A review on Bayesian model-based clustering

Clara Grazian

e-mail: clara.grazian@sydney.edu.au

Abstract: Clustering is an important task in many areas of knowledge: medicine and epidemiology, genomics, environmental science, economics, visual sciences, among others. Methodologies to perform inference on the number of clusters have often been proved to be inconsistent, and introducing a dependence structure among the clusters implies additional difficulties in the estimation process. In a Bayesian setting, clustering is performed by considering the unknown partition as a random object and define a prior distribution on it. This prior distribution may be induced by models on the observations, or directly defined for the partition. Several recent results, however, have shown the difficulties in consistently estimating the number of clusters, and, therefore, the partition. The problem itself of summarising the posterior distribution on the partition remains open, given the large dimension of the partition space. This work aims at reviewing the Bayesian approaches available in the literature to perform clustering, presenting advantages and disadvantages of each of them in order to suggest future lines of research.

Keywords and phrases: Dirichlet process, Mixture models, Clustering, Bayesian analysis, Partitions.

1. Introduction

An important task in statistical modeling is the identification of groups or partitions among observations, aiming to group together those that exhibit greater homogeneity in a specific aspect compared to other clusters. Clustering serves as an initial step in numerous analyses. For instance, in regression, it is common to utilize homogeneous groups to explore associations with particular covariates. However, while clustering is a vital step, it is also a delicate task as the interpretation of the resulting groups is subjective and open to interpretation.

There are two main approaches to clustering: distance-based clustering methods, such as k-means and hierarchical clustering, which define similarity among observations based on a chosen distance metric; and model-based clustering approaches, which assume a probabilistic model and probabilistically assign observations to different clusters. For model-based clustering with a fixed and known number of groups, mixture models are commonly used. However, a drawback of using these models is the need to estimate or select the number of clusters in advance. Model selection techniques, such as the deviance information criterion (DIC) (Celeux et al., 2006) or the integrated likelihood criterion (ICL) (Biernacki et al., 2000), are available. Nevertheless, the performance of each method can vary depending on the specific problem, and different criteria may disagree regarding the true number of components in the underlying model.

Alternatively, it is possible to consider the number of components as an unknown parameter and define a prior distribution for it (Nobile, 2004). In this context, the prior distributions used for the number of components or the component parameters can significantly influence posterior estimation. For instance, studies by Richardson and Green (1997) and Jasra et al. (2005) demonstrate that a Gaussian mixture model with a prior distribution having a large variance on the component means tends to favor smaller values for the posterior distribution of the number of components (and consequently, the number of clusters). Furthermore, research indicates that the posterior distribution on the number of components can diverge when there is misspecification of the component distributions, as shown in studies such as Woo and Sriram (2006), Woo and Sriram (2007), Rodríguez and Walker (2014), and Cai et al. (2021). In this work, it is assumed that the component distributions are correctly specified, and we focus on priors for the number of components.

The nonparametric extensions of finite mixture models, which allow for an infinite number of components, often rely on Dirichlet processes (DP) (Ferguson, 1973). Dirichlet processes have a significant role in Bayesian nonparametrics, not only for clustering but also for density estimation, due to their computationally manageable representations. Notably, the stick-breaking representation (Sethuraman, 1994), the Polya urn representation (Blackwell and MacQueen, 1973), and the Chinese restaurant process (Aldous, 1985) are frequently used. One key characteristic of the Dirichlet process is its ability to assign probability one to a set of countable, discrete distributions. While this property poses a limitation for density estimation, which is usually overcome through the definition of Dirichlet process mixture models, it proves useful in clustering as it automatically groups observations.

Although consistency in L_1 to the true density has been demonstrated for Dirichlet processes in density estimation, achieving the minimax optimal rate up to logarithmic factors (Ghosal et al., 1999; Ghosal and Van Der Vaart, 2001; Lijoi et al., 2005; Tokdar, 2006; Ghosal and Van Der Vaart, 2007; Walker et al., 2007; Kruijer et al., 2010; Wu and Ghosal, 2010; Nguyen, 2013), these results cannot be readily extended to study consistency for the number of clusters. This is because any mixture with k components can be approximated in L_1 by another mixture with $k+1$ components (or generally $k' > k$). Recent work by Miller and Harrison (2014) analytically proves the inconsistency of the posterior distribution on the number of components for a broad class of infinite mixtures, including DP mixture models with various forms of component distributions.

A particular mention should be made regarding clustering in the presence of covariates. This refers to the partitioning of a set of experimental units, where the probability of each partition depends on the covariates. In other words, observations with similar or equal levels of covariates are more likely to be clustered together. In model-based clustering, the dependence on covariates can be incorporated into the cluster probabilities. Examples of such approaches can be found in works by Pawlowsky and Burger (1992), Fernández and Green (2002), Tjelmeland and Lund (2003), Neelon et al. (2014), and Paci and Finazzi (2018). To address biases arising from the sum-to-one constraint of the probability vector, Mastrantonio et al. (2019) propose a logit-Gaussian process. For a recent review of dependent Dirichlet processes, we recommend referring to Quintana et al. (2022).

This study aims to examine the advantages and disadvantages of various approaches to Bayesian model-based clustering found in the literature, in order to provide insights for future research directions.

The paper is organized as follows: Section 2 provides the definitions of random partition models, along with the notation used throughout the paper. Section 3 explores induced random partition models based on both finite and infinite mixture models. Section 4 discusses product partition models and other prior distributions for partitions. Section 5 addresses the problem of clustering populations, while Section 6 introduces approaches to estimate the optimal partition once its posterior distribution is available. Finally, Section 7 presents the concluding remarks of the paper.

2. Random partition models

Model-based clustering involves randomly allocating observations to clusters identified by the model. Let $[n] = \{1, \dots, n\}$ be a set of n indices, and define $\rho_n = (S_1, \dots, S_K)$ as a random partition of the set $[n]$, where $K = |\rho_n|$ represents the number of non-empty and mutually exclusive subsets. The sets S_h are non-empty and mutually exclusive, satisfying $\cup_{S \in \rho_n} S = 1, \dots, n$ and $S_\ell \cap S_h = \emptyset$ for $\ell \neq h$. The random partition $\rho_n \in \mathcal{P}_n$, where $\mathcal{P}_n$ denotes the set of all possible partitions of $[n]$, known as the n-th Bell number. The size of $\mathcal{P}_n$ makes analytical computations infeasible, even for small sample sizes. Moreover, the number of ways to assign n elements to K groups can be determined using the Stirling number of the second kind:

$$\mathcal{S}_{n,K} = \frac{1}{K!} \sum_{h=1}^{K} (-1)^h \binom{K}{h} (K-h)^n$$

The Bell number can be defined as the sum of Stirling numbers: $\sum_{h=1}^{n} \mathcal{S}_{n,h}$.

It is common to represent a partition $\rho_n = (S_1, \ldots, S_K)$ using class label memberships $C_1, \ldots, C_n$, where $C_i \in [K]$ and

$$C_i = h \Leftrightarrow i \in S_h \qquad \text{for } i \in [n] \text{ and } 1 \leq h \leq K.$$

Partition structures are often associated with the exchangeable partition probability function (EPPF) (Pitman, 1995), expressed as

$$P(\rho_n = (S_1, \ldots, S_K)) = p(|S_1|, \ldots, |S_K|) = p(n_1, \ldots, n_K).$$

which is a function $p : \mathbb{N}^* = \cup_{h=1}^{\infty} \mathbb{N}^h \to [0, 1]$ symmetric in its argument, and where $\mathbb{N}$ represents the set of natural numbers. It is worth noting that many works, including this one, use this definition in terms of aggregated probabilities, i.e., probabilities for cluster sizes $\sum_{\rho_n \in \mathcal{A}} P(\rho_n = S_1, \ldots, X_K)$, where $\mathcal{A} = \{\rho_n \in \mathcal{P}^n : \rho_n \text{ has cluster sizes } (n_1, \ldots, n_K)\}$. For an interesting discussion on definitions of the probability mass function of each partition, refer to Lee and Sang (2022).

The function p has the following properties. Let $\mathbf{n}$ be the infinite sequence $(n_1, n_2, \ldots, n_K, 0, 0, \ldots)$. Then, $p(1) = 1$ and

$$p(\mathbf{n}) = \sum_{h=1}^{K(\mathbf{n})+1} p(\mathbf{n}^{h+}) \quad \forall\, \mathbf{n} \in \mathbb{N}^*$$

where $\mathbf{n}^{h+}$ corresponds to $\mathbf{n}$ with the h-th element increased by one unit. Moreover, $K(\mathbf{n})$ represents the number of non-zero components of $\mathbf{n}$. It is important to note the distinction between the number of components and the number of clusters. For a given K, $K(\mathbf{n}) = K_+$ is defined as the number of components that generated the data, i.e., $K_+ = \sum_{h=1}^{K} \mathbb{I} n_h > 0$, where $n_h = \#i : C_i = h$ counts the observations allocated to component h. For further discussion, see Argiento and De Iorio (2022). A notable property of the EPPF is the sample size consistency or addition rule (De Blasi et al., 2015): $p(\rho_n)$ can be derived from $p(\rho_{n+1})$ by marginalizing the last element.

The EPPF is associated with the prediction probability function (PPF), which provides the predictive distribution of a future observation j, given by

$$p_j(\mathbf{n}) = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})} \quad 1 \leq j \leq K + 1.$$

While the definition of a PPF directly derives from an EPPF, the converse is not necessarily true. Lee et al. (2013) provide a necessary and sufficient condition for arbitrary PPFs to define an EPPF.

In the following sections, we review methods to define partition models for ρ_n. Some of these methods result in an analytical form of the EPPF, while others are not associated with a closed-form expression of the EPPF.

3. Induced random partitions models

A possible way to define a random partition probability distribution is by constructing a hierarchical model on the observations and inducing a model on the random partition. The model can be expressed as follows:

$$\begin{aligned} p(y_1, \ldots, y_n | \theta_1, \ldots, \theta_n) &= g(y_1, \ldots, y_n | \theta_1, \ldots, \theta_n) \\ \theta_1, \ldots, \theta_n | F &\sim F \end{aligned} \tag{1}$$

$$F \sim \text{discrete RPM}.$$

where RPM denotes a random probability measure. The discreteness of F implies the presence of ties among the atoms of the process $\theta_1, \dots, \theta_n$. Let $\theta_1^*, \dots, \theta_K^*$ denote the unique values of $\theta_1, \dots, \theta_n$. The partition ρ_n can be redefined as follows: $C_i = h \Leftrightarrow \theta_i = \theta_h^*$. Thus, $S_h = \{i \in [n] : \theta_i = \theta_h^*\}$. Alternatively, the h-th unique value θ_h^* can also be denoted as $\theta_{C_i}^*$. A common choice for inducing a partition on the observations is to use mixture models.

