ISSN 2279-9362

Collegio Carlo Alberto

Flexible clustering via hidden hierarchical Dirichlet priors

Antonio Lijoi, Igor Pruenster and Giovanni Rebaudo

No. 634
December 2020

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Flexible clustering via hidden hierarchical Dirichlet priors

Antonio Lijoi [1,2], Igor Prünster [1,2] and Giovanni Rebaudo[1]

[1] Department of Decision Sciences and BIDSA, Bocconi University, via Röntgen 1, Milan, Italy
[2] Collegio Carlo Alberto, Piazza Arbarello 8, Turin, Italy

April 29, 2020

Abstract

The Bayesian approach to inference stands out for naturally allowing borrowing of information across heterogeneous populations (or studies), 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 introduce a novel nonparametric prior that arises as the composition of two different discrete random structures. We 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 on 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 to test homogeneity between different populations. Finally, we perform a comparison with the nested Dirichlet process and provide illustrative examples on both synthetic and real data.

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 (Lo, 1984; Escobar, 1994; Escobar and West, 1995; Ghosal and van der Vaart, 2017). However, they are not suitable to model data $\{(X_{j,1}, \ldots, X_{j,I_j}) : j = 1, \ldots, J\}$ 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 consider the Collaborative Perinatal Project, a large prospective epidemiologic study conducted from 1959 to 1974 (analyzed in Section 5.1), where pregnant women were enrolled in 12 hospitals and followed over time. Using a DP mixture 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 between centers.

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 on 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); Teh and Jordan (2010); Foti and Williamson (2015) 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. (2019), 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. (2019) 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. (2019).

The goal of this paper is thus to introduce a nonparametric prior, 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 novel model, which we term *Hidden Hierarchical Dirichlet Process* (HHDP). Importantly, our proposal overcomes the above-mentioned theoretical, modeling and computational limitations given 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.

Section 2 concisely reviews the HDP and the NDP with focus on the random partitions they induce. In Section 3 we introduce 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.

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 $\{(X_{j,i})_{i\geq 1} : j = 1, \ldots, J\}$ is *partial exchangeability*, which amounts to assuming homogeneity within each population, though not across different populations. This is characterized by

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

for every finitary permutations $\{\sigma_j : j = 1, \ldots, J\}$ with $\stackrel{\mathrm{d}}{=}$ henceforth denoting equality in distribution. The dependence structure is effectively visualized by the hierarchical formulation

$$\begin{aligned} (X_{j_1,i_1}, \ldots, X_{j_s,i_s}) \mid (G_1, \ldots, G_J) &\stackrel{\mathrm{iid}}{\sim} G_{j_1} \times \cdots \times G_{j_s}, \\ (G_1, \ldots, G_J) &\sim \mathscr{L}, \end{aligned} \tag{1}$$

for any $(j_1, \ldots, j_s) \in \{1, \ldots, J\}^s$ and $s \geq 1$. Here we focus on priors $\mathscr{L}$ defined as composition of discrete random structures and including, as special cases, both the HDP and the NDP. More specifically, we consider $\mathscr{L}$ in (1) that arise as

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

with discrete random probability measures G_j $(j = 1, \ldots, J)$, Q and G_0. The data are denoted by $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J\}$ with $\boldsymbol{X}_j = (X_{j,1}, \ldots, X_{j,I_j})$ 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 $j = 1, \ldots, J$, i.e. $\mathbb{P}(X_{j,i} = X_{j,\ell}) > 0$ 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 $\{1, 2, \ldots, n\}$ 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. (2019). 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 $\boldsymbol{n}_j = (n_{j,1}, \ldots, n_{j,D})$ with $n_{j,d}$ indicating the number of elements in the jth sample that coincide with the dth distinct value in order of arrival. 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

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

with the constraint $\sum_{d=1}^{D} n_{j,d} = I_j$, for each $j = 1, \ldots, J$ and where $\mathbb{X}$ is the space in which the $X_{j,i}$'s take values and $\mathbb{X}_*^D$ is the product space $\mathbb{X}^D$ minus the set of points $\boldsymbol{x} \in \mathbb{X}^D$ that have a tie in at least two coordinates. Obviously for a single population, that is $J = 1$, the standard EPPF is recovered. Note that (3) is interpretable as an extension of a product partition model to a multiple samples framework and, hence, represents an alternative approach to popular covariate–dependent product partition models. See, e.g., Müller et al. (2011), Page and Quintana (2016) and Page and Quintana (2018).

If we further specify $\mathscr{L}(\,\cdot\,|Q)$ and Q such that they give rise to an NDP, then one may have ties also among the population probability distributions $G_1, \ldots, G_J$, i.e. $\mathbb{P}(G_j = G_\kappa) > 0$ for any

$j \neq \kappa$. Therefore, in the framework of (1) and (2), one may investigate two types of clustering: (i) *sample clustering*, which is related to $G_1, \ldots, G_J$ and (ii) *observation 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

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

where $\text{DP}(\,\cdot\,|\alpha, P)$ denotes the law of a DP with concentration parameter $\alpha > 0$ and baseline probability measure P. Here we assume that H is a non–atomic probability measure on $\mathbb{X}$ and we refer to such prior as an J-dimensional HDP denoted by $(G_1, \ldots, G_J) \sim \text{HDP}(\beta, \beta_0; H)$. 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 $\mathscr{L} = \text{HDP}(\beta, \beta_0; H)$ has been determined in Camerlenghi et al. (2019, Ex. 3). 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 are considered in Camerlenghi et al. (2019) and, more recently, in Argiento et al. (2020) and Bassetti et al. (2020). 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. (2019), 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

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

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

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

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 $V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, 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, it clusters populations. Furthermore, a structure similar to the one displayed in (6) holds true for each G_k^*, i.e.

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

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

The discrete structure of the G_k^*'s generates ties across the samples $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ 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 a clustering of the observations $\boldsymbol{X}$.

If the data $\boldsymbol{X}$ are modeled as in (1), with $(G_1, \ldots, G_J) \sim \text{NDP}(\alpha, \beta; H)$, 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, i.e. $G_1 \neq G_2$, they cannot share even a single atom $X_{k,l}^*$ due to the non–atomicity 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 non-atomic 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 far 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. (2019) 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 represents in principle a solution to 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 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. (2019) provides interesting examples. See also Soriano and Ma (2019) and in Christensen and Ma (2020) for stimulating contributions to this literature.

Hence, within the composition structure framework (2), our goal is to obtain a novel prior 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 LDP 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

$$\mathscr{L}(G_j | Q) = Q(G_j), \quad \mathscr{L}(Q | G_0) = \textsc{dp}(Q | \alpha; \textsc{dp}(\beta; G_0)), \quad \mathscr{L}(G_0) = \textsc{dp}(G_0 | \beta_0; H).$$

This leads to the following definition.

