Received: 16 November 2020 | Revised: 11 October 2021 | Accepted: 6 January 2022

DOI: 10.1111/sjos.12578

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Flexible clustering via hidden hierarchical Dirichlet priors

Antonio Lijoi[1] | **Igor Prünster**[1] | **Giovanni Rebaudo**[2]

[1]Department of Decision Sciences and BIDSA, Bocconi University, Milan, Italy

[2]Department of Statistics and Data Sciences, University of Texas at Austin, Austin, Texas, USA

Correspondence
Antonio Lijoi, Department of Decision Sciences and BIDSA, Bocconi University, via Röntgen 1, 20136 Milan, Italy.
Email: antonio.lijoi@unibocconi.it

Funding information
Ministero dell'Istruzione, dell'Università e della Ricerca, Grant/Award Number: 2015SNS29B

[Correction added on 27 May 2022, after first online publication: CRUI funding statement has been added.]

Abstract

The Bayesian approach to inference stands out for naturally allowing borrowing information across heterogeneous populations, with different samples possibly sharing the same distribution. A popular Bayesian nonparametric model for clustering probability distributions is the nested Dirichlet process, which however has the drawback of grouping distributions in a single cluster when ties are observed across samples. With the goal of achieving a flexible and effective clustering method for both samples and observations, we investigate a nonparametric prior that arises as the composition of two different discrete random structures and derive a closed-form expression for the induced distribution of the random partition, the fundamental tool regulating the clustering behavior of the model. On the one hand, this allows to gain a deeper insight into the theoretical properties of the model and, on the other hand, it yields an MCMC algorithm for evaluating Bayesian inferences of interest. Moreover, we single out limitations of this algorithm when working with more than two populations and, consequently, devise an alternative more efficient sampling scheme, which as a by-product, allows testing homogeneity between different populations. Finally, we perform a comparison with the nested

Dirichlet process and provide illustrative examples of both synthetic and real data.

KEYWORDS

Bayesian nonparametrics, clustering, dependent random partitions, hierarchical Dirichlet process, mixture models, nested Dirichlet process, vectors of random probabilities

1 | INTRODUCTION

Dirichlet process (DP) mixtures are well-established and highly successful Bayesian nonparametric models for density estimation and clustering, which also enjoy appealing frequentist asymptotic properties (Escobar, 1994; Escobar & West, 1995; Ghosal & van der Vaart, 2017; Lo, 1984). However, they are not suitable to model data $\{(X_{j,1}, \dots, X_{j,I_j}) : j = 1, \dots, J\}$ that are recorded under J different, though related, experimental conditions. This is due to exchangeability implying a common underlying distribution across populations, a homogeneity assumption which is clearly too restrictive. To make things concrete we consider the Collaborative Perinatal Project (CPP), which is a large prospective epidemiologic study conducted from 1959 to 1974 (analyzed in Section 5.3), where pregnant women were enrolled in 12 hospitals and followed over time. Using a standard DP mixture on the patients enrolled across all 12 hospitals would correspond to ignoring the information on the specific center j where the data are collected and, thus, the heterogeneity across samples. The opposite, also unrealistic, extreme case corresponds to modeling data from each hospital independently, thus ignoring possible similarities among them.

A natural compromise between the aforementioned extreme cases is *partial exchangeability* (de Finetti, 1938), which entails exchangeability within each experimental condition (but not across) and *dependent* population-specific distributions (thus allowing borrowing of information). See Kallenberg (2005) for a detailed account of the topic. In this framework the proposal of dependent versions of the DP date back to the seminal papers of Cifarelli and Regazzini (1978) and MacEachern (1999, 2000). Dependent DPs can be readily used within mixtures leading to several success stories in topic modeling, biostatistics, speaker diarization, genetics, fMRI analysis, and so forth (see Dunson, 2010; Foti & Williamson, 2015; Quintana et al., 2022; Teh & Jordan, 2010 and references therein).

Two hugely popular dependent nonparametric priors, which will also represent the key ingredients of the present contribution, are the hierarchical Dirichlet process (HDP) (Teh et al., 2006) and the nested Dirichlet process (NDP) (Rodríguez et al., 2008). The HDP clusters observations within and across populations. The NDP aims to cluster both population distributions and observations, but as shown in Camerlenghi et al. (2019a), does not achieve this goal. In fact, if there is a cluster of observations shared by different samples, the model degenerates to exchangeability across samples. This issue is successfully overcome in Camerlenghi et al. (2019a) by introducing *latent nested nonparametric priors*. However, while this proposal has the merit of being the first to solve the degeneracy problem, it suffers from other limitations in terms of implementation and modeling: (a) with data from more than two populations the analytical and computational burden implied by the additive structure becomes overwhelming; (b) the model lacks the flexibility needed to capture different weights that common

clusters may feature across different populations. More details can be found in the discussion to Camerlenghi et al. (2019a).

The goal of this paper is thus to devise a principled Bayesian nonparametric approach, which allows to cluster simultaneously distributions and observations (within and across populations). We achieve this by blending peculiar features of both the NDP and the HDP into a model, which we term *Hidden Hierarchical Dirichlet Process* (HHDP). Importantly, the HHDP overcomes the above-mentioned theoretical, modeling, and computational limitations since it, respectively, does not suffer from the degeneracy flaw, is able to effectively capture different weights of shared clusters and allows to handle several populations as showcased in the real data application. Note that the idea of the model was first hinted at in James (2008) and, later, considered in Agrawal et al. (2013) from a mere computational point of view without providing results on distributional properties that are relevant for Bayesian inference. Hence, as a by-product, our theoretical results shed also some light on the topic modeling applications of Agrawal et al. (2013). Additionally, the same model was independently applied in Balocchi et al. (2021) to successfully cluster urban areal units at different levels of resolution simultaneously.

Section 2 concisely reviews the HDP and the NDP with a focus on the random partitions they induce. In Section 3 we define the HHDP and investigate its properties, foremost its clustering structure (induced by a partially exchangeable array of observations). These findings lead to the development of marginal and conditional Gibbs sampling schemes in Section 4. In Section 5 we draw a comparison between HHDP and NDP on synthetic data and present a real data application for our model. Finally, Section 6 is devoted to some concluding remarks and possible future research. Proofs of the results, an additional algorithm and simulation studies are provided in the supplementary material.

2 | BAYESIAN NONPARAMETRIC PRIORS FOR CLUSTERING

The assumption of exchangeability that characterizes widely used Bayesian inferential procedures is equivalent to assuming data homogeneity. This is not realistic in many applied contexts, for instance, for data recorded under J different experimental conditions inducing heterogeneity. A natural assumption that relaxes exchangeability and is suited for arrays of random variables $\{(X_{j,i})_{i\geq 1} : j = 1, \dots, J\}$ is *partial exchangeability*, which amounts to assuming homogeneity within each population, though not across different populations. This is characterized by

$$\{(X_{j,i})_{i\geq 1} : j = 1, \dots, J\} \overset{\mathrm{d}}{=} \{(X_{j,\sigma_j(i)})_{i\geq 1} : j = 1, \dots, J\},$$

for every finitary permutation $\{\sigma_j : j = 1, \dots, J\}$ with $\overset{\mathrm{d}}{=}$ henceforth denoting equality in distribution. Thanks to de Finetti's representation theorem for partially exchangeable arrays, the dependence structure is effectively represented through the following hierarchical formulation

$$\begin{aligned} &X_{j,i}|(G_1, \dots, G_J) \overset{\text{ind}}{\sim} G_j, \qquad (i = 1, \dots, I_j, j = 1, \dots, J). \\ &(G_1, \dots, G_J) \sim \mathcal{L}. \end{aligned} \tag{1}$$

Here we focus on priors $\mathcal{L}$ defined as compositions of discrete random structures and including, as special cases, both the HDP and the NDP. More specifically, we consider $\mathcal{L}$ in (1) that is defined as follows

$$G_j \mid Q \overset{\text{iid}}{\sim} \mathcal{L}(G_j \mid Q) \quad (j = 1, \dots, J); \qquad Q \mid G_0 \sim \mathcal{L}(Q \mid G_0); \qquad G_0 \sim \mathcal{L}(G_0), \tag{2}$$

with discrete random probability measures G_j ($j = 1, \dots, J$), Q and G_0. The data are denoted by $\boldsymbol{X} = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_J\}$ with $\boldsymbol{X}_j = (X_{j,1}, \dots, X_{j,I_j})$ and I_j the size of the jth sample. Discreteness of these random structures entails that with positive probability there are ties within each sample $\boldsymbol{X}_j$ and also across samples $j = 1, \dots, J$, that is, $\mathbb{P}(X_{j,i} = X_{j,\ell}) > 0$ for any $i \neq \ell$, and $\mathbb{P}(X_{j,i} = X_{\kappa,\ell}) > 0$ for any $j \neq \kappa$. Hence, $\boldsymbol{X}$ induces a random partition of the integers $\{1, 2, \dots, n\}$ with $n = I_1 + \cdots + I_J$, whose distribution encapsulates the whole probabilistic clustering of the model and is, therefore, the key quantity to study. Importantly, the random partition can be characterized in terms of the partially exchangeable partition probability function (pEPPF) as defined in Camerlenghi et al. (2019b). The pEPPF is the natural generalization of the concept of exchangeable partition probability function (EPPF) for the exchangeable case (see e.g. Pitman, 2006). More precisely, D is the number of distinct values among the $n = \sum_{j=1}^{J} I_j$ observations in the overall sample $\boldsymbol{X}$. The vector of frequency counts is denoted by $\boldsymbol{n}_j = (n_{j,1}, \dots, n_{j,D})$ with $n_{j,d}$ indicating the number of elements in the jth sample that coincide with the dth distinct value in order of arrival. Clearly, $n_{j,d} \geq 0$ and $\sum_{i=1}^{J} n_{i,d} \geq 1$. One may well have $n_{j,d} = 0$, which implies that the dth distinct value is not recorded in the jth sample, though by virtue of $\sum_{i=1}^{J} n_{i,d} \geq 1$ it must be recorded at least in one of the samples. The dth distinct value is shared by any two samples j and j' if and only if $n_{j,d}\, n_{j',d} \geq 1$. The probability law of the random partition is characterized by the pEPPF defined as