3.1. Finite mixture models

In this section, we introduce several classes of finite mixture models (Frühwirth-Schnatter, 2006; Frühwirth-Schnatter et al., 2019). Generally, a finite mixture model is defined as follows:

$$g(y_i|\pi_1, \dots, \pi_K, \theta_1^*, \dots, \theta_K^*) = \sum_{h=1}^{K} \pi_h f_h(y_i|\theta_h^*) \qquad i = 1, \dots, n. \tag{2}$$

Here, $(\pi_1, \dots, \pi_K)$ are weights satisfying $\pi_h \leq 0$ for $h = 1, \dots, K$ and $\sum_{h=1}^{K} \pi_h = 1$. The term $f_h(\cdot|\theta_h^*)$ represents a probability distribution indexed by component-specific parameters. Typically, the component distributions are assumed to be from the same family, so we have $f_h(\cdot|\theta_h^*) = f(\cdot|\theta_h^*)$.

In a Bayesian framework, prior distributions are assigned to all parameters as follows:

$$\begin{aligned} Y_i|C_i = c_i, \pi_1, \dots, \pi_K, \theta_1^*, \dots, \theta_K^* &\overset{i.i.d.}{\sim} f_{c_i}(y_i|\theta_{c_i}^*) \qquad i = 1, \dots, n \\ C_1, \dots, C_n &\sim Cat(\pi_1, \dots, \pi_K) \\ \theta_1^*, \dots, \theta_K^* &\overset{i.i.d.}{\sim} F_0 \\ \pi_1, \dots, \pi_K &\sim Dir(\gamma, \dots, \gamma). \end{aligned} \tag{3}$$

Here, $C_1, \dots, C_n$ are assigned a multinomial distribution with probabilities given by the vector $(\pi_1, \dots, \pi_K)$. Component parameters are assigned a prior distribution F_0, while the weights are assigned a symmetric Dirichlet prior, i.e., a Dirichlet distribution where all parameters are the same $\gamma_h = \gamma$ for all h.

3.1.1. Overfitted mixtures.

In practical examples, it is common to employ sparse mixture models where a fixed, overfitting value of K is chosen (Rousseau and Mengersen, 2011), along with a symmetric Dirichlet prior distribution for the weights with a small parameter γ. In this approach, although K is fixed, the number of clusters K_+ is a random variable because some components will have weight $\pi_j = 0$. Therefore, the number of clusters is identified as $K_+ = \{h : \pi_h > 0\}$. Grazian and Robert (2018) investigate the properties of Jeffreys' prior in this context and demonstrate consistent estimation of the number of clusters. The prior distribution induced on the random partition by a sparse mixture model approaches the Ewens distribution as $\gamma \to 0$. The Ewens distribution is the distribution induced on the partition by a Dirichlet process (see Section 3.3). Consequently, as we will see for the Dirichlet process, the estimation process for a sparse mixture model tends to concentrate on a large number of small clusters as n increases.

3.1.2. Static mixture models.

Alternatively, it is possible to consider a random number of components $K \sim p_K(k)$, where p_K is a probability mass function on $\mathbb{N}$ such that $\sum_{h=1}^{\infty} p_K(h) = 1$ and $p_K(h) > 0$ for $\forall h$ (Nobile, 1994). Including a prior distribution $p_K(k)$ has the effect that both K_+ and K are random a priori. Kruijer et al. (2010) and Nobile (1994) prove consistency for the number of components when the component distributions are correctly specified. The assumption of correct specification of the component distribution is quite strong, as using Gaussian components usually only approximates the true model of the observations. However, this approximation can lead to the estimation of an increasing number of components as n increases.

A finite mixture model with a prior distribution on the number of components, as in Equation (3), with a fixed value of γ for the Dirichlet prior on the weights, induces a valid EPPF that is available in closed form (Green and Richardson, 2001; McCullagh and Yang, 2008; Miller and Harrison, 2018):

$$p(\rho_n = (S_1, \dots, S_K)) = \sum_{\ell=1}^{\infty} \frac{\ell_{(K)}}{(\gamma\ell)^{(n)}} p_K(\ell) \prod_{s \in (S_1, \dots, S_\ell)} \gamma^{(|s|)} \tag{4}$$

where $b^{(m)} = b(b+1)\dots(b+m-1)$ and $b_{(m)} = b(b-1)\dots(b-m+1)$, $b^{(0)} = 1$ and $b_{(0)} = 1$. This model, with a fixed value of γ for the Dirichlet prior, which does not depend on the number of components, is called static mixture model.

Equation (4) reveals that the EPPF of a finite mixture model is a symmetric function of the cluster size, and the distribution of ρ_n is invariant under permutations of $[n]$. Using Equation (4), we can derive the distribution of the number of components K conditional on the number of clusters K_+:

$$p(K = k | K_+ = k_+) = \frac{1}{\sum_{\ell=1}^{\infty} \frac{\ell_{(k_+)}}{(\gamma\ell)^{(n)}} p_K(\ell)} \frac{k_{(k_+)}}{(\gamma k)^{(n)}} p_K(k)$$

and the distribution of the number of clusters conditional on the number of components:

$$p(K_+ = k_+ | K = k) = \frac{k_{(k_+)}}{(\gamma k)^{(n)}} \sum_{\mathbf{S}: |\mathbf{S}| = k_+} \prod_{s \in (S_1, \dots, S_{k_+})} \gamma^{(|s|)}.$$

where $\mathbf{S} = \{S_h : |S_h| > 0\}$. Finally, the conditional EPPF of a static mixture model is given by:

$$p(|S_1|, \dots, |S_k| | K_+ = k) = \frac{1}{Const_k} \prod_{h=1}^{k} \frac{\Gamma(n_h + \gamma)}{\Gamma(n_h + 1)}$$

where $Const_k$ is the normalizing constant obtained by summing over all labeled cluster sizes whose sum is equal to n. As expected, this conditional EPPF depends on γ, and for $\gamma = 1$, it represents the uniform distribution over all partitions.

Equation (4) is valid when choosing a symmetric Dirichlet prior distribution on $(\pi_1, \dots, \pi_K)$. The choice of the value of γ influences the entropy of the vector of weights: small values of γ are associated to a low entropy while large values of γ are associated to large entropy in $(\pi_1, \dots, \pi_K)$. In case of $\gamma = 1$, Gnedin (2010) derives a form for $p_K(k)$ and Stephens (2000) and Nobile (2004) propose $p_K(k)$ to be Poisson. In particular, when $p_K(k) = Pois(k-1|\lambda)$ and $\gamma = 1$, the finite mixture model has a stick-breaking representation (Argiento and De Iorio, 2022). Richardson and Green (1997) and Miller and Harrison (2018) use $\gamma = 1$ with a uniform prior

$p_K(k)$ over $\{1, 2, \ldots, K_{\max}\}$. Sparse finite mixtures can be considered a special case of finite mixture models with an unknown number of components, because $p_K(k) = \mathbb{I}\{k \leq K_{max}\}$ puts all prior mass on a fixed number of components K_{max}.

The use of $\gamma = 1$ can introduce bias in the estimation of the number of clusters. According to Frühwirth-Schnatter et al. (2021), when $\gamma = 1$, the expected value of K_+ tends to be close to the expected value of K for most of the available prior distributions $p_K(k)$ in the literature (such as Poisson, uniform, geometric, and beta-negative-binomial). However, Grazian et al. (2020) provide a decision-theoretic justification for using a beta-negative-binomial distribution with parameters $(1, \alpha, \beta)$. The choice of parameters allows control over the expectation and variance, but it is recommended to use a value of γ smaller than one, such as $\gamma = \frac{1}{2}$.

Furthermore, Gnedin and Pitman (2006) demonstrate that model (3), with an unknown number of components and a fixed parameter γ, is equivalent to a mixture model with an infinite number of components and a Gibbs-type prior on the random partition. This model is the only finite mixture that induces a Gibbs-type prior (see Section 3.3). Interestingly, in this case, as $n \to \infty$, the number of clusters K_+ behaves similarly to the number of components K:

$$|p(K_+ = k_+|y_1, \ldots, y_n) - p(K = k|y_1, \ldots, y_n)| \to 0 \qquad n \to \infty.$$

3.1.3. Dynamic mixture models.

Assuming the same γ for all K is a specific modeling choice that simplifies the implementation of known algorithms. To extend the static finite mixture model with a constant γ, McCullagh and Yang (2008) introduce the dynamic finite mixture model, where $\gamma_K = \alpha/K$. This means that the parameters of the Dirichlet distribution for the weights of the finite mixture model decrease as the number of components increases. The dynamic model reduces the impact of the experimenter's choice of γ. With increasing K, the symmetric Dirichlet distribution for the mixture weights becomes more concentrated around the boundary of the simplex, resulting in a more conservative estimation of K_+ and allowing the distribution of K_+ to differ from the distribution of K. Specifically, as γ_K increases, the prior variance of the mixture weights decreases, leading to more balanced weights. Conversely, as γ_K decreases (with a larger number of components), the prior variance increases, favoring more unbalanced weights.

For a dynamic mixture model, the EPPF can be expressed as:

$$p(\rho_n = (S_1, \ldots, S_{K_+})) = p_{DP}(\rho_n = (S_1, \ldots, S_{K_+})) \times \sum_{K=K_+}^{\infty} p_K(K) R_{K_+}^{K,\alpha}$$

where $p_{DP}(S_1, \ldots, S_{K_+})$ is the probability mass function of the Ewens distribution (which will be defined in Section 3.3), and

$$R_{K_+}^{K,\alpha} = \prod_{h=1}^{K_+} \frac{\Gamma(n_h + \frac{\alpha}{K})(K - h + 1)}{\Gamma(1 + \frac{\alpha}{K})\Gamma(n_h)K}.$$

The dynamic finite mixture model can be seen as a natural generalization of the Dirichlet process mixture model but does not fall into the class of Gibbs-type priors. Dynamic mixture models are characterized by a slower decrease in the difference between $\mathbb{E}[K_+]$ and $\mathbb{E}[K]$ as α increases compared to the static finite mixture model. This allows for larger differences even for large values of α because γ decreases as K increases, preventing K_+ from increasing too quickly.

The conditional EPPF of a dynamic mixture model can be expressed as:

$$p(\rho_n = (S_1, \ldots, S_k)|K_+ = k) = p(|S_1|, \ldots, |S_k||K_+ = k) = \frac{\sum_{\ell=1}^{\infty} p_K(\ell) \frac{b_{\ell,k}}{\Gamma(\frac{\alpha}{\ell})^k} \prod_{h=1}^{k} \frac{\Gamma(n_h + \frac{\alpha}{\ell})}{\Gamma(n_h+1)}}{\sum_{\ell=1}^{\infty} p_K(\ell) \frac{b_{\ell,k}}{\Gamma(\frac{\alpha}{\ell})^k} Const_k}$$

where $b_{\ell,k}$ is a constant depending on k, and $Const_k$ is the normalizing constant. Unlike the conditional EPPF of the static mixture model, this formula also depends on α and the prior distribution on the number of components, $p_K(\ell)$. This demonstrates that the dynamic mixture model offers more flexibility in defining the prior distribution on the number of clusters. However, the dependence on $p_K(\ell)$ implies that the choice of the prior distribution has a stronger impact on the induced prior distribution on the partitions.