Definition 1. The vector of random probability measures $(G_1, \ldots, G_J)$ is a *hidden hierarchical*

Dirichlet process (HHDP) if

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

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.

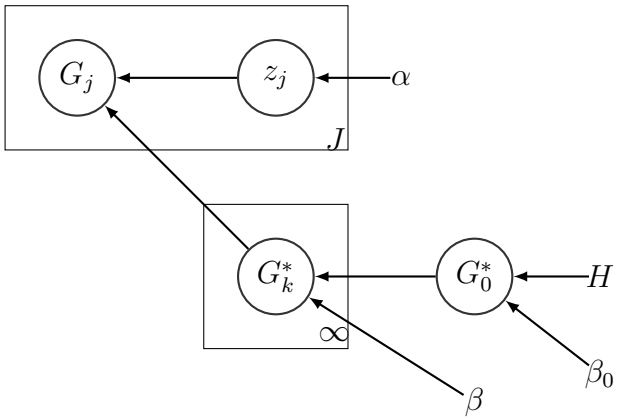

Figure 1: Graphical model representing the dependencies for a $\text{HHDP}(\alpha, \beta, \beta_0; H)$. Here the z_j's are auxiliary integer–valued random variables that assign each G_j to a specific atom G_k^* of Q.

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 $((\omega_{k,l}^*)_{l\geq 1} \mid \boldsymbol{\omega}_0) \overset{\text{iid}}{\sim} \text{DP}(\beta; \boldsymbol{\omega}_0)$, $\boldsymbol{\omega}_0 = (\omega_{0,l})_{l\geq 1} \sim \text{GEM}(\beta_0)$ 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 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 $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$, 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 $(G_1, \ldots, G_J)$, 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 set $A \subset \mathbb{X}$, it is used as a measure of overall dependence between G_j and $G_{j'}$.

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

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

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

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* $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ *are drawn from* $(G_1, \ldots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$*:*

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

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

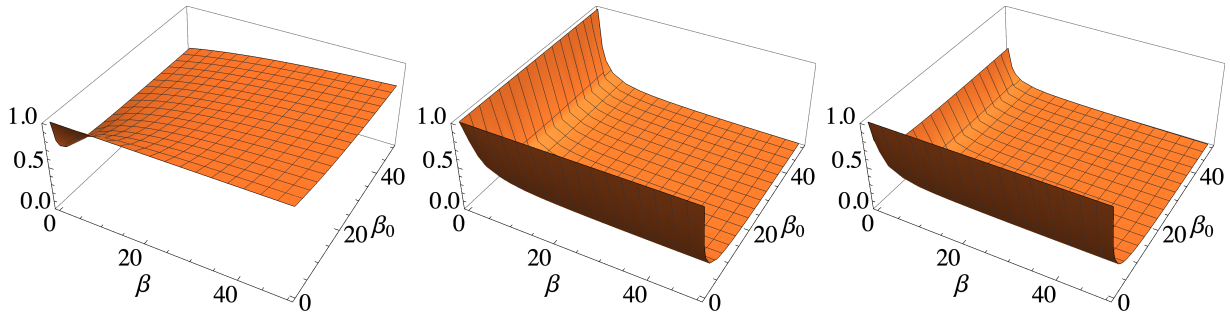

Figure 2: Correlations as function of the hyperparameters β and β_0 with $\alpha = 1$. The left plot represents the correlation between random probabilities $G_j(A)$, the middle one between observations collected in the same population and the right one between observations from different populations.

from different samples, i.e. $j \neq j'$. As for the dependence on the hyperparatmeters (α, β_0, β), $\alpha \to \infty$ forces the G_j's 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 Fig. 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, i.e. 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 to a HHDP, namely the partition of $\{1, \ldots, n\}$, 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 $G_1, \ldots, G_J$. As for the latter, the G_i's are drawn from a discrete random probability measure on $\mathsf{P}_{\mathbb{X}}$ whose weights have a $\text{GEM}(\alpha)$ 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 $\Psi^{(J)}$ is the well–known Ewens sampling formula, namely

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

where $1 \leq R \leq J$, 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 $\mathbb{P}(G_j = G_\kappa \mid \boldsymbol{X}) < 1$,

regardless of the shared clusters across the samples $\boldsymbol{X}$. Moreover, let $\Phi^{(n)}_{D,R}(\cdots;\beta,\beta_0)$ denote the pEPPF of a $\textsc{hdp}(\beta,\beta_0;H)$, namely

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

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

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

$$\Pi^{(n)}_{D}(\boldsymbol{n}_1,\ldots,\boldsymbol{n}_J)=\sum\phi^{(J)}_R(m_1,\ldots,m_R;\alpha)\Phi^{(n)}_{D,R}(\boldsymbol{n}^*_1,\ldots,\boldsymbol{n}^*_R;\beta,\beta_0),\tag{8}$$

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

Given the composition structure underlying the $\textsc{hhdp}(\alpha,\beta,\beta_0;H)$ unsurprisingly the pEPPF (8) 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

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

where $\Phi^{(n)}_{D,2}$ and $\Phi^{(n)}_{D,1}$ are the pEPPF and EPPF of a bivariate and univariate $\textsc{hdp}(\beta,\beta_0;H)$, respectively. 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 drawn from $(G_1,G_2)\sim\textsc{hhdp}(\alpha,\beta,\beta_0;H)$ the posterior probability of degeneracy is*

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

where $\Phi^{(n)}_{D,2}$ *and* $\Phi^{(n)}_{D,1}$ *are the pEPPF and EPPF induced by a bivariate and univariate* $\text{HDP}(\beta, \beta_0; H)$*, 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, *e.g.* 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 for performing inference (see Section 4.1). To develop this sampler and further understand the model behavior, we provide a characterization of the $\text{HHDP}(\alpha, \beta, \beta_0; H)$, which is reminiscent of the popular Chinese restaurant franchise metaphor for the HDP.