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J) = \mathbb{E} \int_{\mathbb{X}_*^D} \prod_{d=1}^{D} \{G_1(\mathrm{d}x_d)\}^{n_{1,d}} \cdots \{G_J(\mathrm{d}x_d)\}^{n_{J,d}}, \tag{3}$$

with the constraint $\sum_{d=1}^{D} n_{j,d} = I_j$, for each $j = 1, \dots, J$ and where $\mathbb{X}$ is the space in which the $X_{j,i}$'s take values and $\mathbb{X}_*^D$ is the collection of vectors in $\mathbb{X}^D$ whose entries are all distinct. We stress that the expected value in (3) is computed with respect to the joint law of the vector of random probabilities $(G_1, \dots, G_J)$, that is the de Finetti measure $\mathcal{L}$ in (1). Hence, the pEPPF may also be interpreted as a marginal likelihood when $(G_1, \dots, G_J)$ directly model the observations according to (1). Obviously, for a single population, that is $J = 1$, the standard EPPF is recovered and (3) is further interpretable as an extension of a product partition model to a multiple samples framework. As such, it provides an alternative approach to popular covariate-dependent product partition models. See, for example, Müller et al. (2011), Page and Quintana (2016, 2018).

If we specify $\mathcal{L}(\cdot \mid Q)$ and Q such that they give rise to an NDP, then one may have ties also among the population probability distributions $G_1, \dots, G_J$, that is, $\mathbb{P}(G_j = G_\kappa) > 0$ for any $j \neq \kappa$. Therefore, in the framework of (1) and (2), one may investigate two types of clustering: (i) *distributional clustering*, which is related to $G_1, \dots, G_J$ and (ii) *observational clustering*, which refers to $\boldsymbol{X}$. The composition of these two clustering structures is the main tool we rely on to devise a simple, yet effective, model that considerably improves over existing alternatives.

2.1 | Hierarchical Dirichlet process

Probably the most popular nonparametric prior for the partially exchangeable case is the HDP of Teh et al. (2006), which can be nicely framed in the composition scheme (2) as

$$\mathcal{L}(G_j|Q) = \mathrm{DP}(G_j|\beta, Q), \quad \mathcal{L}(Q|G_0) = \delta_{G_0}(Q), \quad \mathcal{L}(G_0) = \mathrm{DP}(G_0|\beta_0; H), \tag{4}$$

where $\mathrm{DP}(\cdot\,|\alpha, P)$ denotes the law of a DP with concentration parameter $\alpha > 0$ and baseline probability measure P. Here we assume that H is a nonatomic probability measure on $\mathbb{X}$ and we refer to such prior as the J-dimensional HDP denoted by $(G_1, \ldots, G_J) \sim \mathrm{HDP}(\beta, \beta_0; H)$. Hence, the G_j's share the atoms through G_0 and this leads to the creation of shared clusters of observations (or latent features) across the J groups. The pEPPF induced by a partially exchangeable array in (1) with $\mathcal{L} = \mathrm{HDP}(\beta, \beta_0; H)$ has been determined in Camerlenghi et al. (2019b). It is important to stress that the model is not suited for comparing populations' distributions since $\mathbb{P}(G_j = G_\kappa) = 0$ for any $j \neq \kappa$ (unless the G_j's are degenerate at G_0, in which case all distributions are equal). Similar compositions have been considered in Camerlenghi et al. (2019b) and, later, in Argiento et al. (2020) and Bassetti et al. (2020). Hierarchically dependent mixture hazards have been introduced in Camerlenghi et al. (2021). Anyhow, the HDP and its variations cannot be used to cluster both populations and observations. To achieve this, one has to rely on priors induced by nested structures, the most popular being the NDP.

2.2 | Nested Dirichlet process

The NDP, introduced by Rodríguez et al. (2008), is the most widely used nonparametric prior allowing to cluster both observations and populations. However, as proved in Camerlenghi et al. (2019a), it suffers from a *degeneracy issue*, because even a single tie shared across samples is enough to group the J population distributions into a single cluster.

Like the HDP, also the NDP can be framed in the composition structure (2) as

$$\mathcal{L}(G_j|Q) = Q(G_j), \quad \mathcal{L}(Q|G_0) = \mathrm{DP}(Q|\alpha; G_0), \quad \mathcal{L}(G_0) = \delta_{\mathrm{DP}(\beta;H)}(G_0), \tag{5}$$

where Q is a random probability measure on the space $\mathcal{P}_{\mathbb{X}}$ of probability measures on $\mathbb{X}$ and G_0 is degenerate at the atom $\mathrm{DP}(\beta; H)$, which is the law of a DP on the sample space $\mathbb{X}$. As in (4), H is assumed to be a nonatomic probability measure on $\mathbb{X}$. Henceforth, we write $(G_1, \ldots, G_J) \sim \mathrm{NDP}(\alpha, \beta; H)$. By virtue of the well-known stick-breaking representation of the DP (Sethuraman, 1994) one has

$$Q = \sum_{k\geq 1} \pi_k^* \delta_{G_k^*}, \quad (\pi_k^*)_{k\geq 1} \sim \mathrm{GEM}(\alpha), \quad G_k^* \overset{\mathrm{iid}}{\sim} \mathrm{DP}(\beta; H), \tag{6}$$

where the weights $(\pi_k^*)_{k\geq 1}$ and the random distributions $(G_k^*)_{k\geq 1}$ are independent. Recall that GEM stands for the distribution of probability weights after Griffiths, Engen, and McCloskey, according to the well-established terminology of Ewens (1990). Given a sequence $(V_i)_{i\geq 1}$ such that $V_i \overset{\mathrm{iid}}{\sim} \mathrm{Beta}(1, \alpha)$, this means that $\pi_1^* = V_1$ and $\pi_k^* = V_k \prod_{i=1}^{k-1}(1 - V_i)$, for any $k \geq 2$. Since $\mathbb{P}(G_j = G_\kappa) = 1/(\alpha + 1)$ for any $j \neq \kappa$, Q generates ties among the random distributions G_j's with positive probability and, thus, clusters populations. Furthermore, a structure similar to the one displayed in (6) holds for each G_k^*, that is,

$$G_k^* = \sum_{l\geq 1} \omega_{k,l} \delta_{X_{k,l}^*}, \quad (\omega_{k,l})_{l\geq 1} \overset{\mathrm{iid}}{\sim} \mathrm{GEM}(\beta), \quad X_{k,l}^* \overset{\mathrm{iid}}{\sim} H,$$

and, due to the nonatomicity of H, the $X_{k,l}^*$ are all distinct values.

The discrete structure of the G_k^*'s generates ties across the samples $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ with positive probability. For example, $\mathbb{P}(X_{j,i} = X_{j',i'}) = 1/\{(\alpha+1)(\beta+1)\}$ for any $j \neq j'$. Hence, the G_k^*'s induce the clustering of the observations $\boldsymbol{X}$.

If the data $\boldsymbol{X}$ are modeled as in (1), with $(G_1, \dots, G_J) \sim \text{NDP}(\alpha, \beta; H)$, conditional on a partition of the G_j's the observations from populations allocated to the same cluster are exchangeable and those from populations allocated to distinct clusters are independent. This potentially appealing feature of the NDP is however the one responsible for the above-mentioned *degeneracy issue*. For exposition clarity, consider the case of $J = 2$ populations. If the two populations belong to different clusters, that is, $G_1 \neq G_2$, they cannot share even a single atom $X_{k,l}^*$ due to the nonatomicity of H. Hence, $\mathbb{P}(X_{1,l} = X_{2,l'} | G_1 \neq G_2) = 0$ for any l and l'. Therefore there is neither clustering of observations nor borrowing of information across different populations. On the contrary, $\mathbb{P}(X_{1,i} = X_{2,i'} | G_1 = G_2) = 1/(\beta+1) > 0$. These two findings are quite intuitive. Indeed, $G_1 \neq G_2$ means they are independent realizations of a DP with atoms iid from the same nonatomic probability distribution H and, thus, they are almost surely different. Instead, $G_1 = G_2$ corresponds to all observations coming from the same population distribution, more precisely from the same DP, and ties occur with positive probability. A less intuitive fact is that when a single atom, say $X_{k,l}^*$, is shared between G_1 and G_2 the model degenerates to the exchangeable case, namely $\mathbb{P}(G_1 = G_2 | X_{1,i} = X_{2,i'}) = 1$ and the two populations have (almost surely) equal distributions. Hence, the NDP is not an appropriate specification when aiming at clustering both populations and observations across different populations. This was shown in Camerlenghi et al. (2019a) where, spurred by this anomaly of the NDP, a novel class of priors named *latent nested processes* (LNP) designed to ensure that $\mathbb{P}(G_1 \neq G_2 | X_{1,i} = X_{2,i'}) > 0$ is proposed. However, while this formally solves the problem, it has computational and modeling limitations. On the one hand, the implementation of LNPs with more than two samples is not feasible due to severe computational hurdles. On the other hand, LNPs have limited flexibility since the weights of the common clusters of observations across different populations are the same. This feature is not suited to several applications and the discussion to Camerlenghi et al. (2019a) provides interesting examples. See also Beraha et al. (2021), Christensen & Ma (2020), Denti et al. (2021), Soriano & Ma (2019) for further stimulating contributions to this literature.