3.1.4. Computational aspects.

In terms of computation, Richardson and Green (1997) introduce reversible jump Markov Chain Monte Carlo (RJMCMC) for mixtures with univariate Gaussian components and a fixed parameter γ that does not vary with K. Miller and Harrison (2018) generalize the Chinese restaurant process (CRP) sampler of Jain and Neal (2004) to the case of finite mixtures, where the number of components K is inferred, and the number of clusters K_+ is derived through post-processing. More recently, Frühwirth-Schnatter et al. (2021) introduce a telescoping algorithm, an MCMC algorithm for mixtures that updates the number of clusters K_+ and the number of components K simultaneously during the sampler without resorting to RJMCMC. The telescoping algorithm is implemented in the `fipp` `R` package (Greve et al., 2021).

3.2. Repulsive prior distributions

To reduce the number of estimated clusters, there are various approaches that can be taken. One approach is to modify the prior distribution for the number of components or the partition. Additionally, it is possible to define prior distributions for the component parameters in a way that favors well-separated components. When assuming independent and identically distributed parameters, components can be randomly located in the parameter space, potentially leading to components that are very close to each other. On the other hand, by introducing a repulsive prior distribution, dependence is introduced a priori among the parameters of a mixture model, particularly the location parameters. This results in the parameters no longer being conditionally independent.

To incorporate repulsion, methods such as those proposed by Quinlan et al. (2018) and Xie and Xu (2020) include a penalization term based on pairwise distances between the location parameters. Another approach, suggested by Malsiner-Walli et al. (2017), is to use repulsive mixtures, which encourage components to merge into groups at one hierarchical level while separating groups at another level. The properties of these prior distributions are further studied by Quinlan et al. (2021).

A repulsive distribution can be expressed as $\text{Rep}_K(\theta_K) = \frac{1}{Const_K} \left\{ \prod_{h=1}^{K} f_0(\theta_h) \right\} R_C(\theta_K)$, where f_0 is a probability density function and $Const_K$ is a normalizing constant. The function $R_C(\theta_K)$ is defined as

$$R_C(\theta_K) = \prod_{1 \leq r \leq s \leq K} [1 - C_0\{\rho(\theta_r, \theta_s)\}],$$

where the function $C_0 : [0, \infty) \to (0, 1]$ satisfies the following conditions: i) $C_0(0) = 1$, ii) $C_0(x) \to 0$ as $x \to \infty$, and iii) for any $x, z \geq 0$, if $x < z$ then $C_0(x) > C_0(z)$. The function C_0 is associated with the potential ϕ, given by

$$\phi(\theta_r, \theta_s) = -\log\{1 - C_0(\rho(\theta_r, \theta_s))\}.$$

By assuming a repulsive prior for the location parameters of a mixture model, an interaction structure among them is induced, which is defined by $\rho(\theta_r, \theta_s)$ through C_0. Various repulsive distributions have been proposed in the literature. For instance, Ogata and Tanemura (1985) introduce a repulsive distribution associated with soft repulsion, while Petralia et al. (2012) present a repulsive distribution with a potential, resulting in stronger repulsion. The choice of the repulsive distribution has an impact on the estimated number of clusters. Soft repulsion tends to eliminate singletons while maintaining good density estimation properties, whereas strong repulsion leads to a higher level of parsimony, with only a small number of estimated clusters.

The function $\rho(\cdot, \cdot)$ can be chosen as the Mahalanobis distance, and $C_0(a) = \exp\left(-\frac{1}{2}\frac{a^2}{\tau}\right)$, where τ is a parameter controlling the strength of repulsion. When $\tau \to 0$, the repulsion is weaker, while $\tau \to \infty$ leads to stronger repulsion. The parameter τ can be either estimated or fixed. Treating τ as an unknown parameter and assigning it a prior distribution significantly increases the computational burden and reduces the tractability of the posterior distribution. However, fixing τ may strongly influence the type of repulsion implied.

A repulsive prior distribution on the location parameters of a mixture model induces a prior distribution on the number of clusters, but its explicit expression is not available. Quinlan et al. (2021) prove that when the true cluster locations are separated by a minimum Euclidean distance that favors distinct clusters, and the prior assigns positive mass to arbitrarily small neighborhoods around the true density, the posterior rate of convergence relative to the L_1-metric is $\varepsilon_n = n^{-1/2}\log(n)^{1/2}$.

Alternatively, instead of repulsive priors, Fúquene et al. (2019) propose a "non-local prior" approach for selecting the number of components. A non-local prior distribution for the model with k components assigns vanishing probability as the mixture with k' components becomes equivalent to a mixture with k components when $k = k'$ or $\theta_h = \theta_j$ for some $h \neq j$. This prior distribution only requires identifiability of the model, meaning that $g(y|\theta_k, K = k) = g(y|\theta_{k'}, K = k')$ only when $k = k'$ and $\theta_k = \theta_{i(k')}$ for some permutation $i(k')$ of the component labels in the model with k' components. However, Fúquene et al. (2019) demonstrate through simulation studies and real datasets that this approach may be overly conservative, resulting in an underestimation of the number of components. Nevertheless, it appears to be more robust than other approaches to misspecifications of the component distributions (Cai et al., 2021).

3.3. Infinite mixture models

3.3.1. Dirichlet processes.

An alternative to the model given in Equation (2) is to consider a model with an infinite number of components. One of the main tools used in this context is the Dirichlet process. Dirichlet processes (DPs) are typically denoted as $F \sim DP(\alpha, F_0)$, where $\alpha > 0$ and F_0 is a distribution over $(\Theta, \mathcal{A})$, with Θ being a space and $\mathcal{A}$ a σ-field of subsets. F_0 is commonly referred to as the base distribution, representing the expected value of the process. The second parameter, α, is known as the concentration parameter and can be interpreted as a scale parameter related to the variance of the process.

The success of DPs in Bayesian analysis is due to their conjugacy: if $F \sim DP(\alpha, F_0)$, then the posterior distribution is also a Dirichlet process. It can be proven (Ferguson, 1973) that the support of the random variable F is almost surely the family of discrete distributions. Although this limits its density estimation

properties, this feature introduces a clustering ability in the DP, as demonstrated by the Pólya urn scheme of Blackwell and MacQueen (1973). Consider a sequence of independent and identically distributed (i.i.d.) variables $\theta_1, \theta_2, \ldots \sim F$ where $F \sim DP(\alpha, F_0)$. Since $\theta_{n+1}|F, \theta_1, \ldots, \theta_n \sim F$, for every measurable set $B \subset \Theta$, the predictive distribution of θ_{n+1} given $\theta_1, \ldots, \theta_n$ is

$$\theta_{n+1}|\theta_1, \ldots, \theta_n \sim \frac{1}{\alpha + n}\left(\alpha F_0 + \sum_{i=1}^{n} \delta_{\theta_i}\right). \tag{5}$$

This predictive distribution is a mixture of the base distribution F_0 and the empirical distribution of the atoms already drawn; here, δ_x represents the Dirac mass at x. In other words, it contains point masses at $\theta_1, \ldots, \theta_n$, which implies that there are ties in a sequence of n sequentially drawn atoms. Let $\theta_1^*, \ldots, \theta_{K_+}^*$ denote the unique values drawn from F and let $n_h = \#\{i : \theta_i = \theta_h^*\}$. Equation (5) can then be rewritten as

$$\theta_{n+1}|\theta_1^*, \ldots, \theta_{K_+}^* \sim \frac{1}{\alpha + n}\left(\alpha F_0 + \sum_{h=1}^{K_+} n_h \delta_{\theta_h^*}\right)$$

which demonstrates that the probability of allocating a new location to the h-th component increases as the number of atoms allocated to that component increases. For further information, refer to Aldous (1985) and Pitman (2002), which discuss a similar construction based on random partitions.

3.3.2. Dirichlet processes mixture models.

Mixing a Dirichlet Process (DP) with respect to kernels results in a countable mixture of distributions (Antoniak, 1974). Consequently, it is possible to model a set of random variables $Y_1, \ldots, Y_n$ (which can be possibly absolutely continuous) using the atoms of the Dirichlet process as latent parameters $\{\theta_1, \ldots, \theta_n\}$:

$$\begin{aligned} Y_i|\theta_i &\sim f(y_i|\theta_i) \qquad i = 1, \ldots, n \\ \theta_i|F &\sim F \\ F|\alpha, F_0 &\sim DP(\alpha, F_0), \end{aligned}$$

and since F is almost surely discrete, this model can be rewritten as

$$Y_i \sim \sum_{h=1}^{\infty} \pi_h f(y_i|\theta_h^*) \qquad i = 1, \ldots, n$$

where $\theta_1^*, \theta_2^*, \ldots$ are independent draws from the base distribution F_0. Thus, the clustering properties of the DP naturally extend to the case of DP mixture models.

The DP weights π_h can be constructed using a stick-breaking process (Sethuraman, 1994): $\pi_1 = V_1$ and $\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)$, where $V_1, V_2, \ldots$ are independent beta random variables $V_h \sim Be(a_h, b_h)$. Alternatively, the probabilities π_h can be drawn from any distribution on the simplex, as described in Ongaro and Cattaneo (2004). The DP prior is obtained when $V_h \sim Beta(1, \alpha)$ for all h. The Pitman-Yor (PY) process is obtained when $V_h \sim Be(1-\gamma_a, \gamma_b+\gamma_a h)$ for $\gamma_a \in [0, 1)$ and $\gamma_b > -\gamma_a$ for all h. When $\gamma_a > 0$, the expected value $\mathbb{E}[V_h]$ is a decreasing function of h. Consequently, sets with larger mass π_h are typically associated with smaller indexes h. This characteristic explains why the resulting partition follows heavy-tailed power-law distributions (Goldwater et al., 2006). In this regard, the PY process is more effective in estimating the number of rare or small clusters compared to a DP. When $\gamma_a = 0$, the DP is recovered with a concentration parameter of γ_b.

3.3.3. EPPF and PPF of Dirichlet and Pitman-Yor processes.

Even though the number of components in the DP (or any generalization of it) is infinite, only a random finite subset of components has a probability π_h greater than zero, denoted as $K_+ = \{h : \pi_h > 0\}$. Hence, it is possible to infer the number of underlying clusters, with the probability of assigning an observation to the h-th cluster given by π_h.