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 scheme 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 non–atomic 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 stay distinct if $G_1 \neq G_2$. Moreover, each $X_{j,i}$ identifies the label of the dish that customer i from the j–th population chooses from the shared menu $(X^*_d)_{d\geq 1}$, with the unique dishes $X^*_d \overset{\text{iid}}{\sim} H$. All customers are either assigned to different restaurants, if $G_1 \neq G_2$, or to the same restaurant, if $G_1 = G_2$. Then given such a grouping of the restaurants, the customers are seated according to the CRF applied either to a single restaurant or to two distinct restaurants (Teh et al., 2006; Camerlenghi et al., 2018). Furthermore, each restaurant has infinitely many tables. The first customer i who arrives in a previously unoccupied table chooses a dish that is shared by all the costumers who will join the table afterwards. 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, i.e. 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 by introducing latent variables $T_{j,i}$'s denoting the tables' labels for customer i from population j. 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, e.g. $\ell_{r,\cdot}$ 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 samples 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

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

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 e.g. survival analysis and species sampling. Henceforth, $\mathcal{K}$ is a density kernel and we consider

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

We develop two samplers: (i) a marginal algorithm that relies on the posterior degeneracy probability (Proposition 3) in Section 4.1; (ii) a conditional blocked Gibbs sampler, in the same spirit of the sampler proposed for the NDP by Rodríguez et al. (2008), in Section 4.2. 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). We present 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 $G_1, \ldots, G_J$, 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 4.2, which relies on finite–dimensional approximations of the trajectories of the underlying random probability measure. Its effectiveness in handling $J > 2$ populations is further illustrated in the application of Section 5.1.

4.1 A marginal Gibbs sampler

The marginal Gibbs sampler that updates Δ, the table dish assignments $T_{j,i}$ and D_t can be deduced from the hidden Chinese restaurant franchise presented in Section 3.2. Let $\boldsymbol{S} = \{\Delta, (T_{j,i})_{j,i}, (D_t)_t, (X_{j,i})_{j,i}\}$. Hence, the algorithm can be summarized as follows

(1) Sample the population assignments to the restaurants

$$\mathbb{P}(\Delta = 1 \mid \boldsymbol{X}) = \frac{\Phi^{(n)}_{D,1}(\boldsymbol{n}_1 + \boldsymbol{n}_2)}{\Phi^{(n)}_{D,1}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \alpha\, \Phi^{(n)}_{D,2}(\boldsymbol{n}_1 + \boldsymbol{n}_2)},$$

where $\Phi^{(n)}_{D,2}$, $\Phi^{(n)}_{D,1}$ are the pEPPF and EPPF of a bivariate and univariate $\textsc{hdp}(\beta, \beta_0; H)$, respectively.

(2) Sample the table assignments $T_{j,i}$ and corresponding dishes $D_{T_{j,i}}$ from $p(T_{j,i}, D_{T_{j,i}} \mid \boldsymbol{S}^{-(T_{j,i}, D_{T_{j,i}})}) \propto$

$$\propto \begin{cases} T_{j,i} = t & \frac{q^{-(ji)}_{r,t,\cdot}}{q^{-(ji)}_{r,\cdot,\cdot}+\beta}\ p_{D_t}(\{X_{j,i}\}) \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d & \frac{\beta}{q^{-(ji)}_{r,\cdot,\cdot}+\beta} \frac{\ell^{-(ji)}_{\cdot,d}}{\ell^{-(ji)}_{\cdot,\cdot}+\beta_0}\ p_d(\{X_{j,i}\}) \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d^{\text{new}} & \frac{\beta}{q^{-(ji)}_{r,\cdot,\cdot}+\beta} \frac{\beta_0}{\ell^{-(ji)}_{\cdot,\cdot}+\beta_0}\ p_{d^{\text{new}}}(\{X_{j,i}\}), \end{cases}$$

where $p_d(\{X_{j,i}\})$ is defined by the following equation. For every index set $\mathcal{J}$

$$p_d(\{X_{j,i}\}_{(j,i)\in\mathcal{J}}) = \frac{\int \prod_{j'i'\in\mathcal{J}\cup\mathcal{J}_d} \mathcal{K}(X_{j,i} \mid \theta) \mathrm{d}H(\theta)}{\int \prod_{j'i'\in\mathcal{J}_d\setminus\mathcal{J}} \mathcal{K}(X_{j,i} \mid \theta) \mathrm{d}H(\theta)},$$

where $\mathcal{J}_d = \{(j,i) : D_{T_{j,i}} = d\}$. For instance, $p_d(\{X_{j,i}\})$ is the marginal conditional probability of $X_{j,i}$ in cluster d given the other observation assigned to cluster d.

(3) Sample the dish assignments D_t from

$$p(D_t \mid \boldsymbol{S}^{-t}) \propto \begin{cases} d & \frac{\ell^{-t}_{\cdot,d}}{\ell^{-t}_{\cdot,\cdot}+\beta_0} p_d(\{x_{j,i} : T_{j,i} = t\}) \\ d^{\text{new}} & \frac{\beta_0}{\ell^{-t}_{\cdot,\cdot}+\beta_0} p_{d^{\text{new}}}(\{x_{j,i} : T_{j,i} = t\}). \end{cases}$$

4.2 A conditional blocked Gibbs sampler

A more effective algorithm is based on 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

$$\boldsymbol{\pi}^* \sim \textsc{dir}(\alpha/K, \ldots, \alpha/K), \qquad \boldsymbol{\omega}^*_0 \sim \textsc{dir}\big(\beta_0/L, \ldots, \beta_0/L\big) \tag{12}$$

implying that $(\boldsymbol{\omega}_k^* \mid \boldsymbol{\omega}_0^*) \overset{\text{iid}}{\sim} \text{DIR}(\beta\, \boldsymbol{\omega}_0^*)$, 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)

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

This leads to the following

(1) Sample the unique θ_l^* from

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

(2) Sample distributional cluster probabilities from

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

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

(3) Sample probability weights of the base DP from

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

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

(4) Sample the observational cluster probabilities independently fromfrom

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

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

$$p(z_j = k \mid \boldsymbol{S}^{-\{z_j,\zeta_j\}}) \propto \pi^*_k \prod_{i=1}^{I_j} \sum_{l=1}^{L} \omega^*_{k,l} \mathcal{K}(X_{j,i} \mid \theta^*_l) \qquad (k = 1, \ldots, K),$$

$$p(\zeta_{j,i} = l \mid \boldsymbol{S}^{-\zeta_{j,i}}) \propto \omega^*_{z_j l} \mathcal{K}(X_{j,i} \mid \theta^*_l) \qquad (l = 1, \ldots, L).$$

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 $[\log(\omega_{0,1}/\omega_{0,L}), \ldots, \log($ 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 $\mathcal{K}(\cdot|\theta) = \mathrm{N}(\cdot|\mu, \sigma^2)$ and a conjugate Normal-inverse-Gamma base measure $H(\cdot) = \textsc{nig}(\cdot \mid \mu_0, \lambda_0, s_0, S_0)$ and obtain

$$p(\theta^*_l \mid \boldsymbol{S}^{-\theta^*_l}) = \textsc{nig}(\theta^*_l \mid \mu_l, \lambda_l, s_l, S_l),$$

with $\mu_l = \dfrac{n_l \bar{y}_l + \lambda_0 \mu_0}{\lambda_0 + n_l}$, $S_l = S_0 + \dfrac{1}{2}\left(e^2_l + \dfrac{n_l \lambda_0 (\bar{y}_l - \mu_0)^2}{\lambda_0 + n_l} \right)$, $\lambda_l = \lambda_0 + n_l$, 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 $e^2_l = \sum_{\{j,i:\zeta_{j,i}=l\}} (X_{j,i} - \bar{y}_l)^2$ are the observational cluster sizes, means and deviances, respectively.