Hence, within the composition structure framework (2), our goal is to obtain a prior distribution able to infer the clustering structure of both populations and observations, which is highly flexible and implementable for a large number of populations and associated samples.

3 | HIDDEN HIERARCHICAL DIRICHLET PROCESS

Our proposal consists in blending the HDP and the NDP in a way to leverage on their strengths, namely clustering data across multiple heterogeneous samples for the HDP and clustering different populations (or probability distributions) for the NDP. More precisely we combine these two models in a structure (2) as

$$\mathcal{L}(G_j | Q) = Q(G_j), \quad \mathcal{L}(Q | G_0) = \text{DP}(Q | \alpha; \text{DP}(\beta; G_0)), \quad \mathcal{L}(G_0) = \text{DP}(G_0 | \beta_0; H).$$

This leads to the following definition.

Definition 1. The vector of random probability measures $(G_1, \dots, G_J)$ is a hidden hierarchical Dirichlet process (HHDP) if

$$G_j|Q \overset{\text{iid}}{\sim} Q, \quad Q = \sum_{k\geq 1} \pi_k^* \delta_{G_k^*}, \quad (\pi_k^*)_{k\geq 1} \sim \text{GEM}(\alpha), \quad (G_k^*)_{k\geq 1} \sim \text{HDP}(\beta, \beta_0; H),$$

with $(\pi_k^*)_{k\geq 1}$ and $(G_k^*)_{k\geq 1}$ independent. In the sequel we write $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$.

In terms of a graphical model, the HHDP can be represented as in Figure 1.

The sequence $(G_k^*)_{k\geq 1}$ acts as a hidden, or latent, component that is crucial to avoid the bug of the NDP, namely clustering of populations when they share some observations. Moreover, by extending (4) to $J = \infty$, it can be more conveniently represented as

$$\begin{gathered} G_k^* = \sum_{l\geq 1} \omega_{k,l}\, \delta_{Z_{k,l}}, \quad Z_{k,l}|G_0 \overset{\text{iid}}{\sim} G_0, \quad G_0 = \sum_{l\geq 1} \omega_{0,l}\, \delta_{X_l^*}, \quad X_\ell^* \overset{\text{iid}}{\sim} H, \\ (\omega_{k,l})_{l\geq 1} \overset{\text{iid}}{\sim} \text{GEM}(\beta), \quad (\omega_{0,l})_{l\geq 1} \sim \text{GEM}(\beta_0), \end{gathered} \tag{7}$$

where independence holds true between the sequences $(\omega_{k,l})_{l\geq 1}$ and $(Z_{k,l})_{l\geq 1}$ and between $(\omega_{0,l})_{l\geq 1}$ and $(X_l^*)_{l\geq 1}$. Combining the stick-breaking representation and a closure property of the DP with respect to grouping, one further has

$$G_k^* = \sum_{l\geq 1} \omega_{k,l}^* \delta_{X_l^*}, G_0 = \sum_{l\geq 1} \omega_{0,l} \delta_{X_l^*},$$

where $((\omega_{k,l}^*)_{l\geq 1}|\boldsymbol{\omega}_0) \overset{\text{iid}}{\sim} \text{DP}(\beta; \boldsymbol{\omega}_0)$, $\boldsymbol{\omega}_0 = (\omega_{0,l})_{l\geq 1} \sim \text{GEM}(\beta_0)$ and $X_l^* \overset{\text{iid}}{\sim} H$, for $l \geq 1$.

In this scheme, the clustering of populations is governed, *a priori*, by the NDP layer Q through $(\pi_k^*)_{k\geq 1} \sim \text{GEM}(\alpha)$. However, the aforementioned degeneracy issue of the NDP, a posteriori, is successfully avoided. The intuition is quite straightforward: unlike for the NDP, the distinct distributions G_k^* in the HHDP are dependent and have a common random discrete base measure

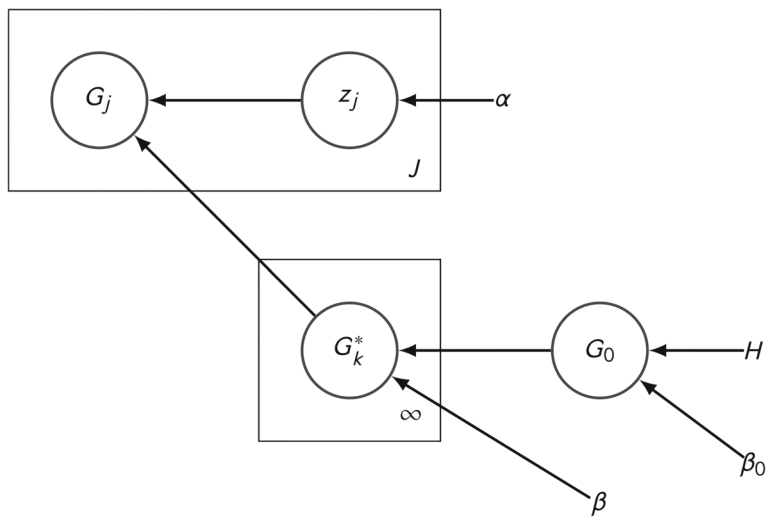

FIGURE 1 Graphical model representing the dependencies for a hidden hierarchical Dirichlet process $(\alpha, \beta, \beta_0; H)$. Here the z_j's are auxiliary integer-valued random variables that assign each G_j to a specific atom G_k^* of Q

G_0, which leads to shared atoms across the G_k^*'s and thus borrowing of information, similarly to the HDP case.

3.1 | Some distributional properties

Given the discreteness of $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$, the key quantity to derive is the induced random partition, which controls the clustering mechanism of the model. However, it is useful to start with a description of pairwise dependence of the elements of the vector $(G_1, \dots, G_J)$, which allows a better understanding of the model and intuitive parameter elicitation. To this end, as customary, we evaluate the correlation between $G_j(A)$ and $G_{j'}(A)$: whenever it does not depend on the specific measurable set $A \subset \mathbb{X}$, it is used as a measure of overall dependence between G_j and $G_{j'}$.

Proposition 1. *If* $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ *and* A *is a measurable subset of* $\mathbb{X}$*, then*

$$\text{Var}[G_j(A)] = \frac{H(A)[1-H(A)](\beta_0+\beta+1)}{(\beta+1)(\beta_0+1)} \qquad (j = 1, \dots, J),$$

$$\text{Corr}[G_j(A), G_{j'}(A)] = 1 - \frac{\alpha\beta_0}{(\alpha+1)(\beta+\beta_0+1)} \qquad (j \neq j').$$

Arguments similar to those in the proof of Proposition 1 lead to determine the correlation between observations, either from the same or from different samples.

Proposition 2. *If* $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ *are from* $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ *according to (1), then*

$$\text{Corr}(X_{j,i}, X_{j',i'}) = \mathbb{P}(X_{j,i} = X_{j,i'}) = \begin{cases} \frac{1}{\beta_0+1} + \frac{\beta_0}{(1+\alpha)(1+\beta)(1+\beta_0)} & (j \neq j') \\ \frac{\beta+\beta_0+1}{(\beta+1)(\beta_0+1)} & (j = j'). \end{cases}$$

The correlation between observations of the same sample depends only on the parameters of the underlying $\text{HDP}(\beta, \beta_0; H)$ that governs the atoms G_k^*: this is not surprising since, whatever the value of the parameter α at the NDP layer, observations from the same sample are exchangeable. Moreover, an appealing feature is that such a correlation is higher than for the case of observations from different samples, that is, $j \neq j'$. As for the dependence on the hyperparameters (α, β_0, β), when $\alpha \to \infty$ the G_js are forced to equal different unique distributions G_k^*, similarly to the NDP case. However, unlike the NDP, this does not imply that the distributions are independent, and the correlation is controlled by the hyperparameters β and β_0 (increasing in β and decreasing in β_0). In Figure 2 we report the aforementioned correlations as functions of β and β_0 with α set equal 1. Finally, if $\alpha \to 0$ the a priori probability to degenerate to the exchangeable case, that is, all G_j's coincide a.s., tends to 1 and so does also $\text{Cor}[G_j(A), G_{j'}(A)]$.