The explicit expression for the EPPF of the Dirichlet Process (DP) is available. If $F \sim DP(\alpha, F_0)$, then

$$P(\rho_n = (S_1, \ldots, S_{K_+})) = p(|S_1|, \ldots, |S_{K_+}|) = \frac{\alpha^{K_+} \prod_{h=1}^{K_+} (n_h - 1)!}{\prod_{i=1}^{n} (\alpha + i - 1)!}. \tag{6}$$

which is known as the Ewens distribution. This equation can be generalized to the case of the Pitman-Yor process (PY) as follows:

$$P(\rho_n = (S_1, \ldots, S_{K_+})) = p(|S_1|, \ldots, |S_{K_+}|) = \frac{\left(\prod_{h=1}^{K_+ - 1} (\alpha + h\gamma_a)\right) \left(\prod_{h=1}^{K_+} (1 - \gamma_a)^{(n_h - 1)}\right)}{(1 + \alpha)^{n-1}}.$$

and this distribution is known as the Ewens-Pitman distribution. Other Gibbs-type priors also have explicit EPPFs (Gnedin and Pitman, 2006; Ho et al., 2007; Lijoi et al., 2008; Gnedin, 2010; Cerquetti, 2013; De Blasi et al., 2015).

There is a second family of Pitman-Yor (PY) processes, where $b < 0$ and $a = K_+|b|$ with $K_+ \in \mathbb{N}$ (and known), (Gnedin, 2010; De Blasi et al., 2015). The EPPF of this representation of the PY process is given by:

$$p(\rho_n = (S_1, \ldots, S_{K_+})) = p(|S_1|, \ldots, |S_{K_+}|) = \frac{\Gamma(a)}{\Gamma(n + a)} \prod_{h=1}^{K_+} (a + b(h - 1)) \frac{\Gamma(n_h - b)}{\Gamma(1 - b)}.$$

A static finite mixture with K components and a symmetric Dirichlet prior on the weights, with all parameters equal to γ (fixed and known), is obtained by mixing a $PY(-\gamma, K\gamma)$ process prior over the concentration parameter $\alpha = K\gamma$, and fixing $b = -\gamma$ (Gnedin and Pitman, 2006). On the other hand, the dynamic finite mixture model is derived by mixing a $PY\left(-\frac{\alpha}{K}, \alpha\right)$ prior over $b = -\frac{\alpha}{K}$, while the concentration parameter is fixed at $a = \alpha$.

Also, the PPF of a DP is available in closed form, which is given by the Pólya urn representation (Blackwell and MacQueen, 1973):

$$p_j(n_1, \ldots, n_{K_+}) = \left(\frac{n_j}{n + \alpha}\right) \mathbb{I}\{1 \leq j \leq K_+\} + \left(\frac{\alpha}{n + \alpha}\right) \mathbb{I}\{j = K_+ + 1\}. \tag{7}$$

This expression can be generalized to the PY process $PY(\gamma_a, \gamma_b, F_0)$ as follows:

$$p_j(n_1, \ldots, n_{K_+}) = \left(\frac{n_j - \gamma_a}{n + \gamma_b}\right) \mathbb{I}\{1 \leq j \leq K_+\} + \left(\frac{\gamma_b + K_+\gamma_a}{n + \gamma_b}\right) \mathbb{I}\{j = K_+ + 1\}. \tag{8}$$

The conditional EPPF for a DP mixture model, induced for a given number of clusters $K_+ = k$, is given by:

$$p_{DP}(\rho_n = (S_1, \ldots, S_k) | K_+ = k) = p(|S_1|, \ldots, |S_{K_+}| | K_+ = k) = \frac{1}{Const_\infty} \prod_{h=1}^{k} \frac{1}{n_h},$$

where $Const_\infty$ is the normalizing constant taken as the summation with respect to all the labeled cluster sizes for which the sum is equal to n. This conditional EPPF favors unbalanced partitions with some small values of n_h due to the inverse dependence on n_h, for $h = 1, \ldots, k$ (Antoniak, 1974; Miller and Harrison, 2018). More precisely, this conditional EPPF has the form of a discrete Dirichlet distribution for a finite mixture model when $K_+ = k$, while it is an improper Dirichlet distribution (with all parameters equal to zero) in the case of an infinite mixture model. This means that as $n \to \infty$, the distribution of $(S_1, \ldots, S_k)|K_+ = k$ concentrates all its mass at the corners of the simplex in the case of a DP mixture model.

3.3.4. Inconsistency of the posterior distribution on the number of clusters.

For DP mixture models, Antoniak (1974) also provides the induced prior distribution on K_+ as $p_{K_+}(k_+) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \mathcal{S}_{n,k_+}$ where $\mathcal{S}_{n,k_+}$ is the Stirling number of the first kind. The number of clusters grows as $K_+ \sim \alpha \log(n)$. This reveals another perspective showing that the DP mixture model favors the estimation of many small clusters (Antoniak, 1974; Argiento et al., 2009; Onogi et al., 2011) when the concentration parameter is fixed.

These results have led Miller and Harrison (2014) to analytically prove that the posterior distribution on the number of clusters does not concentrate on any finite value as the sample size n increases:

$$\limsup_{n \to \infty} P(K_+ = k | y_1, \ldots y_n) < 1$$

with probability one. This holds for a large class of models, including DP and PY processes with components from a broad range of distribution families. Suppose ρ_n^k is a partition with k components, and define $\rho_n^{k'}$ as a partition with $k' = k + 1$ components. Then $\rho_n^{k'}$ can be generated by splitting one element in ρ_k to be in its own cluster (a singleton). The reason behind these results lies in the fact that the likelihood for a model with k clusters, $f(y_1, \ldots, y_n|(S_1, \ldots, S_k))$, is of the same order as the likelihood with $k' = k + 1$ clusters (where one observation is removed from an existing cluster to create a singleton), $f(y_1, \ldots, y_n|(S_1, \ldots, S_{k'}))$. However, the induced prior distribution on the partition provided by a PY process tends to favor models with additional clusters. Therefore, a PY mixture model tends to give preference to models with small clusters.

3.3.5. Other parameters.

Finally, the parameters of the base distribution F_0 and the concentration parameter α can also be considered as random variables with their own prior distribution. In particular, α plays a crucial role in the distribution induced on the number of clusters and partitions. Several works suggest using a Gamma distribution, $\alpha \sim \text{Ga}(a, b)$ (Escobar and West, 1995; Jara et al., 2007). The standard "non-informative" choice of setting a and b to be close to zero results in a highly informative prior for the number of clusters, with concentration around one and infinity (Dorazio, 2009). An alternative proposal by Frühwirth-Schnatter et al. (2021) is to use the F-distribution, $\alpha \sim \mathcal{F}(\nu_l, \nu_r)$, where the parameters control different characteristics of the prior distribution. A small ν_r gives fat tails, while a small ν_l tends to favor models with a small number of clusters.

3.3.6. Computational aspects.

Computationally, Ishwaran and James (2001) propose computational methods for the most general case of $V_h \sim \text{Beta}(a_h, b_h)$. Given the stick-breaking representation of the Dirichlet process, there are two main ways to perform posterior inference through Gibbs sampling. The first one is associated with the P'olya urn Gibbs sampler (Escobar, 1994; MacEachern, 1994; Escobar and West, 1995; MacEachern, 1998). The second one is the blocked Gibbs sampler proposed by Ishwaran and James (2001).

3.4. Dirichlet process mixtures in presence of covariates

DPs are based on the assumption that data are infinitely exchangeable, meaning that the ordering of data items does not matter. However, this assumption can be unrealistic, and many works have attempted to model more structured data. In particular, it is important to define a model in which the distribution F_x changes smoothly with respect to $x \in \mathcal{X}$, such that $F_{x_1} \to F_{x_2}$ as $x_1 \to x_2$. The dependent Dirichlet process (DDP) (MacEachern, 2000) is a generalization of the DP that creates a distribution on the set of countable mixture distributions. DDP introduces dependence among collections of distributions, where the dependence is driven by a covariate x. This is achieved by allowing the atoms θ_h^* to be replaced by a process $\theta_{\mathcal{X}}$, which provides the atom for each value of the covariate. Similarly, the random variable V_h in the stick-breaking construction can be replaced by a process $V_{\mathcal{X}}$, which determines the mass assigned to $\theta_{\mathcal{X}}$ at each level of the covariate. For a recent review on this topic, the reader is referred to Quintana et al. (2022). While highly investigated for density estimation, these processes are much less studied for clustering (while often used in practice for that purpose) and analytical expressions of the EPPF may not be available.

3.4.1. Single-π dependent Dirichlet processes.

An important class of DDPs is the single-π DDP, which offers a significant simplification in terms of computation. The key idea behind the single-π DDP is that the mass $\pi_h(x)$ does not vary with x. As a consequence of this restriction, the model can be regarded as a countable mixture of stochastic processes, with mixing weights that align with those of a single Dirichlet process model. The single-π DDP is useful for smoothing the prediction distribution across the covariate space. However, it is not suitable for clustering tasks because the DP probabilities are not dependent on the covariate.

3.4.2. Single-θ dependent Dirichlet processes.

Several authors have proposed an extension of the stick-breaking construction introduced by Sethuraman (1994) that allows the probabilities π_h to vary with the covariate. For example, Reich and Fuentes (2007), Dunson and Park (2008), and Warren et al. (2012) have explored this idea. In this extended framework, the random variable depending on covariates follows a model given by:

$$Y_i \sim \sum_{h=1}^{\infty} \pi_h(x) \delta_{\theta_h^*} \qquad i = 1, \dots, n \tag{9}$$

The weights $\pi_h(x)$ are constructed using a stick-breaking process, where $\pi_1(x) = V_1(x)$ and $\pi_h(x) = V_h(x) \prod_{j=1}^{h-1} (1 - V_j(x))$ for $h > 1$. However, in this extended construction, the variables $V_h(x)$ are allowed to vary according to a kernel function that smooths over the covariate space. Specifically, $V_h(x) = w_h(x) V_h$, where $V_h \sim \text{Beta}(a_h, b_h)$ and w_h is a kernel function that is constrained within the interval $[0, 1]$. This formulation incorporates dependence in the allocation probabilities, defining clusters characterized by kernel functions.

The model described in Equation (9) represents a case of the so-called "single-atom" DDP, where the atoms θ_h^* are independent with a marginal distribution F_0. Fuentes and Reich (2013) further extend this model by incorporating dependence among the atoms $\theta^*(x)$ of the DP using a Gaussian process as the base distribution.

The dependence imposed on the probabilities $\pi_1(x), \pi_2(x), \dots$ can also be described through a model. Chung and Dunson (2009) propose a construction that relies on a probit representation of the variables used to construct the clustering probabilities, instead of using beta random variables. On the other hand, Papageorgiou

et al. (2015) propose to directly model the mixture weights through a probit model. Ren et al. (2011) propose incorporating dependence on the weights of the mixture components through a logistic regression, via the use of a kernel depending on the distance between covariates level. Another possible construction uses geometric weights (Fuentes-García et al., 2009). Griffin and Steel (2006) define the mixing weights as transformations of i.i.d. random variables. However, they introduce dependence by inducing an ordering of the i.i.d. random variables at each covariate level, so that distributions at similar covariate levels are associated with similar orderings.