5 Illustration

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. In doing so we rely on blocked Gibbs sampler of Section 4. The data are simulated from the same scenarios considered in Camerlenghi et al. (2019). More precisely, we consider two populations and the data in each population are iid from a mixture of two normals:

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

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

Scen 2. We simulate the data in the two populations independently from a mixture of two normals with one shared component

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

Scen 3. We simulate the data in the two populations independently from a mixture of two normals having the same components with different weights

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

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 $H(\cdot) = \textsc{nig}(\cdot \mid \mu_0, \lambda_0, s_0, S_0)$. We set standard values of the hyperparameters in terms of the mean $\bar{y}$ and variance $\mathrm{Var}(y)$ of the data, i.e. $\mu_0 = \bar{y}$, $\lambda_0 = 1/(3\,\mathrm{Var}(y))$, $s_0 = 1$ and $S_0 = 4$. In drawing the comparison between (11) and the $\textsc{ndp}(\alpha, \beta; H)$, we further set $\alpha = \beta = 1$. Furthermore, we set the concentration parameters all equal to 1. In Appendix A.5 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 $\mathrm{Var}[G_j(A)]$ 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 Appendix we perform an empirical analysis trying different levels of L and K which corroborates the fact that the approximation error is negligible in term 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 Meil (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 $\mathbb{P}(\zeta_{j,i} = \zeta_{j',i'} \mid \boldsymbol{X})$ 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 the MCMC iterations. These are visualized through heat maps in Fig. 5, with colors ranging from white, if the probability is 0, to dark, 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 Fig. 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 2, the heat maps of the posterior co-clustering probabilities in Fig. 4 and the posterior distributions of the overall number of components in Table 1 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 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.

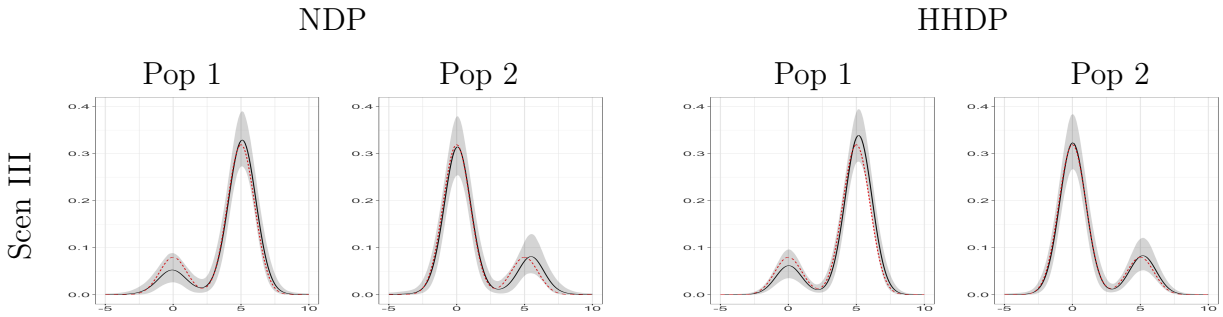

Figure 3: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under the third scenario.

Scen	Model	Overall number of components									
		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

Table 1: Posterior distributions of the number of overall components estimated 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

Table 2: Frequencies of observations in the two populations allocated to the point estimate of the clustering that minimizes the VI loss with the two models under different scenarios.

5.1 Collaborative perinatal project data

A multi-center application is the focus of this section. We consider a data set from the Collaborative Perinatal Project (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 pre–pregnancy measurements, we focus on the birth weight $X_{j,i}$ for non-smoking woman i in center j. We assume the following Gaussian mixture model:

$$
\begin{aligned}
X_{j,i} \mid \mu_{j,i}, \sigma_{j,i} &\overset{\text{ind}}{\sim} \mathrm{N}(\mu_{j,i}, \sigma_{j,i}) && (i = 1, \ldots, I_j, \quad j = 1; \ldots, 12), \\
\mu_{j,i}, \sigma_{j,i} \mid G_j &\overset{\text{ind}}{\sim} G_j && (i = 1, \ldots, I_j, \quad j = 1; \ldots, 12).
\end{aligned}
$$

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 cap-

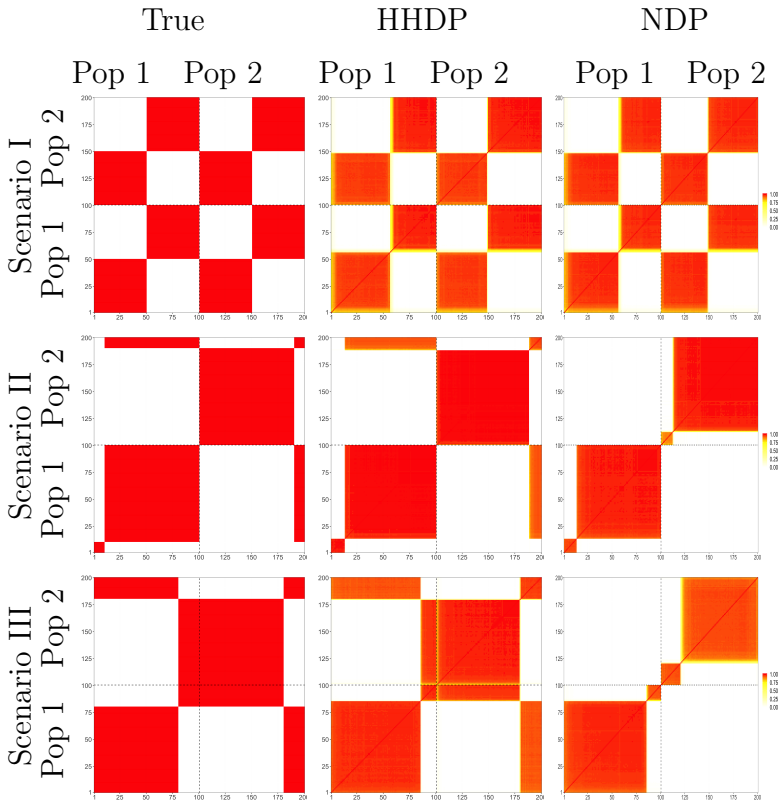

Figure 4: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different scenarios.

ture differences between the centers since different groups of hospitals can serve different women. Canale et al. (2019) provide a further analysis of the CPP data.