We now investigate the random partition structure associated with a HHDP, namely the partition of $\{1, \dots, n\}$, with $n = \sum_{j=1}^{J} I_j$, induced by a partially exchangeable sample $\boldsymbol{X}$ modeled as in (1). Since a $\text{HHDP}(\alpha, \beta, \beta_0; H)$ arises from the composition of two discrete random structures, it is clear that the partition induced by $\boldsymbol{X}$ will depend on the partition, say $\Psi^{(J)}$, of the random probability measures $G_1, \dots, G_J$. As for the latter, the G_i's are drawn from a discrete random probability measure on $\mathcal{P}_{\mathbb{X}}$ whose weights have a GEM(α) distribution and whose atoms are almost surely different since they are sampled from an $\text{HDP}(\beta, \beta_0; H)$. Then the probability distribution of $\Psi^{(J)}$

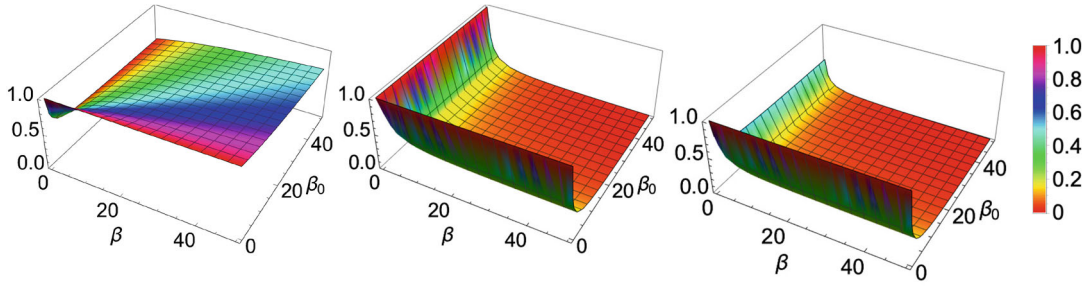

FIGURE 2 Correlations as functions of the hyperparameters β and β_0 with $\alpha = 1$. The left plot represents the correlation between random probabilities $G_j(A)$, the middle one between observations collected in the same population and the right one between observations from different populations

is the well-known Ewens sampling formula, namely

$$\mathbb{P}[\Psi^{(J)} = \{B_1, \dots, B_R\}] = \varphi_R^{(J)}(m_1, \dots, m_R) = \frac{\alpha^R}{\alpha_{(J)}} \prod_{r=1}^{R} (m_r - 1),$$

where $\{B_1, \dots, B_R\}$ is a partition of $\{1, \dots, J\}$, with $1 \le R \le J$, the frequencies $m_r = \text{card}(B_r)$ are such that $\sum_{r=1}^{R} m_r = J$ and $\alpha_{(J)} = \Gamma(\alpha + J)/\Gamma(\alpha)$. This structure a priori implies, as in the NDP case, that $\mathbb{P}(G_j = G_\kappa) \in (0, 1)$ for any $j \neq \kappa$. However, unlike the NDP, a posteriori the HHDP yields $\mathbb{P}(G_j = G_\kappa | \boldsymbol{X}) < 1$, regardless of the shared clusters across the samples $\boldsymbol{X}$. Moreover, let $\Phi_{D,R}^{(n)}(\cdots; \beta, \beta_0)$ denote the pEPPF of a HDP$(\beta, \beta_0; H)$, namely

$$\Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0) = \mathbb{E} \int_{\mathbb{X}_*^D} \prod_{d=1}^{D} \hat{G}_1(\mathrm{d}x_d)^{n_{1,d}^*} \cdots \hat{G}_R(\mathrm{d}x_d)^{n_{R,d}^*},$$

where $(\hat{G}_1, \dots, \hat{G}_R) \sim \text{HDP}(\beta, \beta_0; H)$, $D \in \{1, \dots, n\}$ and $\sum_{r=1}^{R} \sum_{d=1}^{D} n_{r,d}^* = n$. An explicit expression of $\Phi_{D,R}^{(n)}$ has been established in Camerlenghi et al. (2019b), even beyond the DP case. Now we can state the pEPPF induced by $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ in (1), where $\mathcal{L}$ is the law of a HHDP$(\alpha, \beta, \beta_0; H)$.

Theorem 1. *The random partition induced by the partially exchangeable array* $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ *drawn from* $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$*, according to (1), is characterized by the following pEPPF*

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J) = \sum \varphi_R^{(J)}(m_1, \dots, m_R; \alpha) \Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0), \tag{8}$$

where the sum runs over all partitions $\{B_1, \dots, B_R\}$ *of* $\{1, \dots, J\}$ *and* $n_{r,d}^* = \sum_{j \in B_r} n_{j,d}$ *for each* $r \in \{1, \dots, R\}$, $d \in \{1, \dots, D\}$.

Given the composition structure underlying the HHDP$(\alpha, \beta, \beta_0; H)$, the pEPPF (8) unsurprisingly is a mixture of pEPPF's induced by different HDPs. For ease of interpretation consider the case of $J = 2$ populations and note that the pEPPF boils down to

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{1}{\alpha+1}\Phi_{D,1}(\boldsymbol{n}_1+\boldsymbol{n}_2) + \frac{\alpha}{\alpha+1}\Phi_{D,2}(\boldsymbol{n}_1, \boldsymbol{n}_2), \tag{9}$$

where $\Phi_{D,1}^{(n)}$ is the EPPF of a single HDP$(\beta, \beta_0; H)$, namely $J = 1$, while $\Phi_{D,2}^{(n)}$ is the pEPPF of a HDP$(\beta, \beta_0; H)$ with two samples, namely $J = 2$. Clearly (9) arises from mixing with respect to partitions of $\{G_1, G_2\}$ in either $R = 1$ and $R = 2$ groups, where the former corresponds to exchangeability across the two populations. Still for the case $J = 2$, a straightforward application of the pEPPF leads to the posterior probability of gathering the two probability curves, G_1 and G_2, in the same cluster thus making the two samples exchangeable, or homogeneous.

Proposition 3. *If the sample* $\{\boldsymbol{X}_j : j = 1, 2\}$ *is from* $(G_1, G_2) \sim$ HHDP$(\alpha, \beta, \beta_0; H)$*, according to (1), the posterior probability of degeneracy is*

$$\mathbb{P}(G_1 = G_2 | \boldsymbol{X}) = \frac{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1+\boldsymbol{n}_2)}{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1+\boldsymbol{n}_2) + \alpha\ \Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)}, \tag{10}$$

where $\Phi_{D,1}^{(n)}$ *and* $\Phi_{D,2}^{(n)}$ *are the EPPF and the pEPPF induced by the* HDP$(\beta, \beta_0; H)$ *for a single exchangeable sample and for two partially exchangeable samples, respectively*.

The pEPPF is a fundamental tool in Bayesian calculus and it plays, in the partially exchangeable framework, the same role of the EPPF in the exchangeable case. Indeed, the pEPPF governs the learning mechanism, for example, the strength of the borrowing information, clustering, and, in view of Proposition 3, it allows to perform hypothesis testing for distributional homogeneity between populations. Finally, one can obtain a Pólya urn scheme that is essential for inference and prediction. See Section 2 of the supplementary material. In the next section, we provide a characterization of the HHDP$(\alpha, \beta, \beta_0; H)$ that is reminiscent of the popular Chinese restaurant franchise metaphor for the HDP and allows us to devise a suitable sampling algorithm and further understand the model behavior.

3.2 | The hidden Chinese restaurant franchise

The marginalization of the underlying random probability measures, as displayed in Theorem 1, can be characterized in terms of a *hidden Chinese restaurant franchise* (HCRF) metaphor. This representation sheds further light on the HHDP and clarifies the sense in which it generalizes the well-known Chinese restaurant (CRP) and franchise (CRF) processes induced by the DP and the HDP, respectively. For simplicity we consider the case $J = 2$.

As with simpler sampling schemes, all restaurants of the franchise share the same menu, which has an infinite number of dishes generated by the nonatomic base measure H. However, unlike the standard CRF, the restaurants of the franchise are merged into a single one if $G_1 = G_2$, while they differ if $G_1 \neq G_2$. Moreover, each $X_{j,i}$ identifies the label of the dish that customer i from the jth population chooses from the shared menu $(X_d^*)_{d\geq 1}$, with the unique dishes $X_d^* \overset{\text{iid}}{\sim} H$. If $G_1 \neq G_2$, customers may be assigned to different restaurants and when $G_1 = G_2$, they are all seated in the same restaurant. Given such a grouping of the restaurants, the customers are, then, seated according to the CRF applied either to a single restaurant or to two distinct restaurants (Camerlenghi et al., 2018; Teh et al., 2006). Furthermore, each restaurant has infinitely many tables. The first customer i who arrives at a previously unoccupied table chooses a dish that is shared by

all the customers who will join the table afterward. It is to be noted that distinct tables within each restaurant and across restaurants may share the same dish. An additional distinctive feature, compared to the CRF, is that tables can be shared across populations when they are assigned to the same restaurant, that is, when $G_1 = G_2$. Accordingly, the allocation of each customer $X_{j,i}$ to a specific restaurant clearly depends on having either $G_1 = G_2$ or $G_1 \neq G_2$.

The sampling scheme simplifies if latent variables $T_{j,i}$'s, denoting the tables' labels for customer i from population j, are introduced. We stress that, if $G_1 \neq G_2$, the number of shared tables across the two populations is zero, given the populations $j = 1, 2$ are assigned to different restaurants, labeled $r = 1, 2$, respectively. Conversely, if $G_1 = G_2$, one may have shared tables across populations, since they are assigned to the same restaurant $r = 1$.

Now define $q_{r,t,d}$ as the frequencies of observations sitting at table t eating the dth dish, for a table specific to restaurant r. Moreover, D_t is the dish label corresponding to table t and $\ell_{r,d}$ the frequency of tables serving dish d in restaurant r. Marginal frequencies are represented with dots, for example, $\ell_{r,\cdot}$ is the number of tables in restaurant r. Throughout the symbol $\boldsymbol{x}^{-i}$ identifies either a set or a frequency obtained upon removing the element i from $\boldsymbol{x}$. Finally, Δ stands for an indicator function such that $\Delta = 1$ if $G_1 = G_2$, while $\Delta = 0$ if $G_1 \neq G_2$.