Sudderth and Jordan (2008) extend the Pitman-Yor (PY) process to incorporate information provided by a covariate through a latent variable with a thresholded Gaussian distribution. The priors on the stick-breaking proportions $V_h \sim \text{Beta}(1-\gamma_a, \gamma_b + h\gamma_a)$ are then transformed into corresponding random thresholds. When $\gamma_a = 0$, the model of Duan et al. (2007) is formally recovered. Rodríguez et al. (2010) propose a latent stick-breaking process in which observations at different spatial locations are dependent but share a common marginal distribution.

3.4.3. Difficulties in introducing dependence on covariates.

While stick-breaking methods are appealing from a computational perspective, they encounter a natural difficulty. The stick-breaking construction of clustering probabilities involves transforming the variables on which the dependence is defined. This transformation modifies the structure of dependence, thereby making it challenging to control the dependence among the clustering probabilities. For a discussion of this problem in the context of finite mixture models, refer to Mastrantonio et al. (2019). The vector of probabilities in stick-breaking methods is compositional, which complicates the interpretability of the dependence structure. Since the elements of a compositional vector are defined on the simplex, the covariance between each element h and the sum of all elements in any finite K-dimensional sequence of the Dirichlet Process (DP) is given by:

$$\text{Cov}(\pi_h, \pi_1 + \cdots + \pi_h + \cdots + \pi_K) = 0.$$

This is because $\pi_1 + \cdots + \pi_h + \cdots + \pi_K = 1$. Consequently, we have:

$$-\text{Var}(\pi_h) = \sum_{\substack{\ell=1 \\ h \neq \ell}}^{K} \text{Cov}(\pi_\ell, \pi_h).$$

In other words, at least one element on the right side of the equation must be negative, and correlations are not allowed to vary freely in the range $(-1, 1)$. Therefore, the sum-to-one constraint for any sub-sequence of the DP induces negative correlations among the probabilities (Aitchison, 1986).

In more detail, let $(\pi_1(x), \pi_2(x), \pi_3(x))^T$ be a subsequence of a DP. By definition, such a vector follows a Dirichlet distribution. Consider two values of the covariate, x_1 and x_2, and suppose the beta variables of the stick-breaking construction are described by the following matrix:

$$\begin{pmatrix} v_1(x_1) & v_2(x_1) \\ v_1(x_2) & v_2(x_2) \end{pmatrix}.$$

Let p_1 and P_1 denote the probability density function and the cumulative distribution function of $V_1(x)$, respectively. Similarly, let p_2 and P_2 denote the probability density function and the cumulative distribution function of $V_2(x)$. The joint cumulative distribution function of $(V_1(x_1), V_1(x_2))$ can be defined using a copula

representation C_1 with density c_1. Similarly, define C_2 and c_2 for the bivariate random variable $(V_2(x_1), V_2(x_2))$. Then, the joint density of the four variables can be written as:

$$p\begin{pmatrix} V_1(x_1) & V_2(x_1) \\ V_1(x_2) & V_2(x_2) \end{pmatrix}\begin{pmatrix} v_1(x_1) & v_2(x_1) \\ v_1(x_2) & v_2(x_2) \end{pmatrix} = c_1\begin{pmatrix} P_1(v_1(x_1)) \\ P_1(v_1(x_2)) \end{pmatrix} p_1(v_1(x_1))p_1(v_1(x_2))\cdot$$
$$\cdot\, c_2\begin{pmatrix} P_2(v_2(x_1)) \\ P_2(v_2(x_2)) \end{pmatrix} p_2(v_2(x_1))p_2(v_2(x_2)).$$

The vector $[V_1(x), V_2(x), \ldots]$ represents the vector on which the dependence is constructed. However, the focus is on the corresponding vector $[\pi_1(x), \pi(x), \ldots]$, which is associated with $[V_1(x), V_2(x), \ldots]$ through a one-to-one map from the space $\mathcal{V}^2$ to $[0,1]^3$. This mapping is achieved using a matrix of transformations B.

$$B = \begin{bmatrix} \pi_1(x_1) = v_1(x_1) & \pi_1(x_2) = v_1(x_2) \\ \pi_2(x_1) = v_2(x_1)(1 - v_1(x_1)) & \pi_2(x_2) = v_2(x_2)(1 - v_1(x_2)) \\ \pi_3(x_1) = 1 - \pi_1(x_1) - \pi_2(x_1) & \pi_3(x_2) = 1 - \pi_1(x_2) - \pi_2(x_2) \end{bmatrix}.$$

The two-dimensional space of (V_1, V_2) is expanded to the three-dimensional space of (π_1, π_2, π_3). By employing a change of variables, it becomes possible to derive the joint density of the probability vectors

$$p\begin{pmatrix} \pi_1(x_1) & \pi_2(x_1) & \pi_3(x_1) \\ \pi_1(x_2) & \pi_2(x_2) & \pi_3(x_2) \end{pmatrix} = c^*\begin{pmatrix} P_1(\pi_1(x_1)) & P_2\left(\frac{\pi_2(x_1)}{1-\pi_1(x_1)}\right) & P_g(g^{-1}(\pi_3(x_1))) \\ P_1(\pi_1(x_2)) & P_2\left(\frac{\pi_2(x_2)}{1-\pi_1(x_2)}\right) & P_g(g^{-1}(\pi_3(x_2))) \end{pmatrix}\cdot$$
$$\cdot p_1(\pi_1(x_1))p_1(\pi_1(x_2))p_2\left(\frac{\pi_2(x_1)}{1-\pi_1(x_1)}\right)p_2\left(\frac{\pi_2(x_1)}{1-\pi_1(x_1)}\right)\cdot p_g(g^{-1}(\pi_3(x_1)))h_g(g^{-1}(\pi_3(x_2)))\cdot|\mathbb{J}(B^{-1})|.$$

Here, c^* represents the copula of the augmented three-dimensional variable, and $|\mathbb{J}(B^{-1})|$ denotes the determinant of the Jacobian matrix of the inverse one-to-one map. The determinant of the Jacobian matrix depends on functions of $(\pi_1(x_1), \pi_2(x_1), \pi_1(x_2), \pi_2(x_2))$, given the deterministic definition of $(\pi_3(x_1), \pi_3(x_2))$. As a result, the structure of the dependence, described by the copula density $c^*(\cdot)$, is altered.

To illustrate this characteristic, Figure 1 presents the results of a simulation involving $V_h(x)$ drawn from beta distributions, with dependence expressed through a Clayton copula parameterized at one. These variables have been transformed into compositional vectors using the stick-breaking construction. The scatterplots resulting from 10^5 simulations are displayed in Figure 1. The left side depicts the variables $V_1(x)$ and $V_2(x)$, while the right side illustrates the variables $\pi_1(x)$ and $\pi_2(x)$. As $\pi_3(x)$ is deterministically derived, it is not shown. It is apparent that the dependence structure originally present in the beta random variables is not preserved in the probabilities.

3.4.4. Distance dependent Chinese restaurant process.

Alternatively, Blei and Frazier (2011) introduce the distance dependent Chinese restaurant process (ddCRP), which directly models the probability of assigning observations to available clusters. The underlying assumption of the ddCRP is that data points that are close to each other, based on some form of distance, are more likely to be clustered together. Consequently, the assignment is based on the distances between observations, connecting each observation with others rather than with the atoms of the DP.

Let C_i denote the assignment for the i-th observation, and d_{ij} be a distance measure between observation i and observation j. D represents the set of all distance measurements between observations, and m is a decay

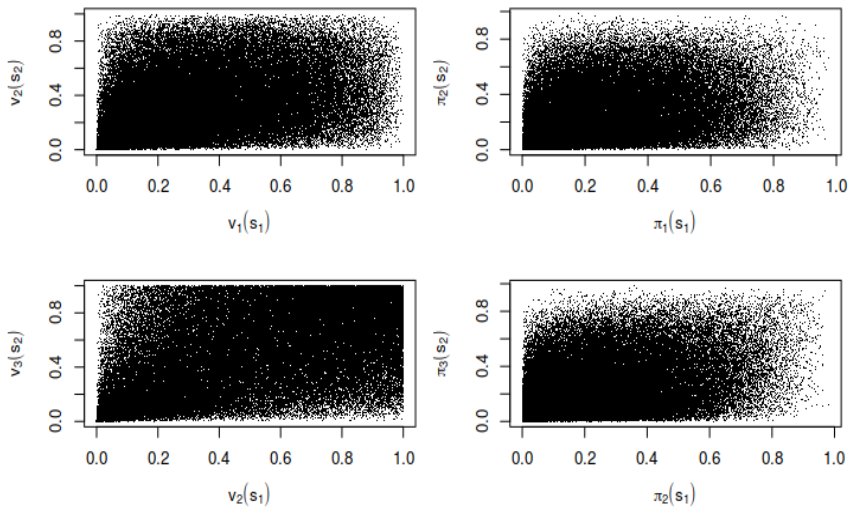

FIGURE 1. *Experiment to show the change in the dependence structure implied by the stick-breaking construction: on the left side the beta random variables are shown, on the right side the probabilities obtained through the stick-breaking transformation of the beta variables are shown.*

function. The probability of assigning observation i to cluster j given the distances d_{ij}, the decay function m, and the concentration parameter α, is proportional to:

$$p(C_i = j|\{d_{ij}\}, m, \alpha) \propto \begin{cases} m(d_{ij}) & j \neq i \\ \alpha & j = i. \end{cases}$$

The assignment of an observation depends solely on the distance d_{ij}, which can represent various measures such as time difference, Euclidean distance for spatial points, and so on. The choice of $m(\cdot)$ determines the behavior of the process and should possess several key properties, including being non-increasing, non-negative, having finite values, and satisfying $m(\infty) = 0$.

Additionally, Ghosh et al. (2011) define a hierarchical version of the ddCRP that clusters groups of observations. This hierarchical version allows for sharing of cluster components across groups, and within-group clustering depends on distances among locations.

3.4.5. Dirichlet process mixtures for spatial data.

Dirichlet process mixtures have also been widely applied to spatial data. The formal definition presented in the previous sections remains valid, where the spatial location serves as the covariate. For instance, in the work of Gelfand et al. (2005), the atoms θ_h are allowed to depend on the location s, denoted as $\theta_h = \theta_h(s)$. Spatial

dependence is then introduced through a Gaussian process, resulting in the representation:

$$F(s) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_{hD}}.$$

Since the probabilities π_h do not depend on the locations $s \in D$, the distribution $F(s)$ follows a Dirichlet process, specifically $F(s) \sim DP(\alpha, F_0)$, where F_0 represents an n-variate distribution. In this context, $F(s)$ represents a random distribution for each level s, and each realization θ_{hD} corresponds to a surface over the spatial domain D.