The heat map of the co-clustering posterior probability for the 12 hospitals is shown in Fig. 5. 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 to the two clusters are reported. Note that these 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 weight.

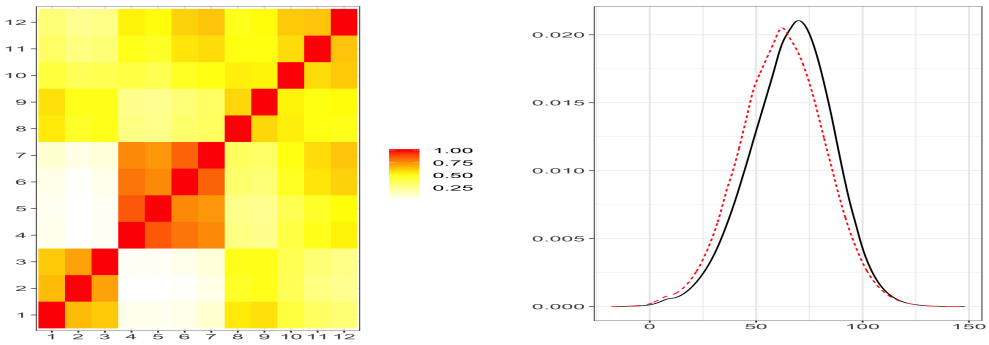

Figure 5: Heat map of the estimated posterior probability of co-clustering of hospitals and estimated population cluster specific posterior densities for the CPP data.

6 Discussion

As highlighted in the recent literature, NDP mixture models are not an appropriate tool for clustering simultaneously population distributions and observations. Our new proposal, the HHDP, overcomes the issues plaguing the NDP, while preserving tractability and clustering flexibility. We also derive sampling schemes allowing efficient inference when the number of populations increases. This work lays the foundation for future research. First, it is natural to move beyond DPs combining other discrete nonparametric priors, as the Pitman-Yor process and normalized completely random measures, and studying the induced clustering. Moreover, it is of interest to investigate and tailor the HHDP to perform inference on survival and functional data.

A Appendix

A.1 Proof of Proposition 1

Note that $G_1^*(A) \mid G_0^* \sim \text{BETA}(\beta G_0^*(A), \beta(1 - G_0^*(A)))$ and $G_0^*(A) \sim \text{BETA}(\beta_0 H(A), \beta_0(1 - H(A)))$. Hence,

$$\mathbb{E}G_0^*(A) = H(A), \quad \text{Var}[G_0^*(A)] = \frac{H(A)[1 - H(A)]}{\beta_0 + 1}$$

and since $G_j \stackrel{\mathrm{d}}{=} G_1^*$,

$$\mathbb{E}G_j(A) = \mathbb{E}\mathbb{E}[G_1^*(A) \mid G_0^*] = \mathbb{E}G_0^*(A) = H(A)$$
$$\mathrm{Var}[G_j(A)] = \mathbb{E}\mathrm{Var}[G_1^*(A) \mid G_0^*] + \mathrm{Var}[G_0^*(A)] = \frac{H(A)[1-H(A)](\beta_0+\beta+1)}{(\beta+1)(\beta_0+1)}.$$

Mixed moments are also easy to determine, as $\mathbb{E}G_1^*(A)G_2^*(A) = \mathbb{E}\mathbb{E}[G_1^*(A) \mid G_0^*]\mathbb{E}[G_2^*(A) \mid G_0^*] = \mathbb{E}G_0^*(A)^2$ and

$$\begin{aligned}\mathbb{E}G_j(A)G_{j'}(A) &= \mathbb{E}[G_1(A)G_2(A) \mid G_1 = G_2]\,\mathbb{P}(G_1 = G_2) + \mathbb{E}[G_1(A)G_2(A) \mid G_1 \neq G_2]\,\mathbb{P}(G_1 \neq G_2)\\ &= \frac{1}{1+\alpha}\mathbb{E}[G_1^*(A)^2] + \frac{\alpha}{\alpha+1}\mathbb{E}[G_1^*(A)G_2^*(A)]\\ &= \frac{1}{1+\alpha}\mathbb{E}[G_1^*(A)^2] + \frac{\alpha}{\alpha+1}\mathbb{E}[G_0^*(A)^2].\end{aligned}$$

One, then, obtains

$$\mathrm{Cov}[G_j(A), G_{j'}(A)] = \mathbb{E}[G_j(A)G_{j'}(A)] - H(A)^2 = \frac{1}{1+\alpha}\mathrm{Var}[G_1^*(A)] + \frac{\alpha}{\alpha+1}\mathrm{Var}[G_0^*(A)]$$

and

$$\mathrm{Cor}[G_j(A), G_{j'}(A)] = \frac{1}{1+\alpha} + \frac{\alpha}{\alpha+1}\frac{\mathrm{Var}[G_0^*(A)]}{\mathrm{Var}[G_1^*(A)]} = \frac{\beta_0+\beta+1+\alpha\beta+\alpha}{(1+\alpha)(\beta_0+\beta+1)}$$

so that the conclusion follows. □

A.2 Proof of Proposition 2

Note that $X_{j,i} \stackrel{\mathrm{d}}{=} X_d^*$. Thus,

$$\mathrm{Cov}(X_{j,i}, X_{j',i'}) = \mathbb{P}(X_{j,i} = X_{j',i'})\,\mathrm{Var}(X_d^*)$$

Moreover, if $j = j'$, then

$$\begin{aligned}\mathbb{P}(X_{j,i'} = X_{j,i}) &= \mathbb{P}(X_{j,i'} = X_{j,i} \mid T_{j,i} = T_{j,i'})\,\mathbb{P}(T_{j,i} = T_{j,i'}) + \mathbb{P}(X_{j,i'} = X_{j,i} \mid T_{j,i} \neq T_{j,i'})\,\mathbb{P}(T_{j,i} \neq T_{j,i'})\\ &= \frac{1}{\beta+1} + \mathbb{P}(D_{T_{j,i'}} = D_{T_{j,i}} \mid T_{j,i} \neq T_{j,i'})\,\frac{\beta}{\beta+1} = \frac{\beta+\beta_0+1}{(\beta+1)(\beta_0+1)}\end{aligned}$$

If $j \neq j'$, then

$$\begin{aligned}\mathbb{P}(X_{j,i} = X_{j',i'}) &= \mathbb{P}(X_{j,i} = X_{j',i'} \mid G_j = G_{j'})\,\mathbb{P}(G_j = G_{j'}) + \mathbb{P}(X_{j,i} = X_{j',i'} \mid G_j \neq G_{j'})\,\mathbb{P}(G_j \neq G_{j'})\\ &= \mathbb{P}(X_{j,i'} = X_{j,i})\mathbb{P}(G_j = G_{j'}) + \mathbb{P}(D_{T_{j',i'}} = D_{T_{j,i}} \mid T_{j,i} \neq T_{j',i'})\,\mathbb{P}(G_j \neq G_{j'})\\ &= \frac{1}{\beta_0+1} + \frac{\beta_0}{(1+\alpha)(1+\beta)(1+\beta_0)}\end{aligned}$$

and the conclusion follows. □

A.3 Proof of Theorem 1

In order to prove Theorem 1 we first state the following auxiliary result.