The stepwise structure of the sampling procedure reflects the composition of the three layers $\mathcal{L}(G_j|Q)$, $\mathcal{L}(Q|G_0)$ and $\mathcal{L}(G_0)$ in (7) relying on a conditional CRF. First, one sample the populations' clustering Δ and, given the allocations of the populations to the restaurants, one has a CRF. Hence, the algorithm becomes

(1) Sample the population assignments to the restaurants from $\mathbb{P}(\Delta = 1) = 1/(\alpha + 1)$.
(2) Sequentially sample the table assignments $T_{j,i}$ and corresponding dishes $D_{T_{j,i}}$ from

$$p(T_{j,i}, D_{T_{j,i}}|\boldsymbol{T}^{-(ji+)}, \boldsymbol{X}^{-(ji+)}, \Delta) \propto \begin{cases} T_{j,i} = t & \frac{q_{r,t,\cdot}^{-(ji+)}}{q_{r,\cdot,\cdot}^{-(ji+)}+\beta} \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji+)}+\beta} \frac{\ell_{\cdot,d}^{-(ji+)}}{\ell_{\cdot,\cdot}^{-(ji+)}+\beta_0} \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d^{\text{new}} & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji+)}+\beta} \frac{\beta_0}{\ell_{\cdot,\cdot}^{-(ji+)}+\beta_0}, \end{cases}$$

where $(ji+) = \{(ji') : i' \geq i\} \cup \{(j'i') : j' \geq j\}$ is the index set associated to the future random variables not yet sampled.

4 | POSTERIOR INFERENCE FOR HHDP MIXTURE MODELS

Thanks to the results of Section 3, we now devise MCMC algorithms for drawing posterior inferences with mixture models driven by a HHDP. Though the samplers are tailored to mixture models, they are easily adapted to other inferential problems such as, for example, survival analysis and species sampling. Henceforth, $\mathcal{K}$ is a density kernel and we consider

$$\begin{aligned} X_{j,i}|\theta_{j,i} &\overset{\text{ind}}{\sim} \mathcal{K}(\cdot|\theta_{j,i}), \qquad (i = 1, \ldots, I_j \quad j = 1, \ldots, J), \\ \theta_{j,i}|G_j &\overset{\text{ind}}{\sim} G_j, \qquad (i = 1, \ldots, I_j, \quad j = 1, \ldots, J), \\ (G_1, \ldots, G_J) &\sim \text{HHDP}(\alpha, \beta, \beta_0; H). \end{aligned} \tag{11}$$

We develop two samplers: (i) a marginal algorithm that relies on the posterior degeneracy probability (Proposition 3) in Section 2 of the supplementary material; (ii) a conditional blocked Gibbs sampler, in the same spirit of the sampler proposed for the NDP by Rodríguez et al. (2008), in Section 4.1. As for (i), the underlying random probability measures G_0 and G_k^*'s are integrated out leading to urn schemes that extend the class of Blackwell-MacQueen Pólya urn processes. In such a way we generalize the a posteriori sampling scheme of the Chinese restaurant process for the DP mixture (Neal, 2000) and the one of the Chinese restaurant franchise for the HDP mixture (Teh et al., 2006). In the supplementary material, we describe the marginal sampler for the case of $J = 2$ populations. Even if in principle it can be generalized in a straightforward way, it is computationally intractable for a larger number of populations. Similarly to the hidden Chinese restaurant franchise situation, one has to evaluate the posterior probability of all possible groupings of $G_1, \dots, G_J$, which boils down to $\mathbb{P}(G_1 = G_2|\boldsymbol{X})$ when $J = 2$ but becomes involved for $J > 2$.

This shortcoming is overcome by the conditional algorithm we derive in Section 4.1, which relies on finite-dimensional approximations of the trajectories of the underlying random probability measure. Its effectiveness in dealing with $J > 2$ populations is further illustrated in the synthetic data example 5.2 and in the application of Section 5.3.

4.1 | A conditional blocked Gibbs sampler

A more effective algorithm is based on a simple blocked conditional procedure. To this end, we use a finite approximation of the DP in the spirit of Muliere and Tardella (1998) and Ishwaran and James (2001). However, instead of truncating the stick-breaking representation of the DP, we use a finite Dirichlet approximation. See Ishwaran and Zarepour (2002). Therefore, we approximate $\boldsymbol{\pi}^*, \boldsymbol{\omega}_0^*$, with a K- and an L-dimensional Dirichlet distribution, respectively. More precisely, we consider the following approximation

$$\boldsymbol{\pi}^* \sim \mathrm{DIR}(\alpha/K, \dots, \alpha/K), \qquad \boldsymbol{\omega}_0^* \sim \mathrm{DIR}\,(\beta_0/L, \dots, \beta_0/L)\,, \tag{12}$$

implying that $(\boldsymbol{\omega}_k^*|\boldsymbol{\omega}_0^*) \overset{\text{iid}}{\sim} \mathrm{DIR}(\beta\ \boldsymbol{\omega}_0^*)$, for $k \geq 1$.

Introduce the auxiliary variables z_j and $\zeta_{j,i}$ which represent the distributional and observational cluster memberships, respectively, such that $z_j = k$ and $\zeta_{j,i} = l$ if and only if $G_j = G_k^*$ and $\theta_{j,i} = \theta_l^*$. Henceforth, $\boldsymbol{S} = \{(\theta_l^*)_{l=1}^L, \boldsymbol{\pi}^*, \boldsymbol{\omega}_0^*, (\boldsymbol{\omega}_k^*)_{k=1}^K, (z_j)_{j=1}^J, (\zeta_{j,i})_{j,i}, (X_{j,i})_{j,i}\}$ and, in order to identify the full conditionals of the Gibbs sampler, we note that under the finite Dirichlet approximation (12)

$$p(\boldsymbol{S}) = p(\boldsymbol{\pi}^*)p(\boldsymbol{\omega}_0^*)\left[\prod_{l=1}^{L} p(\theta_l^*)\right]\left[\prod_{k=1}^{K} p(\boldsymbol{\omega}_k^*|\boldsymbol{\omega}_0^*)\right]\left\{\prod_{j=1}^{J} p(z_j|\boldsymbol{\pi}^*)\left[\prod_{i=1}^{I_j} p(X_{j,i}|\theta_{\zeta_{j,i}}^*)p(\zeta_{j,i}|\boldsymbol{\omega}_{z_j}^*)\right]\right\}.$$

This leads to the following

(1) Sample the unique θ_l^* from

$$p(\theta_l^*|\boldsymbol{S}^{-\theta_l^*}) \propto H(\theta_l^*) \prod_{\{j,i:\zeta_{j,i}=l\}} \mathcal{K}(X_{j,i}|\theta_l^*).$$

(2) Sample distributional cluster probabilities from

$$p(\boldsymbol{\pi}^*|\boldsymbol{S}^{-\boldsymbol{\pi}^*}) = \text{DIR}(\boldsymbol{\pi}^*|\alpha/K + m_1, \dots, \alpha/K + m_K),$$

with $m_k = \sum_{j=1}^{J} \mathbb{1}\{z_j = k\}$.

(3) Sample probability weights of the base DP from

$$p(\boldsymbol{\omega}_0^*|\boldsymbol{S}^{-\boldsymbol{\omega}_0^*}) \propto \prod_{l=1}^{L} \left[\frac{(\omega_{0,l}^*)^{\beta_0/L-1} \xi_l^{\beta\omega_{0,l}^*}}{\Gamma(\beta_0\omega_{0,l}^*)^K} \right], \tag{13}$$

with $\xi_l = \prod_{k=1}^{K} \omega_{k,l}^*$.

(4) Sample the observational cluster probabilities independently from

$$p(\boldsymbol{\omega}_k^*|\boldsymbol{S}^{-\boldsymbol{\omega}_k^*}) = \text{DIR}(\boldsymbol{\omega}_k^*|\beta\boldsymbol{\omega}_0^* + \boldsymbol{n}_k),$$

with $n_{k,l} = \sum_{\{j:z_j=k\}} \sum_{i=1}^{I_j} \mathbb{1}\{\zeta_{j,i} = l\}$.

(5) Sample distributional and observational cluster membership from

$$p(z_j = k|\boldsymbol{S}^{-\{z_j,\zeta_j\}}) \propto \pi_k^* \prod_{i=1}^{I_j} \sum_{l=1}^{L} \omega_{k,l}^* \mathcal{K}(X_{j,i}|\theta_l^*) \qquad (k = 1, \dots, K),$$

$$p(\zeta_{j,i} = l|\boldsymbol{S}^{-\zeta_{j,i}}) \propto \omega_{z_j l}^* \, \mathcal{K}(X_{j,i}|\theta_l^*) \qquad (l = 1, \dots, L).$$

Importantly, all the full conditional distributions are available in simple closed forms, with the exception of the distributions of $\boldsymbol{\omega}_0^*$ and, possibly, of θ_l^*. To update $\boldsymbol{\omega}_0^*$ we perform a Metropolis-Hastings step, where we work on the unconstrained space $\mathbb{R}^{L-1}$ after the transformation $[\log(\omega_{0,1}/\omega_{0,L}), \dots, \log(\omega_{0,L-1}/\omega_{0,L})]$ and we adopt a component-wise adaptive random walk proposal following Roberts and Rosenthal (2009). The update of the unique atoms θ_l^* is standard, as with the DP mixture model in the exchangeable case.