Both Duan et al. (2007) and Gelfand et al. (2007) extend the Spatial Dirichlet Process (SDP) by considering varying mixture weights, which allow observations to depend on different surfaces at different locations. Here, the weights can be defined to allow for site-specific selection of surfaces, enabling similar weights to be assigned to sites that are close together. It is worth noting that due to the lack of techniques for deriving a posterior distribution for the weights, this process is not suitable for clustering purposes. However, this limitation allows for prediction to be performed.

One of the main limitations of single-θ processes is that such models require replications at each spatial location. Consequently, these models are not suitable when the data represent only one surface. To address this limitation, Reich and Fuentes (2007) propose a spatial extension of the stick-breaking construction that allows the probabilities π_h to vary spatially without relying on the presence of replications. This model is a case of single-π process, where probabilities are defined through a kernel function smoothing over space. Grazian (2023) proposes a spatio-temporal extension of this model, also comparing a single-θ model with a model when both the weight and the atoms are allowed to depend on space and time. While the conceptual introduction of dependence on both the atoms and the weights of the process may seem straightforward, the computational time required increases drastically.

Another computational limitation of the kernel stick-breaking prior is related to the dimensionality of the data. To make the algorithm feasible, these approaches usually select a limited area around each point location. In order to perform dimension reduction, an approach developed by Reich et al. (2012) based on Bayesian variable selection can be employed, where only informative spatial locations are included in the definition of the kernels.

Although the method implemented by Reich and Fuentes (2007) is designed for point-referenced data, generalizations to account for areal data have also been proposed, such as the areally-weighted stick-breaking process and the areally-referenced Dirichlet process by Li et al. (2010). Similar methodologies can be applied to model spatial-varying regression coefficients (Cai et al., 2013).

4. Prior distributions on the partitions

Using an induced model on the partitions may have some drawbacks: spatial correlation can sometimes be counter-intuitive (Wall, 2004), and local features may not be adequately captured. Alternatively, it is possible to directly define a prior distribution on the partition by re-expressing a Gibbs-type prior model as follows:

$$\begin{aligned} Y_i|\theta^*_{C_i}, \rho_n &\sim f(y_i|\theta^*_{C_i}) \qquad i = 1, \dots, n \\ \theta^*_1, \dots, \theta^*_K &\sim F_0 \\ C_i = h &\Leftrightarrow i \in S_h \qquad i = 1, \dots, n \text{ and } h = 1, \dots, K \\ \rho_n &\sim p(\rho_n). \end{aligned}$$

In this model, the prior distribution $p(\rho_n)$ is defined directly. However, this construction does not guarantee sample size consistency and may not result in a valid EPPF.

4.1. Product partition models

A popular method for defining a prior distribution directly on partitions is provided by product partition models (Hartigan, 1990; Barry and Hartigan, 1992). In this approach, the prior distribution is given by:

$$p(\rho_n = (S_1, \dots, S_K)) = c_0 \prod_{h=1}^{K} c(S_h).$$

Here, $c(S_h) \geq 0$ is a non-negative function called the cohesion of S_h, which represents a measure of the strength of the prior assumption that elements in S_h should be clustered together. The constant c_0 is a normalizing constant that sums over all possible partitions.

The prior distributions for the parameters of the model can be defined as:

$$\pi(\theta_1^*, \dots, \theta_K^*) = c_0 \sum_{(S_1,\dots,S_K)\in\mathcal{P}_n} \prod_{h=1}^{K} c(S_h)\pi_h(\theta_h).$$

If the subjects are exchangeable, product partition models are conjugate, and the posterior distribution $p(\rho_n|y_1, \dots, y_n)$ is still a product model

$$p(\rho_n = (S_1, \dots, S_K)|y_1, \dots, y_n) = c_0' \prod_{h=1}^{K} c(S_h) f(y_h^*|\theta_{S_h}^*)$$

where $y_h^* = \{y_i, i \in S_h\}$. Additionally, they have the property of sample size consistency, i.e., the prior distribution of the partition of the first $n-1$ indices can be obtained by integrating the joint distribution with respect to the n-th index.

Quintana and Iglesias (2003) demonstrate the connection between product partition models and DP mixture models. The Pólya urn representation of the DP implies that the induced prior distribution for the partition ρ_n is given by:

$$p(\rho_n) = \frac{\alpha^K \prod_{h=1}^{K} \Gamma(|S_h|)}{\prod_{i=1}^{n}(\alpha + i - 1)} \propto \prod_{h=1}^{K} \alpha\Gamma(|S_h|)$$

where α is the mass parameter of the DP. This representation is proportional to the product over partition components, with $c(S) = \alpha\Gamma(|S|)$. Thus, the DP can be viewed as a product partition model, and as a result, product partition models yield a valid EPPF.

4.2. Product partition models in presence of covariates

4.2.1. Using a similarity function.

The cohesion function $c(S_h)$ can be modified to include an additional regression function. Let $w(x_h^*)$ denote a nonnegative similarity function that measures the homogeneity of x_i within cluster S_h, where $x_h^* = \{x_i, i \in S_h\}$. The modified expression for the prior distribution of the partition ρ_n conditioned on the covariates $x_1, \dots, x_n$ is:

$$p(\rho_n|x_1, \dots, x_n) = c_{0x} \prod_{h=1}^{K} c(S_h) \cdot w(x_h^*)$$

where c_{0x} is the normalization constant. The similarity function $w(\cdot)$ introduces a penalty for cluster size, such that $\lim_{n_h \to \infty} w(x_h^*) = 0$. To facilitate calculations, $w(\cdot)$ can be defined by marginalizing over an auxiliary model $q(\cdot)$:

$$w(x_h^*) = \int \prod_{i \in S_h} q(x_i|\xi_h) q(\xi_h) d\xi_h.$$

where ξ_h represents a set of parameters of the auxiliary model. It is important to note that this representation does not necessarily imply that the covariates are random, but it is a convenient way to introduce correlations among similar values of the covariates. Under the assumptions of i) symmetry with respect to permutations and ii) scaling across sample size, which means the similarity of any cluster is the average of the augmented clusters, i.e., $w(x^*) = \int w(x^*, x) dx$, it can be proven that $w(x^*)$ is necessarily proportional to the marginal distribution of x^* under a hierarchical auxiliary model. This is a direct application of De Finetti's representation theorem. In practical applications, the auxiliary model $q(\cdot)$ can be chosen in a conjugate form to evaluate the integral analytically.

In the case where p covariates are available, the similarity function $w(\cdot)$ can be easily extended to include all covariates as $w(x_{h1}^*, \ldots, x_{hp}^*) = \prod_{\ell=1}^p q_\ell(x_{h\ell}^*)$, where $x_{h\ell}^* = \{x_{i\ell} : i \in S_h\}$. This allows for incorporating multiple covariates into the similarity measure. The random partition model maintains coherence across different sample sizes when observations are independent across clusters and exchangeable within clusters.

4.2.2. Using a covariate-dependent cohesion function.

Park and Dunson (2010) generalize the previous class of product partition models by introducing a new definition:

$$p(\rho_n = (S_1, \ldots, S_K)|x_1^*, \ldots, x_K^*) \propto \prod_{h=1}^K c(S_h, x_h^*)$$

for $h = 1, \ldots, K$. The posterior distribution of $(S_1, \ldots, S_K)$ remains a product partition model with an updated cohesion function. Similar to Müller et al. (2011), there is a direct influence of the covariates on the definition of the partition distribution. This representation remains sample size consistent.

The predictive model follows a similar approach: when considering a new observation $(n+1)$, it is assigned to either a new cluster or one of the existing clusters. The assignment probabilities are proportional to the marginal likelihoods evaluated at the covariate value of the new observation. These probabilities can vary across clusters, indicating that if the covariate value of observation $n+1$ is close to the value x_h^* of a subject in cluster h, then the subject will be allocated with a higher probability to cluster h.

Similar to Müller et al. (2011), the covariates x_i are assumed to follow an auxiliary model, which can be chosen to be conjugate for computational efficiency.

4.2.3. Spatial extensions.

In a spatial setting, Page and Quintana (2016) propose a flexible location-dependent product partition model that incorporates spatial information by considering the likelihood of assigning locations that are far apart to the same cluster. Let $s_1, \ldots, s_n$ denote n locations, which can be either two-dimensional coordinates or areal locations. To extend the product partition model to include spatial information, the cohesion function can be modified as follows:

$$p(\rho_n|s_1^*, \ldots, s_K^*) \propto \prod_{h=1}^K c(S_h, s_h^*).$$

where $s_h^* = \{s_i : i \in S_h\}$. One possible approach is to define the cohesion function similar to the one used in the DP:

$$c(S, s^*) = \begin{cases} \alpha \times \Gamma(|S|) & \text{if } S \text{ is spatially connected} \\ 0 & \text{otherwise.} \end{cases}$$

However, this model is computationally challenging to approximate. Page and Quintana (2016) propose four alternative functions that define the cohesion as a decreasing function of the distance between locations. Some of these functions exhibit sample size consistency, while others do not. Unlike the DP, Page and Quintana (2016) do not derive an analytic formula for the expected number of clusters in the spatial extension of the product partition models because the expectations depend on the distances among locations. However, they provide experimental results demonstrating that the expected number of clusters may grow at a slower or faster rate compared to a standard DP, depending on the specific cohesion function chosen.

4.3. Alternatives to product partition models

Alternatively to product partition models or induced partitions, it is possible to define other distributions for the random partition. There are several possible prior distributions to consider.

4.3.1. Uniform prior.

The simplest one is the uniform prior:

$$p(\rho_n = (S_1, \dots, S_K)) = \frac{1}{\mathcal{B}_n}$$

where $\mathcal{B}_n$ is the Bell number.

4.3.2. Predictive approach.

Jensen and Liu (2008) propose a prior distribution that favors the allocation of a new observation to already existing clusters. It is defined as follows:

$$P(C_{n+1} = h | C_1, \dots, C_n, \alpha) = \frac{n_h}{\alpha + n}$$
$$P(C_{n+1} = \text{new} | C_1, \dots, C_n, \alpha) = \frac{\alpha}{\alpha + n}$$

where α is the mass parameter of a DP.

4.3.3. Hierarchical prior.

Casella et al. (2014) propose a hierarchical uniform prior where the prior distribution on the partitions is conditioned on the number of components, which influences the number of clusters. The prior distribution is defined as:

$$p(K, C_1, \dots, C_n) = p(C_1, \dots, C_n | K) p(K).$$

The prior distribution on $(C_1, \dots, C_n)$ can be chosen to be uniform, while $p(K)$ can be chosen in such a way that it assigns small support to the case where $K = n$.