Lemma 1. *The random partition induced by the samples* $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ *drawn from* $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ *given a particular partition of distributions* $\Psi^{(J)} = \{B_1, \dots, B_R\}$ *is characterized by the pEPPF*

$$\Pi_D^{(n)}\left(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \dots, B_R\}\right) = \Phi_{D,R}^{(n)}\left(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0\right),$$

where $n_{r,d}^* = \sum_{j \in B_r} n_{j,d}$ *for each* $r = 1, \dots, R$, $d = 1, \dots, D$ *and* $\Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0)$ *is the pEPPF associated to a R-dimensional* $\text{HDP}(\beta, \beta_0; H)$.

Now we can write

$$\begin{aligned}
&\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \dots, B_R\}) = \\
&= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} G_1^{n_{1,d}}(\mathrm{d}x_d) \dots G_J^{n_{J,d}}(\mathrm{d}x_d) \mid \Psi^{(J)} = \{B_1, \dots, B_R\}\right] = \\
&= \mathbb{E}\left[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} G_1^{*n_{1,d}^*}(\mathrm{d}x_d) \dots G_R^{*n_{R,d}^*}(\mathrm{d}x_d)\right] = \Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0),
\end{aligned} \tag{14}$$

with $\mathbb{X}_*^D = \mathbb{X}^D \setminus \{\boldsymbol{x} : x_i = x_j \text{ for some } i \neq j\}$ and $(G_1^*, \dots, G_R^*) \sim \text{HDP}(\beta, \beta_0; H)$. Moreover, note that the R unique values among $(G_1, \dots, G_J)$ are not necessarily the first $(G_1^*, \dots, G_R^*)$ but since $(G_k^*)_{k \geq 1}$ are exchangeable the third equality holds.

Therefore, by applying Lemma 1

$$\begin{aligned}
\Pi_D^{(n)}\big(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J\big) &= \sum p\big(\Psi^{(J)} = \{B_1, \dots, B_R\}\big) \Pi_D^{(n)}\big(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \dots, B_R\}\big) = \\
&= \sum \phi_R^{(J)}\big(m_1, \dots, m_R; \alpha, \big) \Phi_{D,R}^{(n)}\big(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \beta, \beta_0\big)
\end{aligned} \tag{15}$$

A.4 Proof of Proposition 3

In order to derive the posterior probability of degeneracy we write the marginal likelihood as

$$p(\boldsymbol{X}) = \Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{d=1}^{D} H(\mathrm{d}X_d^*),$$

where $\{X_1^*, \ldots, X_D^*\}$ are the D unique values among $\boldsymbol{X}$ and $\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is the pEPPF associated to the proposed model (8), that is

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \mathbb{P}(G_1 = G_2)\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \mathbb{P}(G_1 \neq G_2)\Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2),$$

Finally, we prove the proposition by applying Bayes theorem

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

where $\Phi_{D,1}^{(n)}$ and $\Phi_{D,2}^{(n)}$ are the pEPPF and the EPPF of a bivariate and univariate $\text{HDP}(\beta, \beta_0; H)$, respectively.

More precisely, following Camerlenghi et al. (2019, 2018) we can derive the pEPPF $\Phi_{D,2}^{(n)}$ and the EPPF $\Phi_{D,1}^{(n)}$ of a bivariate and univariate $\text{HDP}(\beta, \beta_0; H)$, respectively. In particular

$$\Phi_{D,1}^{(n)}(\boldsymbol{n}^*) = \frac{\beta_0^D}{(\beta)_n} \sum_{\boldsymbol{\ell}^*} \frac{\beta^{|\boldsymbol{\ell}^*|}}{(\beta_0)_{|\boldsymbol{\ell}^*|}} \prod_{d=1}^{D} (\ell_d^* - 1)! |\mathfrak{s}(n_d^*, \ell_d^*)|, \tag{16}$$

where $|\mathfrak{s}(n, \ell)|$ is the signless Stirling numbers of the first kind and the sum runs over all vectors $\boldsymbol{\ell}^* = (\ell_1^*, \ldots, \ell_D^*)$ such that $\ell_d^* \in \{1, \ldots, n_d^*\}$, $|\ell^*| = \sum_{d=1}^D \ell_d^*$ and

$$\Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{\beta_0^D}{\prod_{j=1}^J (\beta)_{I_j}} \sum_{\boldsymbol{\ell}} \frac{\beta^{|\boldsymbol{\ell}|}}{(\beta_0)_{|\boldsymbol{\ell}|}} \prod_{d=1}^{D} (\ell_{\cdot l} - 1)! \prod_{j=1}^{2} |\mathfrak{s}(n_{j,d}, \ell_{j,d})|, \tag{17}$$

where $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$, with each $\boldsymbol{\ell}_j = (\ell_{j,1}, \ldots, \ell_{j,D}) \in \times_{d=1}^D \{1, \ldots, n_{j,d}\}$ and $|\ell| = \sum_{j=1}^2 \sum_{d=1}^D \ell_{j,d}$.

A.5 Sensitivity analysis for the hyperparameters specification

Here we study the robustness with respect to the specification of hyperparameters in relation to the comparison between the NDP and the HHDP mixture models presented in Section 5. The results are reported in terms of density estimates in Fig. 6 and probabilities of co-clustering of the observations in Fig. 7 using the finite–dimensional approximations of the DPs with $L = K = 50$ and different hyperparameter specifications. The sensitivity analysis is performed by selecting different values for the concentration parameters. This allows to verify the robustness of the results comparing the two models. We report the results for the data simulated according to scenario III, in which the two populations share both the Gaussian components, but with different mixture weights.