In Section 5 we assume a Gaussian kernel $\mathcal{K}(\cdot|\theta) = \text{N}(\cdot|\mu, \sigma^2)$ and a conjugate Normal-inverse-Gamma base measure $H(\cdot) = \text{NIG}(\cdot|\mu_0, \lambda_0, s_0, S_0)$ and obtain

$$p(\theta_l^*|\boldsymbol{S}^{-\theta_l^*}) = \text{NIG}(\theta_l^*|\mu_l, \lambda_l, s_l, S_l),$$

with $\mu_l = \frac{n_l\bar{y}_l + \lambda_0\mu_0}{\lambda_0 + n_l}$, $S_l = S_0 + \frac{1}{2}\left(e_l^2 + \frac{n_l\lambda_0(\bar{y}_l - \mu_0)^2}{\lambda_0 + n_l}\right)$, $\lambda_l = \lambda_0 + n_l$, and $s_l = n_l/2 + s_0$, where $n_l = \sum_{j=1}^{J}\sum_{i=1}^{I_j} \mathbb{1}\{\zeta_{j,i} = l\}$, $\bar{y}_l = \sum_{\{j,i:\zeta_{j,i}=l\}} X_{j,i}/n_l$, and $e_l^2 = \sum_{\{j,i:\zeta_{j,i}=l\}} (X_{j,i} - \bar{y}_l)^2$ are the observational cluster sizes, means and deviances, respectively.

5 | ILLUSTRATION

In this section, we compare the performance of our proposal (11) with the same model where the HHDP is replaced by a NDP as in (5), on synthetic data involving $J = 2$ and $J = 4$ populations. Note that for the latter, the implementation of the latent nested prior process mixture of Camerlenghi et al. (2019a) is not feasible, while the proposed HHDP mixture model can easily handle

that level of complexity. The inferential results that we display are obtained by relying on the blocked Gibbs sampler of Section 4.

5.1 | Inference with two populations

The data are simulated from the same scenarios considered in Camerlenghi et al. (2019a). More precisely, we consider two populations and the data in each population are iid from a mixture of two normals:

Scenario 1. We simulate the data from the two populations independently from the same density

$$X_{1,i} \overset{\text{d}}{=} X_{2,i'} \overset{\text{iid}}{\sim} 0.5\text{N}(0,1) + 0.5\text{N}(0,1).$$

Scenario 2. We simulate the data in the two populations independently from mixtures of two normals with one shared component

$$X_{1,i} \overset{\text{iid}}{\sim} 0.9\text{N}(5,0.6) + 0.1\text{N}(10,0.6) \qquad X_{2,i'} \overset{\text{iid}}{\sim} 0.1\text{N}(5,0.6) + 0.9\text{N}(0,0.6).$$

Scenario 3. We simulate the data in the two populations independently from mixtures of two normals having the same components, though with different weights

$$X_{1,i} \overset{\text{iid}}{\sim} 0.8\text{N}(5,1) + 0.2\text{N}(0,1) \qquad X_{2,i'} \overset{\text{iid}}{\sim} 0.2\text{N}(5,1) + 0.8\text{N}(0,1).$$

In all these scenarios we consider balanced sample sizes $I_1 = I_2 = 100$ and an HHDP mixture model (11), with $\alpha = 1$, $\beta = 1$, $\beta_0 = 1$ and $H(\cdot) = \text{NIG}(\cdot|\mu_0, \lambda_0, s_0, S_0)$. We set standard values of the hyperparameters in terms of the mean $\bar{y}$ and variance $\text{Var}(y)$ of the data, that is, $\mu_0 = \bar{y}$, $\lambda_0 = 1/(3\ \text{Var}(y))$, $s_0 = 1$ and $S_0 = 4$. In drawing the comparison between (11) and the NDP$(\alpha, \beta; H)$, we further set $\alpha = \beta = 1$. Furthermore, we set the concentration parameters all equal to 1. In Section 3 we perform a sensitivity analysis with respect to hyperparameters' specifications as done, for instance, by Zuanetti et al. (2018) for the NDP. The mean measure of the marginal underlying random distributions $\mathbb{E}[G_j(A)] = H(A)$ is the same for all populations. Also variances are comparable (see Proposition 1) since $\text{Var}[G_j(A)]$ equals $H(A)[1 - H(A)]/2$ for the NDP and $3H(A)[1 - H(A)]/4$ for the HHDP. The sensitivity analysis leads, for all the considered settings, to the same conclusions in terms of comparison of the two models. Moreover, we fix the dimensions of the finite approximations $L = K = 50$ in (12) and we do the same for the truncation levels in the algorithm of Rodríguez et al. (2008). In the supplementary material, we perform an empirical analysis trying different levels of L and K which corroborates the fact that the approximation error is negligible in terms of inferential results.

Inference is based on 10,000 iterations with the first half discarded as burn-in. As for the output, besides obtaining density estimates for the two populations we also determine the point estimate of the clustering of observations that minimizes the variation of information (VI) loss function. See Meilă (2007) and Wade and Ghahramani (2018) for detailed discussions on VI and point summaries of probabilistic clustering. Additionally, we estimate the probability that observations co-cluster, namely $\mathbb{P}(\zeta_{j,i} = \zeta_{j',i'}|\boldsymbol{X})$ through the average over MCMC draws

$$\frac{\sum_{b=1}^{B} \mathbb{1}\{\zeta_{j,i}^{b} = \zeta_{j',i'}^{b}\}}{B},$$

where B is the number of MCMC iterations. These are visualized through heatmaps as in Figure 4, with colors ranging from white, if the probability is 0, to dark red, if the probability is 1. Our analysis is completed by reporting the estimated distributions of the numbers of mixture components in each scenario.

As expected, both models yield accurate estimates of the true densities in all scenarios. In Figure 3 we report the true and estimated models under the third scenario. In terms of clustering, in the first scenario both models correctly cluster together the two populations, thus degenerating to the exchangeable case as they should. However, in the second and third scenarios the NDP makes the two samples $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ independent, therefore preventing borrowing of information across the two populations. As the distributions have a shared component, the only way for the NDP to recover correctly the true densities is by missing such a component. Had it been detected, the density estimates of the two populations would have been equal and, thus, far from the truth. The point estimate of the observations' clustering in Table 1, the heatmaps of the posterior co-clustering probabilities in Figure 4 and the posterior distributions of the overall number of occupied components in Table 2 showcase the theoretical findings, namely that the NDP in the second and third scenarios cannot learn the shared components. Hence, it overestimates the total number of occupied components and does not cluster observations across populations. In contrast, the HHDP model is able to cluster observations across populations, learns the shared components and borrows information also when the model does not degenerate to the exchangeable case.

5.2 | Inference with more than two populations

Here we consider $J = 4$ populations and deal with the same scenario discussed in Beraha et al. (2021). More precisely, we simulate independently across populations $I_j = 100$ (for $j = 1, \dots, 4$) observations as follows

$$X_{1,i} \overset{\mathrm{d}}{=} X_{2,i} \overset{\mathrm{iid}}{\sim} 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(5,1) \quad X_{3,i} \overset{\mathrm{iid}}{\sim} 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(-5,1) \quad X_{4,i} \overset{\mathrm{iid}}{\sim} 0.5\mathrm{N}(-5,1) + 0.5\mathrm{N}(5,1)$$

Our prior corresponds to a Gaussian mixture model with the same specification for the HHDP used in the previous Section with $J = 2$ population. Figure 5 shows that the HHDP mixture model

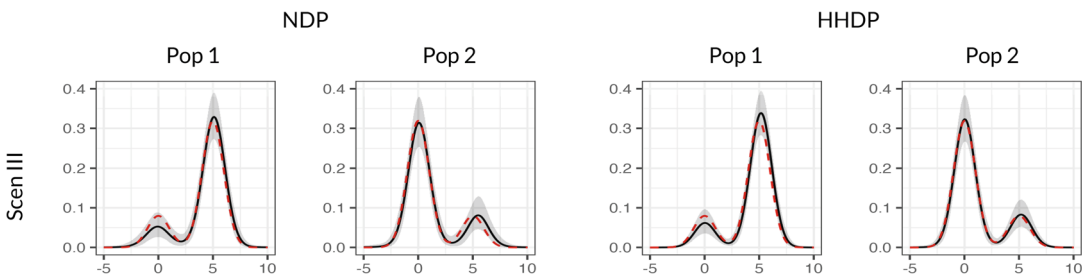

FIGURE 3 True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under the third scenario

TABLE 1 Frequencies of observations in the two populations allocated to the point estimate of the clustering that minimizes the variation of information loss with the two models under different scenarios

	Scenario I				Scenario II							Scenario III					
	NDP		HHDP		NDP				HHDP			NDP				HHDP	
Population	**1**	**2**	**1**	**2**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**
1	56	44	56	44	87	13	0	0	87	13	0	85	15	0	0	85	15
2	48	52	48	52	0	0	88	12	12	0	88	0	0	80	20	21	79

Abbreviations: HHDP, hidden hierarchical Dirichlet process; NDP, nested Dirichlet process.

