive Mixtures." *arXiv preprint arXiv:2303.02438*. 29

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs Partitions and Stirling Triangles." *Journal of Mathematical sciences*, 138: 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 5

Gray, R. J. (1994). "A Bayesian Analysis of Institutional Effects in a Multicenter Cancer Clinical Trial." *Biometrics*, 50(1): 244–253. 1

Graziani, R., Guindani, M., and Thall, P. F. (2015). "Bayesian Nonparametric Estimation of Targeted Agent Effects on Biomarker Change to Predict Clinical Outcome." *Biometrics*, 71(1): 188–197. MR3335363. doi: https://doi.org/10.1111/biom.12250. 2

Green, P. J. and Richardson, S. (2001). "Modelling Heterogeneity With and Without the Dirichlet Process." *Scandinavian Journal of Statistics*, 28(2): 355–375. MR1842255. doi: https://doi.org/10.1111/1467-9469.00242. 9

Griffin, J. E. and Steel, M. F. (2006). "Order-based Dependent Dirichlet Processes." *Journal of the American Statistical Association*, 101(473): 179–194. MR2268037. doi: https://doi.org/10.1198/016214505000000727. 28

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). "Stochastic Variational Inference." *Journal of Machine Learning Research*, 14(40): 1303–1347. MR3081926. 29

Hubert, L. and Arabie, P. (1985). "Comparing Partitions." *Journal of Classification*, 2: 193–218. 15

Lee, S. Y., Lei, B., and Mallick, B. (2020). "Estimation of COVID-19 Spread Curves Integrating Global Data and Borrowing Information." *PloS One*, 15(7): 1–17. 1

Li, X., Guindani, M., Ng, C. S., and Hobbs, B. P. (2021). "A Bayesian Nonparametric Model for Textural Pattern Heterogeneity." *Journal of the Royal Statistical Society – Series C*, 70(2): 459–480. MR4226677. doi: https://doi.org/10.1111/rssc.12469. 2

Lijoi, A., Nipoti, B., and Prünster, I. (2014a). "Bayesian Inference with Dependent Normalized Completely Random Measures." *Bernoulli*, 20(3): 1260–1291. MR3217444. doi: https://doi.org/10.3150/13-BEJ521. 8

Lijoi, A., Nipoti, B., and Prünster, I. (2014b). "Dependent Mixture Models: Clustering and Borrowing Information." *Computational Statistics & Data Analysis*, 71: 417–433. MR3131980. doi: https://doi.org/10.1016/j.csda.2013.06.015. 8

Lijoi, A., Prünster, I., and Rebaudo, G. (2023a). "Flexible Clustering Via Hidden Hierarchical Dirichlet Priors." *Scandinavian Journal of Statistics*, 50(1): 213–234. MR4558734. doi: https://doi.org/10.1111/sjos.12578. 2, 7

Lijoi, A., Prünster, I., and Rigon, T. (2020). "The Pitman-Yor multinomial process for mixture modelling." *Biometrika*, 107(4): 891–906. MR4186494. doi: https://doi.org/10.1093/biomet/asaa030. 29

Lijoi, A., Prünster, I., and Rigon, T. (2023b). "Finite-Dimensional Discrete Random

Structures and Bayesian Clustering." *Journal of the American Statistical Association*, 1–13. 29

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based Clustering Based on Sparse Finite Gaussian Mixtures." *Statistics and Computing*, 26: 303–324. MR3439375. doi: https://doi.org/10.1007/s11222-014-9500-2. 5

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 5

Nobile, A. (2004). "On the Posterior Distribution of the Number of Components in a Finite Mixture." *The Annals of Statistics*, 32(5): 2044 – 2073. MR2102502. doi: https://doi.org/10.1214/009053604000000788. 5

Papastamoulis, P. (2016). "label.switching: An R Package for Dealing with the Label Switching Problem in MCMC Outputs." *Journal of Statistical Software*, 69(1): 1–24. 19

Pitman, J. (1995). "Exchangeable and Partially Exchangeable Random Partitions." *Probability Theory and Related Fields*, 102(2): 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 8

Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII*. Lecture Notes in Mathematics N. 1875. Springer, Berlin. 8

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). "The Dependent Dirichlet Process and Related Models." *Statistical Science*, 37(1): 24–41. MR4371095. doi: https://doi.org/10.1214/20-sts819. 28

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66(336): 846–850. 15

Rebaudo, G., Lin, Q., and Müller, P. (2021). "Separate Exchangeability as Modeling Principle in Bayesian Nonparametrics." *arXiv preprint arXiv:2112.07755*, 1–27. 29

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures With an Unknown Number of Components (With Discussion)." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4): 731 – 792. MR1483213. doi: https://doi.org/10.1111/1467-9868.00095. 5

Rodríguez, A. and Dunson, D. B. (2014). "Functional Clustering in Nested Designs: Modeling Variability in Reproductive Epidemiology Studies." *Annals of Applied Statistics*, 8(3): 1416–1442. MR3271338. doi: https://doi.org/10.1214/14-AOAS751. 2

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The Nested Dirichlet Process." *Journal of the American Statistical Association*, 103(483): 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 2, 4, 7

Rodríguez, C. E. and Walker, S. G. (2014). "Label Switching in Bayesian Mixture Models: Deterministic Relabeling Strategies." *Journal of Computational and Graphi-*

cal Statistics, 23(1): 25–45. MR3173759. doi: https://doi.org/10.1080/10618600.2012.735624. 19

Rousseau, J. and Mengersen, K. (2011). "Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models." *Journal of the Royal Statistical Society – Series B*, 73(5): 689–710. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 5, 11, 14

Sethuraman, J. (1994). "A Constructive Definition of Dirichlet Priors." *Statistica Sinica*, 4: 639–650. MR1309433. 4

Stephens, M. (2000). "Dealing With Label Switching in Mixture Models." *Journal of the Royal Statistical Society – Series B*, 62(4): 795–809. MR1796293. doi: https://doi.org/10.1111/1467-9868.00265. 19

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 2

Wade, S. and Ghahramani, Z. (2018). "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)." *Bayesian Analysis*, 13: 559–626. MR3807860. doi: https://doi.org/10.1214/17-BA1073. 15

Wang, C. and Rosner, G. L. (2019). "A Bayesian Nonparametric Causal Inference Model for Synthesizing Randomized Clinical Trial and Real-World Evidence." *Statistics in Medicine*, 38(14): 2573–2588. MR3962129. doi: https://doi.org/10.1002/sim.8134. 2

Zuanetti, D. A., Müller, P., Zhu, Y., Yang, S., and Ji, Y. (2018). "Clustering Distributions With the Marginalized Nested Dirichlet Process." *Biometrics*, 74(2): 584–594. MR3825345. doi: https://doi.org/10.1111/biom.12778. 2