We perform inference with the model as in Section 5 with the following specifications for the concentration parameters:

- **Parameters 1:** all the concentration parameters are set equal to 1, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 1, \beta = 1; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 1, \beta = 1, \beta_0 = 1; H)$, respectively.
- **Parameters 0.1:** all the concentration parameters are set equal to 0.1, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 0.1, \beta = 0.1; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 0.1, \beta = 0.1, \beta_0 = 0.1; H)$, respectively.
- **Parameters 3:** all the concentration parameters are set equal to 3, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 3, \beta = 3; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 3, \beta = 3, \beta_0 = 3; H)$, respectively.

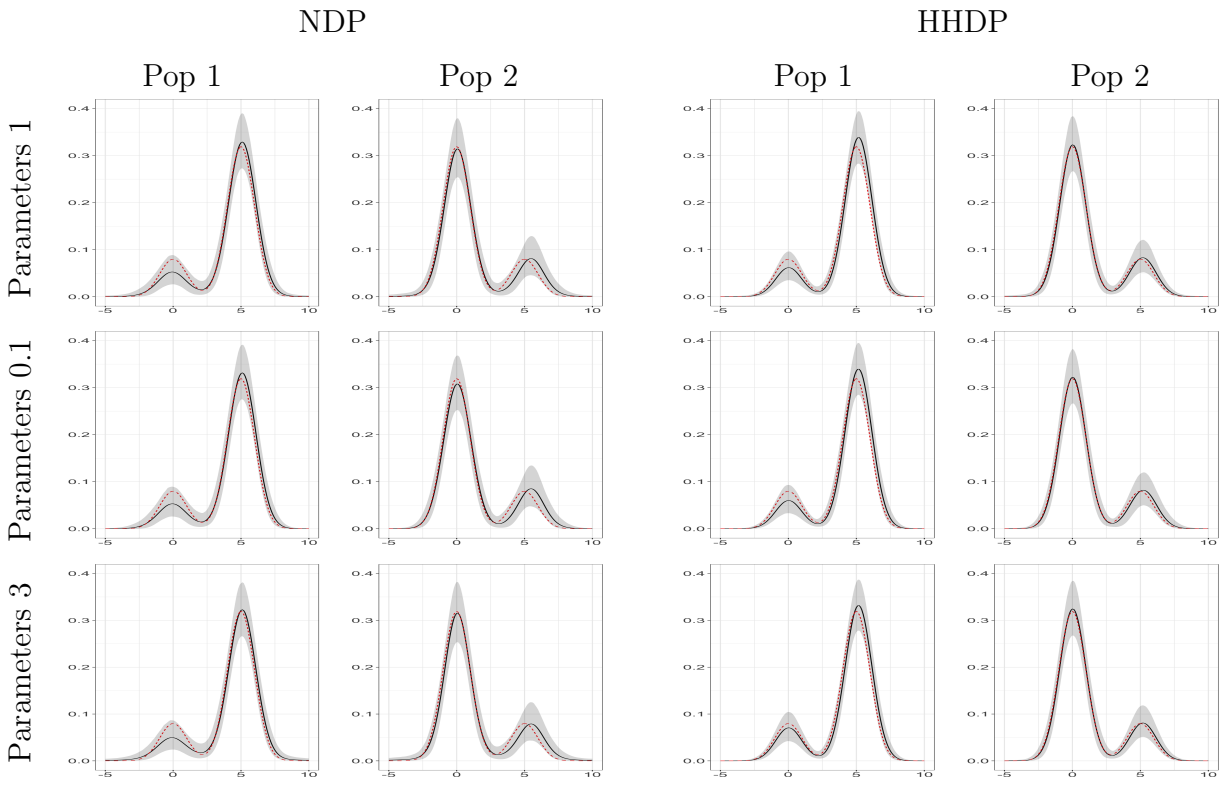

Figure 6: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under different hyperparameters specifications.

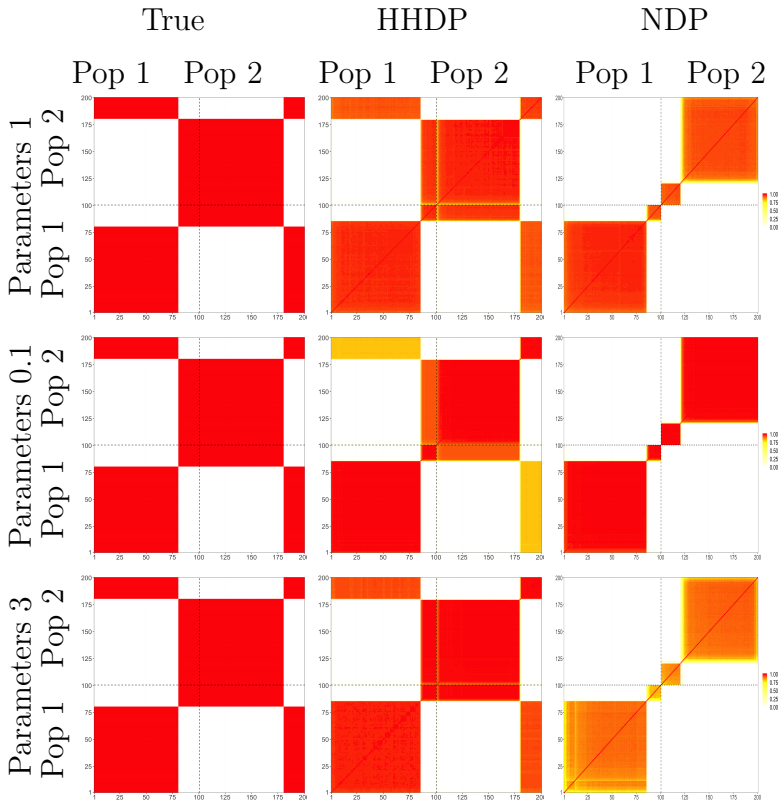

Figure 7: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different hyperparameters specifications.

Importantly the density estimates are essentially the same under the different hyperparameters specifications. Probabilities of co-clustering change under the different hyperparameter settings coherently with the theory developed in Section 3.1. However, in all the scenarios both models do not degenerate to the exchangeable case. This implies that the NDP cannot cluster observations across populations, while the HHDP overcomes this issue. Therefore, the results of the comparison between the two models presented in Section 5 are essentially the same.

A.6 Choice of the finite dimensional approximations

We now present the inferential results in terms of density estimates in Fig. 6 and probability of co-clustering of the observations in Fig. 7 for the two specifications in Section 5. We report the results for the data simulated according to scenario III with the following finite dimensional

approximations of the DPs:

- $L = K = 50$;
- $L = K = 30$;
- $L = K = 70$.