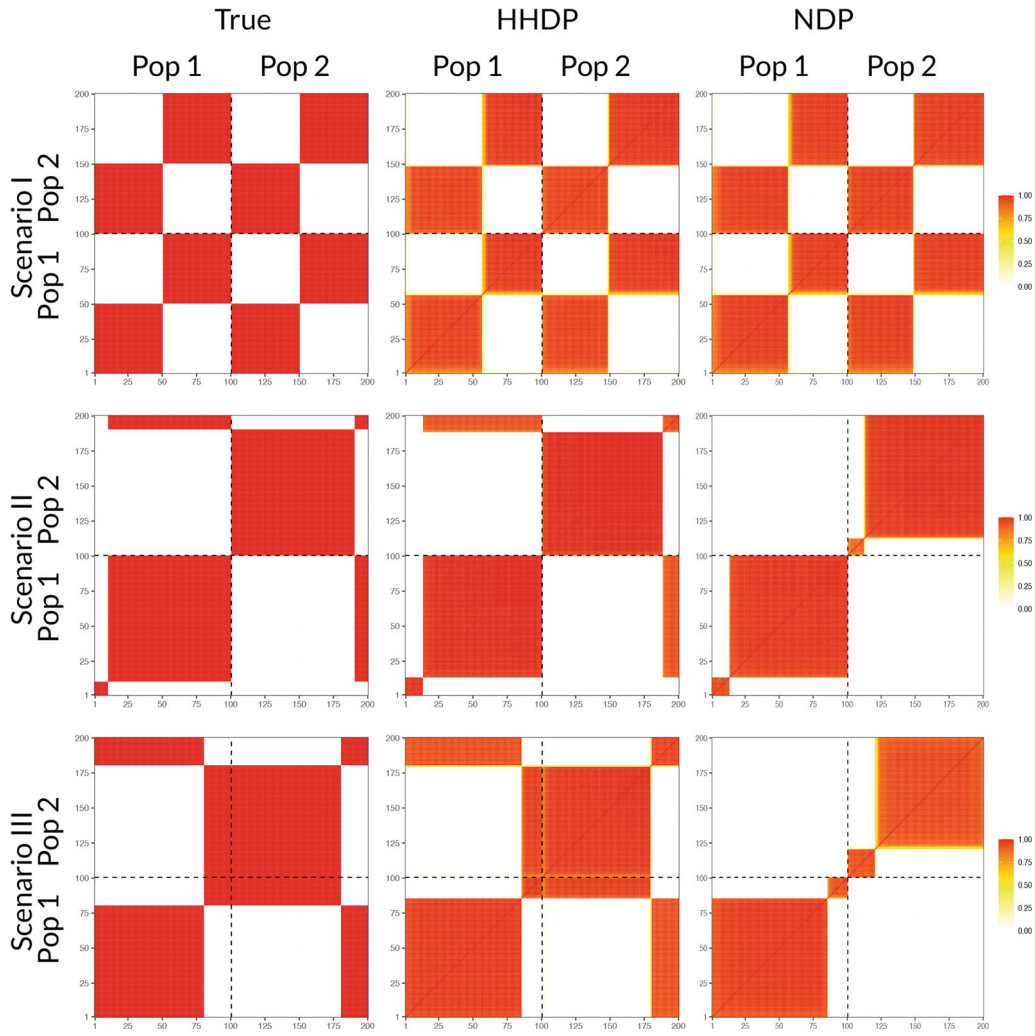

FIGURE 4 Heatmaps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, under the hidden hierarchical Dirichlet process and the nested Dirichlet process models, for the three different scenarios in Section 5.1

TABLE 2 Posterior distributions of the number of overall occupied components estimated with the two models under different scenarios

		Overall number of components									
Scenario	**Model**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**≥10**
I	NDP	0	0.4090	0.3615	0.1647	0.0492	0.0136	0.0020	0	0	0
	HHDP	0	0.5374	0.3743	0.0788	0.0080	0.0016	0	0	0	0
II	NDP	0	0	0	0.2959	0.3906	0.2151	0.0700	0.0256	0.0024	0.0004
	HHDP	0	0	0.5742	0.3339	0.0796	0.0116	0.0008	0	0	0
III	NDP	0	0	0	0.1331	0.3055	0.2947	0.1743	0.0608	0.0232	0.0084
	HHDP	0	0.5010	0.3966	0.0856	0.0164	0.0004	0	0	0	0

Abbreviations: HHDP, hidden hierarchical Dirichlet process; NDP, nested Dirichlet process.

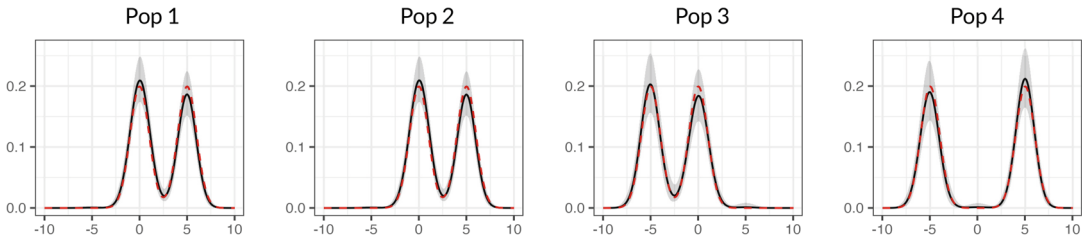

FIGURE 5 True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under the fourth scenario

is able to recover the data generating densities also in this scenario. In terms of clustering of populations the point estimate that minimizes the VI loss coincides with the data generating truth. Figure 6 reports the heatmaps of the posterior co-clustering probabilities of the four populations that show little uncertainty around the correct point estimate, for example the estimated probability that populations 1 and 2 are correctly clustered together is 0.9858.

Finally, the point estimate of the observations' clustering in Table 3 shows the HHDP model is able to cluster observations across populations, learns the shared components and borrows information also when there are more than two populations.

5.3 | CPP data

A multicenter application is the focus of this section. We consider a dataset from the CPP, a large prospective epidemiologic study conducted from 1959 to 1974. Pregnant women were enrolled in 12 hospitals between 1959 and 1966 and were followed over time. Among several prepregnancy measurements, we focus on the birth weight $X_{j,i}$ for nonsmoking woman i in center j. We assume the following Gaussian mixture model:

$$X_{j,i}|\mu_{j,i},\sigma_{j,i} \overset{\text{ind}}{\sim} \mathrm{N}(\mu_{j,i},\sigma_{j,i}) \qquad (i=1,\dots,I_j, \quad j=1,\dots,12),$$
$$\mu_{j,i},\sigma_{j,i}|G_j \overset{\text{ind}}{\sim} G_j \qquad (i=1,\dots,I_j, \quad j=1,\dots,12).$$

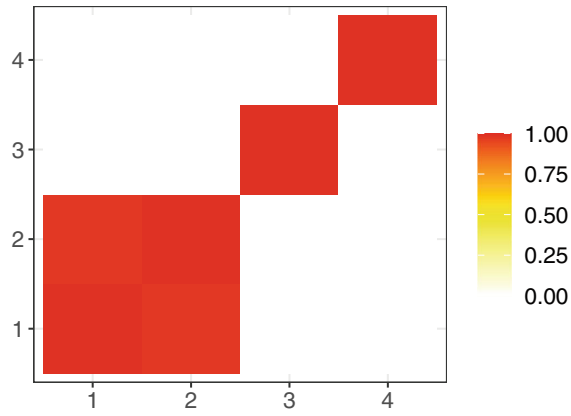

FIGURE 6 Heatmap of the estimated posterior probabilities of co-clustering of the population estimated with the hidden hierarchical Dirichlet process mixture model under the fourth scenario in Section 5.2

TABLE 3 Frequencies of observations in the four populations allocated to the point estimate of the clustering that minimizes the variation of information loss with hidden hierarchical Dirichlet process under the fourth scenario

Observational cluster	1	2	3
Pop 1	53	47	0
Pop 2	56	44	0
Pop 3	48	0	52
Pop 4	0	52	48

The same HHDP prior used for the previous synthetic data is placed the vector of random distributions. This model specification is coherent with what is suggested by Dunson (2010) for the CPP data. Indeed, it is known that the pregnancy outcome can vary substantially for women from different ethnicity and socioeconomic groups. Therefore, we specify a model allowing to capture differences between the centers since different groups of hospitals can serve different women. Canale et al. (2019) provide further analysis of the CPP data.

The heatmap of the co-clustering posterior probability for the 12 hospitals is shown in Figure 7. Such probabilities imply that the clustering point estimate of the hospitals that minimizes the VI loss has two blocks and, in the same figure, the mean posterior densities associated with the two clusters are reported. Given the partition of the hospitals, the posterior mean densities are evaluated based on all patients belonging to hospitals in each of the two partition groups. The heatmap shows the posterior distribution of the clustering of the hospitals and can be used to perform uncertainty quantification. As expected, the lack of well-separated data generating mixtures of Gaussians entails more uncertainty around the point estimate of the clustering of the populations with respect to the numerical experiments. However, the heatmap shows that the point estimate of the clustering of distributions is a reliable summary. More precisely, the point estimate that minimizes the VI loss entails that the first cluster of hospitals includes the hospitals with (reordered) labels 1,2,3: these are well-separated from the remaining hospitals

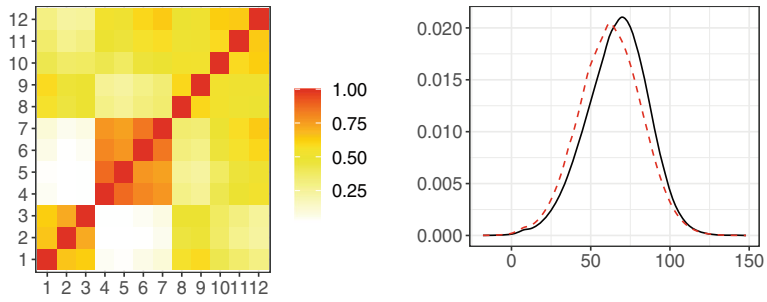

FIGURE 7 Heatmap of the estimated posterior probability of co-clustering of hospitals and estimated population cluster-specific posterior densities for the Collaborative Perinatal Project data

TABLE 4 Posterior distributions of the number of clusters shared and not shared across the two clusters of hospitals

Number of observational clusters	0	1	2	3	4	5
Only in the second cluster of hospitals	0.3530	0.3670	0.2040	0.0640	0.0100	0.0020
Only in the first cluster of hospitals	0.7750	0.1850	0.0340	0.0060	0	0
Shared across clusters of hospitals	0	0.1680	0.4800	0.2660	0.0780	0.0080

according to the posterior probabilities of co-clustering in the heatmap. The heatmap shows also that another meaningful point estimate of the clustering of the hospitals is the finer partition $\{\{1,2,3\},\{4,5,6,7\},\{8,9,10,11,12\}\}$. However, the VI loss suggests a more parsimonious clustering of the hospitals in two blocks, that is $\{\{1,2,3\},\{4,5,6,7,8,9,10,11,12\}\}$. Note that in the second cluster of hospitals (red dashed density in Figure 7) the distribution of the birth weights is slightly shifted on lower values and the two mean densities are similar in the two clusters of populations. Coherently the proposed model allows to borrow information across clusters of hospitals for estimating the posterior mean densities of the birth weights. Furthermore the model can be used to identify clusters of women shared in the two different clusters of hospitals. Indeed, Table 4 shows that some clusters of observations are shared across different clusters of hospitals, thus allowing the borrowing of information for estimating the densities of the birth weights in the two groups.