4.3.4. Ewens-Pitman prior.

Dahl et al. (2017) propose an Ewens-Pitman attraction (EPA) distribution, which allocates observations based on their "attraction" to existing clusters. The attraction to a given cluster is determined by pairwise similarities between the current observation and the observations already in the cluster. The allocation process sequentially assigns items to subsets, creating a partition. Let $(j_{(1)}, \dots, j_{(n)})$ be permutations of $1, \dots, n$ such that the i-th allocated observation is j_i. The resulting partition of n observations has $K^{(i)}$ subsets. To make allocation decisions, a similarity function $\lambda(j_{(i)}, j_{(\ell)})$ is required. It is common to define the similarity function as a function of the distance between observations, i.e., $\lambda(j_{(i)}, j_{(\ell)}) = f(d_{i\ell})$, where f is a non-increasing function. As the function $f(d) \to 0$, the EPA distribution becomes increasingly different from the Ewens distribution, which is the partition distribution of a DP. This means that the EPA distribution favors partitions that group items with small distances, contrasting the behavior of the DP.

The EPA distribution is defined as the product of increasing conditional probabilities:

$$p(\rho_n = (S_1, \dots, S_K)) | \alpha, \delta, \lambda, (j_{(1)}, \dots, j_{(n)})) = \prod_{i=1}^{n} p_i(\alpha, \delta, \lambda, (j_{(1)}, \dots, j_{(n)}))$$

where

$$p_i(\alpha, \delta, \lambda, (j_{(1)}, \dots, j_{(n)})) = \begin{cases} \frac{i-1-\delta K^{(i-1)}}{\alpha+i-1} \frac{\sum_{j_\ell \in S} \lambda(j_i, j_\ell)}{\sum_{\ell=1}^{i-1} \lambda(j_i, j_\ell)} & S \in \rho_n^{(i-1)} \\ \frac{\alpha+\delta K^{(i-1)}}{\alpha+i-1} & S \text{ being a new subset.} \end{cases}$$

Here, S represents the cluster to which observation i is allocated, and $\rho_n^{(i-1)}$ is the partition of the first $(i-1)$ observations. The ratio of the similarity functions provides the "attraction" of j_i to the observations allocated to S. The distribution is invariant to permutations of the observations and also to scale changes in the similarity function λ. The parameter $\delta \in [0, 1)$ is a discount parameter, while α is a concentration parameter.

The EPA distribution also produces a probability distribution on the number of clusters K_+, which can be derived in closed form. This distribution does not depend on the similarity $\lambda(j_{(i)}, j_{(\ell)})$ or the permutation $(j_{(1)}, \dots, j_{(n)})$. The expected number of clusters is $\mathbb{E}[K_+] = \sum_{i=1}^{n} w_i$, where $w_1 = 1$ and $w_i = \frac{\alpha+\delta \sum_{\ell=1}^{i-1} w_\ell}{\alpha+i-1}$ for $i > 1$. As n increases, the average number of clusters increases, with the rate of growth depending on α and δ. In particular, as α and δ approach zero, the average number of clusters increases more slowly.

The PPF of the EPA distribution is given by:

$$p_j(n_1, \dots, n_K) = \left(\frac{n_h - \delta K n_h}{n + \alpha} \right) \mathbb{I}\{1 \le h \le K\} + \left(\frac{\alpha + \delta K}{n + \alpha} \right) \mathbb{I}\{h = K + 1\}.$$

For $\delta = 0$ and $\lambda(j_{(i)}, j_{(\ell)})$ constant for all i, ℓ, this PPF corresponds to the PPF of a DP, i.e., the Ewens distribution; see Equation (7). However, there is no way to recover the PPF of the Ewens-Pitman distribution; see Equation (8). The EPA distribution applies the discount δ proportionally to the relative size of the cluster and the number of clusters, whereas the Ewens-Pitman distribution applies a uniform discount γ_a to small and large clusters. This difference increases the entropy of the derived partitions for the EPA distribution and decreases the proportion of singletons. However, the distribution on the number of clusters is the same for both distributions.

Similarly to product partition models, the EPA distribution is symmetric. However, unlike product partition models, it is not marginally invariant. Nevertheless, in product partition models, the hyperparameters are often fixed to constant values for computational feasibility. On the other hand, the EPA model can easily treat the

hyperparameters as random variables and estimate them from the data points. Furthermore, a characteristic of the EPA distribution is that it allocates probability among partitions within a given number of clusters, but it does not redistribute probability among sets of partitions with different numbers of clusters.

4.3.5. Informative prior distributions.

None of these prior distributions allows for the incorporation of prior information about the grouping. Paganin et al. (2021) propose an approach where it is possible to incorporate prior information on the partition. The prior for $S = (S_1, \ldots, S_K)$ can be defined as:

$$p(S|S^{(0)}) \propto p_0(S^{(0)}) e^{-d(S,S^{(0)};\psi)}$$

where $S^{(0)} = (S_1^{(0)}, \ldots, S_1^{(0)})$ and ψ is a penalisation parameter. As ψ increases, the model favors partitions S that are similar to $S^{(0)}$, but not in a uniform way, as the space of partitions $\mathcal{P}_n$ is not uniform. In other words, for a fixed configuration, there is a heterogeneous number of partitions. The function $d(\cdot, \cdot)$ is a distance measure on the partitions, such as the Variation of Information by Meilă (2007), and $p_0(S^{(0)})$ is a baseline EPPF. The baseline EPPF can be chosen as the uniform EPPF:

$$p_0(S^{(0)}) = \frac{1}{\mathcal{B}_n}$$

where $\mathcal{B}_n$ is the Bell number. The concentration around $S^{(0)}$ depends on n.

The distance $d(S, S^{(0)})$ takes a finite number of discrete values $\tau_1, \ldots, \tau_L$, where L depends on $S^{(0)}$ and $d(\cdot, \cdot)$. Let $S_\ell(S^{(0)}) = S \in \mathcal{P}n : d(S, S^{(0)}) = \tau\ell$ be the set of partitions having the same fixed distance from $S^{(0)}$. The exponential term penalizes partitions in the same set $S_\ell(S^{(0)})$ equally for a given τ_ℓ, but the resulting probability may differ depending on the baseline EPPF $p_0(S^{(0)})$.

5. Clustering populations

Several hierarchical models have been introduced based on the DP to cluster observations at multiple levels. To address the clustering of populations, it is beneficial to introduce the concept of partial exchangeability. A set of random variables $(Y_{1,1}, Y_{1,2}, \ldots, Y_{2,1}, Y_{2,2}, \ldots)$ is considered partially exchangeable if, for all sample sizes $n_1, n_2 \geq 1$, and all permutations $(i_{(1)}, \ldots, i_{(n_1)})$ and $(j_{(1)}, \ldots, j_{(n_2)})$ of $(1, 2, \ldots, n_1)$ and $(1, 2, \ldots, n_2)$, the distribution

$$f(y_{1,1}, \ldots, y_{1,n_1}, y_{2,1}, \ldots, y_{2,n_2}) = f(y_{1,i_{(1)}}, \ldots, y_{1,i_{(n_1)}}, y_{2,j_{(1)}}, \ldots, y_{2,j_{(n_2)}}),$$

remains the same. In other words, the distribution of joint samples remains invariant under permutations within each sample. The concept of partial exchangeability can be extended by considering random variables Y_{ij}, where $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$, and each $Y_{ij} \sim f_j$, with $f_j \sim H$ epresenting a prior on the space of random measures. The entire sequence of random variables is exchangeable if the probability measure H assigns probability one to $\{(f_1, \ldots, f_J) \in \mathcal{F}_{\mathcal{Y}}^J : f_1 = f_2 = \ldots = f_J\}$. The opposite of exchangeability is independence. However, in practical situations, it can be useful to consider intermediate scenarios where random measures are similar but not exactly identical. The situation of partial exchangeability is associated with an allocation distribution described by a partial EPPF or pEPPF.

5.1. Different levels of clustering

5.1.1. Focus on the first level of clustering.

Reich and Bondell (2011) utilize a separable structure to identify genomic clusters and their relationship to spatial locations in order to investigate the interplay between natural selection and environmental factors. In their model, the allele frequencies for individual i at locus ℓ follow a multinomial distribution, conditioned on the cluster assignment:

$$Y_{i\ell}|C_i = c_i \sim \mathcal{M}\text{ultinomial}(2, \omega_{c_i\ell})$$

where $\omega_{c_i\ell}$ represents the vector of allele probabilities at locus ℓ in cluster c_i. These probabilities are assigned a stick-breaking prior. Separately, the spatial locations are modeled nonparametrically, conditioned on the cluster assignment:

$$\begin{aligned} s_i|C_i = c_i &\sim F_{c_i} \\ f_{c_i}(s_i) &= \sum_{h=1}^{\infty} \pi_{c_i h} \mathcal{N}(s_i|\mu_{c_i h}, \Sigma_{c_i}). \end{aligned}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean vector μ and covariance matrix Σ. Lastly, the cluster assignment is modeled as a categorical variable:

$$C_i \sim \mathcal{C}\text{at}(q_1, \dots, q_K),$$

where K could potentially be infinite and the probabilities $q_1, \dots, q_K$ are assigned a stick-breaking construction. In the study by Reich and Bondell (2011), there are multiple clustering features at each step, but the primary focus lies in the clustering of spatial locations, characterized by the allocation variable $(C_1, \dots, C_n)$.

5.1.2. Clustering for meta-analysis.

Müller et al. (2004) introduce a hierarchical approach for meta-analysis problems. The random distribution of observations is defined as a mixture of a common measure F_0, representing the shared component across all populations, and a random measure F_j specific to population j. In this framework, the model can be described as follows:

$$\begin{aligned} Y_{1,j}, \dots, Y_{n_j,j} &\sim f(y_{1,j}, \dots, y_{n_j,j}|\theta_{ij}) \qquad j = 1, \dots, J \\ \theta_{ij} &\sim H_j \qquad i = 1, \dots, n_j, \quad j = 1, \dots, J \\ H_j &= \gamma F_0 + (1-\gamma) F_j \\ F_j &\sim \text{discrete RPM}, \end{aligned} \tag{10}$$

where $0 \leq \gamma \leq 1$ represents the weight determining the dependence among populations and the amount of information borrowed by the estimation procedure from other probability measures. Thus, all data contribute to learning F_0, while $(y_{1,j}, \dots, y_{n_j,j})$ contributes to the specific learning of F_j. Kolossiatis et al. (2013) propose selecting γ in a way that ensures H_j is marginally a DP.

As a prior distribution for γ, Müller et al. (2004) suggest:

$$p(\gamma) = w_0 \delta_0(\gamma) + w_1 \delta_1(1-\gamma) + (1 - w_0 - w_1) Beta(\gamma|a_\gamma, b_\gamma)$$

where w_0 and w_1 assign non-zero probability to $\gamma = 0$, corresponding to independent H_j, and $\gamma = 1$, corresponding to exchangeable observations across populations. Additionally, γ can follow a beta distribution with parameters a_γ and b_γ with positive probability. A similar model has been employed by Wang et al. (2019) to combine information from randomized and registry studies for causal inference. There exist several generalizations of this approach.