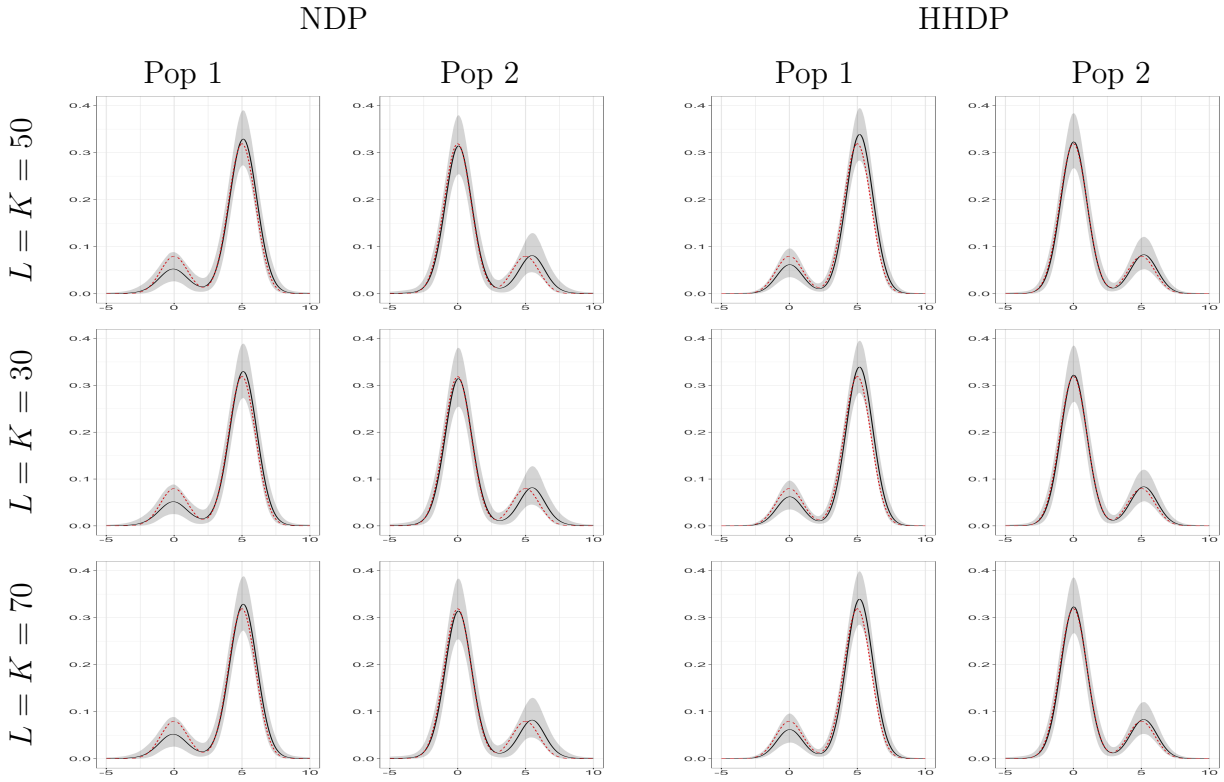

Figure 8: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under different truncation levels.

Under all the different finite dimensional approximations the inference is qualitatively the same, corroborating the idea that the finite dimensional approximations $L = K = 50$ proposed for the comparison of the NDP and HHDP in Section 5 induce a negligible error in our analysis.

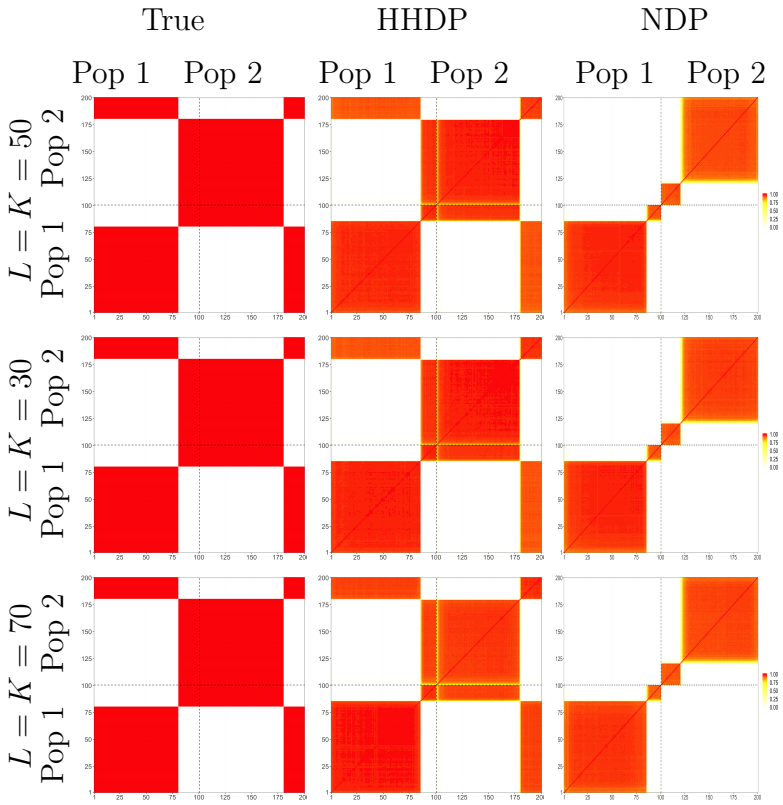

Figure 9: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different finite dimensional approximations.

References

Argiento, R., A. Cremaschi, and M. Vannucci (2020). Hierarchical Normalized Completely Random Measures to Cluster Grouped Data. *J. Amer. Statist. Assoc. 115*(529), 318–333.

Bassetti, F., R. Casarin, and L. Rossini (2020). Hierarchical species sampling models. *Bayesian Analysis*. Forthcoming.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez (2019). Latent nested nonparametric priors. *Bayesian Analysis 14*, 1303–1356. (With discussion).

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

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

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

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

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

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

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *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. and M. West (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 *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Dordrecht: Springer.

Foti, N. J. and S. A. Williamson (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Trans. Pattern Anal. Mach. Intell. 37*(2), 359–371.

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

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

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

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*, pp. 50–55.

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

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

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

Müller, P., F. Quintana, and G. L. Rosner (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist. 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. and F. A. Quintana (2016). Spatial product partition models. *Bayesian Anal. 11*(1), 265–298.

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

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

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

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

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*(2), 639–650.

Soriano, J. and L. Ma (2019). Mixture modeling on related samples by ψ-stick breaking and kernel perturbation. *Bayesian Anal. 14*(1), 161–180.

Teh, Y. W. and M. I. Jordan (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, pp. 158–207. Cambridge University Press.

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

Wade, S. and Z. Ghahramani (2018). Bayesian cluster analysis: point estimation and credible balls. *Bayesian Analysis 13*(2), 559–626.

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