6 | DISCUSSION

As highlighted in the recent literature, NDP mixture models are often not an appropriate tool for clustering simultaneously population distributions and observations. In contrast, the HHDP, overcomes the issues plaguing the NDP, while preserving tractability and clustering flexibility even when the number of populations J is larger than 2. We have further devised sampling schemes allowing for efficient inference and prediction. This work paves the way for future intriguing research directions that we plan to address in forthcoming work. First, it is natural to move beyond DPs and consider models based on alternative discrete nonparametric priors, such as the Pitman–Yor process and normalized completely random measures, while studying the induced clustering. The characterization of the HHDP in terms of the induced random partition

suggests a nice connection of our work with recent and exciting advances on time-dependent random partition models such as those proposed, for example, in Page et al. (2022) and Zanini et al. (2019). Indeed, these papers define a general framework that can be tailored to HHDP priors for generating time-dependent models suited for analyzing, for example, longitudinal data thus allowing for the investigation of the joint evolution of observational and distributional clustering through time. The theory we have developed in Sections 3 and 4 provides the necessary tools for successfully carrying out such a program. Moreover, the general composition scheme, where we have embedded the HHDP, seems a promising and effective approach for addressing other interesting inferential problems, beyond density estimation and clustering. Finally, the general scheme that we have introduced in (2) seems an appropriate specification for capturing the inherent complexity and heterogeneity of data that arise when drawing predictions with multivariate species sampling models and when performing inferences in survival and functional data analysis. These will be the object of forthcoming work.

ACKNOWLEDGMENTS

The authors are grateful to two Referees for their insightful comments and suggestions that have led to a significant improvement of the manuscript. Most of the paper was completed while G. Rebaudo was a PhD student at Bocconi University. Open Access Funding provided by Universita Bocconi within the CRUI-CARE Agreement.

ORCID

Antonio Lijoi https://orcid.org/0000-0001-6159-2650
Igor Prünster https://orcid.org/0000-0003-2860-1476
Giovanni Rebaudo https://orcid.org/0000-0003-4619-9302

REFERENCES

Agrawal, P., Tekumalla, L. S. & Bhattacharya, I. (2013). Nested Hierarchical Dirichlet process for nonparametric entity-topic analysis. Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (Vol. 8189, pp. 564-579). LNAI, Springer, Berlin, Heidelberg.

Argiento R., Cremaschi A., Vannucci M. (2020). Hierarchical Normalized Completely Random Measures to Cluster Grouped Data. *Journal of the American Statistical Association*, *115*, (529), 318–333.

Balocchi, C., George, E. I., & Jensen, S. T. (2021). Clustering areal units at multiple levels of resolution to model crime incidence in Philadelphia. *Preprint arXiv: 2112.02059*.

Bassetti F., Casarin R., Rossini L. (2020). Hierarchical Species Sampling Models. *Bayesian Analysis*, *15*, (3), 809–838.

Beraha M., Guglielmi A., Quintana F. A. (2021). The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions. *Bayesian Analysis*, *16*, (4), 1187–1219.

Camerlenghi F., Dunson D. B., Lijoi A., Prünster I., Rodríguez A. (2019a). Latent Nested Nonparametric Priors (with Discussion). *Bayesian Analysis*, *14*, (4), 1303–1356.

Camerlenghi F., Lijoi A., Orbanz P., Prünster I. (2019b). Distribution theory for hierarchical processes. *The Annals of Statistics*, *47*, (1), 67–92.

Camerlenghi F., Lijoi A., Prünster I. (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics*, *45*, (4), 1062–1091.

Camerlenghi F., Lijoi A., Prünster I. (2021). Survival analysis via hierarchically dependent mixture hazards. *The Annals of Statistics*, *49*, (2), 863–884.

Canale, A., Corradin, R., & Nipoti, B. (2019). Importance conditional sampling for Bayesian nonparametric mixtures. *Preprint at arXiv: 1906.08147*.

Christensen J., Ma L. (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*, (1), 127–153.

Cifarelli, D. M., & Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambiabilita parziale e impiego di medie associative. *Quaderni Istituto Matematica Finanziaria dell'Universita di Torino*.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles*, *739*, 5–18.

Denti F., Camerlenghi F., Guindani M., Mira A. (2021). A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. *Journal of the American Statistical Association*.

Dunson, D. B. (2010). *Nonparametric Bayes applications to biostatistics*. In N. L. Hjort, C. Holmes, P. Müller & S. G. Walker, *Bayesian Nonparametrics* (pp. 223–273). Cambridge University Press.

Escobar M. D. (1994). Estimating Normal Means with a Dirichlet Process Prior. *Journal of the American Statistical Association*, *89*, (425), 268–277.

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

Ewens, W. J. (1990). *Population genetics theory - the past and the future*. In S. Lessard, *Mathematical and statistical developments of evolutionary theory* (pp. 177–227). Springer.

Foti N. J., Williamson S. A. (2015). A Survey of Non-Exchangeable Priors for Bayesian Nonparametric Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*, (2), 359–371.

Ghosal, S., & van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

Ishwaran H., James L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, *96*, (453), 161–173.

Ishwaran H., Zarepour M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, *30*, (2), 269–283.

James, L. (2008). Discussion of nested Dirichlet process paper by Rodríguez, Dunson and Gelfand. *Journal of the American Statistical Association*, *483*, 1131.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles*. Springer.

Lo A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, *12*, (1), 351–357.

MacEachern, S. N. (1999). *Dependent nonparametric processes*. In *ASA proceedings of the section on Bayesian statistical science* (Vol. *Sci*, pp. 50–55). American Statistical Association.

MacEachern, S. N. (2000). Dependent Dirichlet processes (Technical report). The Ohio State University.

Meilă M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, *98*, (5), 873–895.

Muliere P., Tardella L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, *26*, (2), 283–297.

Müller P., Quintana F., Rosner G. L. (2011). A Product Partition Model With Regression on Covariates. *Journal of Computational and Graphical Statistics*, *20*, (1), 260–278.

Neal R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *9*, (2), 249–265.

Page G. L., Quintana F. A. (2016). Spatial Product Partition Models. *Bayesian Analysis*, *11*, (1), 265–298.

Page G. L., Quintana F. A. (2018). Calibrating covariate informed product partition models. *Statistics and Computing*, *28*, (5), 1009–1031.

Page G. L., Quintana F. A., Dahl D. B. (2021). Dependent Modeling of Temporal Sequences of Random Partitions. *Journal of Computational and Graphical Statistics*.

Pitman, J. (2006). *Combinatorial stochastic processes*. Springer.

Quintana F. A., Müller P., Jara A., MacEachern S. N. (2022). The Dependent Dirichlet Process and Related Models. *Statistical Science*, *37*, (1), 24–41.

Roberts G. O., Rosenthal J. S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, *18*, (2), 349–367.

Rodríguez A., Dunson D. B., Gelfand A. E. (2008). The Nested Dirichlet Process. *Journal of the American Statistical Association*, *103*, (483), 1131–1154.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Soriano J., Ma L. (2019). Mixture Modeling on Related Samples by ψ-Stick Breaking and Kernel Perturbation. *Bayesian Analysis*, *14*, (1), 161–180.

Teh, Y. W., & Jordan, M. I. (2010). *Hierarchical Bayesian nonparametric models with applications*. In N. L. Hjort, C. Holmes, P. Müller & S. G. Walker, *Bayesian Nonparametrics* (pp. 158–207). Cambridge University Press.

Teh Y. W., Jordan M. I., Beal M. J., Blei D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*, (476), 1566–1581.

Wade S., Ghahramani Z. (2018). Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). *Bayesian Analysis*, *13*, (2), 559–626.

Zanini C. T. P., Müller P., Ji Y., Quintana F. A. (2019). A Bayesian random partition model for sequential refinement and coagulation. *Biometrics*, *75*, (3), 988–999.

Zuanetti D. A., Müller P., Zhu Y., Yang S., Ji Y. (2018). Clustering distributions with the marginalized nested Dirichlet process. *Biometrics*, *74*, (2), 584–594.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lijoi, A., Prünster, I., & Rebaudo, G. (2023). Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics*, *50*(1), 213–234. https://doi.org/10.1111/sjos.12578