5.1.3. Extensions of Müller et al. (2004).

Dunson (2006) extends the work of Müller et al. (2004) by introducing latent trait distributions. Caron et al. (2007) propose an alternative approach to incorporate temporal dependence using a generalized Pólya urn, which represents time-varying DP mixtures. Caron et al. (2014) utilize a mixture of DPs for heterogeneous ranking data with nonparametric Plackett-Luce components, where each component was parameterized by a random measure, such as a gamma process. Billio et al. (2019) suggest employing model (10) for parameter blocks in high-dimensional vector autoregressive models.

A characteristic of model (10) is that the atoms are different for each population, even if they originate from the same component in the mixture. Conversely, Gutiérrez et al. (2019) propose an approach where the atoms for two populations can assume the same values, which is particularly useful when testing equality between two or more random measures. Finally, Lijoi et al. (2014) propose a model where F_0 and F_j are independent normalized completely random measures (NCRM). Although these two approaches may appear similar, the approach proposed by Lijoi et al. (2014) cannot be interpreted as a generalization of the approach proposed by Müller et al. (2004). The main difference lies in the fact that the measures H_j in Müller et al. (2004) are not guaranteed to be marginally DPs, while in the approach by Lijoi et al. (2014), they are ensured to be marginally NCRM.

5.2. Nested processes

The models presented in Section 5.1 lack formal definitions of the induced partition model. One possible approach to defining partitions of populations is through nested models, which involve nesting discrete random probability measures. Nested DPs have been introduced by Rodriguez et al. (2008) to perform both clustering among observations and clustering among distributions. For the case of $d = 2$ populations, the model is defined as:

$$\begin{aligned}
(Y_{i_1,1}, Y_{i_2,2}) | f_1, f_2 &\overset{i.i.d}{\sim} f_1 \times f_2 \qquad i_1 = 1, \dots, n_1, \quad i_2 = 1, \dots, n_2 \\
f_1, f_2 | H &\overset{i.i.d}{\sim} H \\
H &\overset{d}{=} \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*} \\
\theta_h^* &\overset{i.i.d}{\sim} F = \sum_{\ell=1}^{\infty} w_\ell \delta_{\psi_\ell^*} \\
\psi_\ell^* &\overset{i.i.d}{\sim} F_0 = DP(\alpha, F_{00}).
\end{aligned}$$

Here, for $h = 1, 2, \dots$, π_h are independent of θ_h^*, and for $\ell = 1, 2, \dots$, w_ℓ are independent of ψ_ℓ^*. The extension to J populations is straightforward. Rodriguez et al. (2008) propose using DPs at both levels of the hierarchy, but other processes can also be employed. The advantage of using DPs is that the weights at both levels can

be constructed using the stick-breaking representation, which offers computational efficiency. As H is almost surely discrete, f_1 and f_2 can be equal with positive probability, implying that $Y_{i_1,1}$ and $Y_{i_2,2}$ can have the same distribution.

Camerlenghi et al. (2019a) demonstrate that nested DPs are unable to flexibly and realistically cluster populations. Specifically, if f_1 and f_2 have at least one common atom, the posterior distribution of (f_1, f_2) degenerates to the case where $f_1 \stackrel{d}{=} f_2$. This characteristic, where clusters can be either entirely common among populations or entirely distinct, is not unique to nested DPs but is present in all nested processes.

To address this degeneracy issue, Camerlenghi et al. (2019a) introduce latent nested processes where the nesting structure is applied to the underlying completely random measures. This allows for a representation in which each distribution f_j can be expressed as a mixture:

$$f_j = \frac{\mu_j + \mu_S}{\mu_j(\mathcal{Y}) + \mu_S(\mathcal{Y})} = \gamma_j \frac{\mu_j}{\mu_j(\mathcal{Y})} + (1-\gamma_j)\frac{\mu_S}{\mu_S(\mathcal{Y})} \qquad j = 1, 2, \ldots, J.$$

Here, $\mu_1, \mu_2, \ldots, \mu_J, \mu_S$ are normalized random measures with independent increments, and $\mathcal{Y}$ represents the sample space. This representation allows f_j to be a mixture of a population-specific component μ_j and a common component μ_S. As a result, two distributions, f_1 and f_2, can share some related atoms (and clusters) while also having distinct clusters specific to each population. The value of γ_j determines the degree of relatedness between populations: when $\gamma_j = 1$ for a specific population j, the populations are independent; when $\gamma_j = 0$ for all populations, the populations are exchangeable. Furthermore, the latent nested process can represent all intermediate situations between independence and full exchangeability. To test equality among populations, one can examine the posterior distribution of $\mathbb{I}[\mu_j = \mu_\ell]$ for $j \neq \ell$. Additionally, the latent nested process induces a partial EPPF, which is a linear combination of the EPPF corresponding to the fully exchangeable case and the EPPF corresponding to unconditional independence.

It is possible to incorporate a dispersion parameter, either scalar or infinite-dimensional, that governs the variability among samples within the same populations and can be assigned a hyperprior. In this case, the completely random measure can be defined as $\frac{\mu_j}{\mu_j(\mathcal{Y})\times\Omega}$, where Ω represents the space over which the dispersion parameter is defined Christensen and Ma (2020).

The main drawback of this model is its computational cost. While the model allows for a latent representation associated with the allocation to each cluster, each step of the corresponding MCMC requires approximating integrals, which can be computationally demanding when using Monte Carlo integration. This slows down the estimation procedure and makes generalizing the model to the case of $J > 2$ populations infeasible in realistic situations.

An alternative approach is to select μ_j and μ_S as independent gamma processes, with μ_j being independent and identically distributed. In this case, $\frac{\mu_j}{\mu_j(\mathcal{Y})}$ and $\frac{\mu_S}{\mu_S(\mathcal{Y})}$ are draws from two independent DPs, and the resulting process is a latent nested DP (Beraha et al., 2021). Another option is to use a Pitman-Yor process instead of a DP, which introduces more flexibility and allows for extending the model in the presence of covariates. For example, μ_j and μ_S can be defined as gamma processes in $\mathcal{Y} \times \mathcal{X}$, where $\mathcal{X}$ represents the covariate space.

Another potential drawback of the latent nested process is that μ_S includes all the common atoms across populations, but these common atoms must have the same weight across populations. This implies that different distributions sharing the same clusters should also have observations allocated to those clusters in the same proportions. This constraint can be limiting in certain scenarios.

To address this limitation, one possible solution is to introduce a weight matrix that relates the weights to a matrix of indicators, where each row represents a population. While this approach is straightforward to implement, it results in a more complex mathematical model and lacks a closed-form expression for the pEPPF.

For further information, refer to Liu and Müller (2019) and Soriano and Ma (2019), which discuss these issues in the context of latent nested processes.

5.3. Hierarchical processes

Unlike the approaches taken by Müller et al. (2004) and Rodriguez et al. (2008), Teh et al. (2005) propose a different model that incorporates information sharing among observations through a common prior. In this model, the mean and covariance of each component are shared across all samples, while the mixture weights remain unique. Each random measure F_j is distributed according to a specific $DP(\alpha_0, F_0)$, where the base measure F_0 has a Dirichlet process prior, denoted as $DP(\gamma, F_{00})$. The clustering in this model arises from the shared clusters among groups of observations.

To elaborate further, the hierarchical Dirichlet process is a nonparametric prior process in which observations Y_{ij}, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, are distributed according to a generic distribution $f(\theta_{ij})$, where θ_{ij} follows the distribution F_j with a Dirichlet process prior. Each F_j is conditionally independent given the base measure F_0, and F_0 itself follows a Dirichlet process prior. The model can be summarized as follows:

$$\begin{aligned} Y_{ij}|\theta_i &\sim f(\theta_i) \qquad i = 1, \ldots, n_j \quad j = 1, \ldots, J \\ \theta_{ij}|F_j &\sim F_j \\ F_j|\alpha, F_0 &\sim DP(\alpha, F_0) \\ F_0|\gamma, F_{00} &\sim DP(\gamma, F_{00}). \end{aligned}$$

This hierarchical model allows individual F_j to share atoms. This sharing is evident from the stick-breaking construction:

$$F_0 = \sum_{h=1}^{\infty} w_h \delta_{\theta_h^*} \qquad \text{and} \qquad F_j = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*}.$$

Thus, the model for each j relies on groups sharing the same mixture atoms θ_h^*, but with different mixing weights $(\pi_1, \pi_2, \ldots)$. The hierarchical construction of the HDP allows the definition of clusters at different levels. Recently, Camerlenghi et al. (2019b) characterize the posterior distribution of this prior process.

The hierarchical DP (HDP) has been widely successful and applied in various fields. It has been used in cytometry (Cron et al., 2013), genomics (Sohn et al., 2009), social networks (Airoldi et al., 2008), imaging (Sivic et al., 2005), health sciences (Gaba and Mittal, 2020), topic models (Gerlach et al., 2018), neuroimaging (Jbabdi et al., 2009; Wang et al., 2011), visual scenes (Kivinen et al., 2007), and many other domains.

The distinction between the HDP and the nested DP and its extensions lies in their clustering properties. The nested DP constructs clusters of individuals across different groups, where the random measures either share the same atoms and weights or have no sharing. In contrast, the HDP allows random measures to share the same atoms but with different weights. As a result, the nested DP enables clustering at both the level of observations and the level of distributions, while the HDP only facilitates clustering at the level of observations.

In order to apply the HDP for population clustering, Beraha et al. (2021) introduce the semi-HDP, which incorporates a baseline distribution as a mixture between a DP and a non-atomic measure. This construction reduces the computational burden compared to Camerlenghi et al. (2019a) for dimensions $J > 2$, and allows for population clustering through a random partition model. The model can be represented as follows:

$$Y_{ij}|F_1, \ldots, F_J, C_1, \ldots, C_J \overset{ind}{\sim} \int_{\Theta} f(\cdot|\theta) F_{c_j}(d\theta) \qquad i = 1, \ldots, n_j, \quad j = 1, \ldots, J$$

$$
\begin{aligned}
C_1, \ldots, C_J &\sim Cat(\pi_1, \ldots, \pi_J) \\
F_1, \ldots, F_J | F_0 &\sim DP(\alpha, F_0) \\
F_0 &= \gamma G_0 + (1-\gamma) G \\
G &\sim DP(\kappa, G_{00}) \\
\gamma &\sim Be(a_\gamma, b_\gamma),
\end{aligned}
$$

where $Cat(\pi_1, \ldots, \pi_J)$ is a categorical distributions, with weights $\pi_1, \ldots, \pi_J$, that means that there are at most J populations; the vector $\pi_1, \ldots, \pi_J$ can be assigned a (Dirichlet) prior distribution. F_j is a discrete random probability measure, i.e. $F_j = \sum_{h=1}^{\infty} w_{jh} \delta_{\theta^*_{jh}}$ where w_{jh} are given a stick-breaking construction and $\theta^*_{jh} \sim F_0$. F_0 is mixture between a DP G and a fixed probability measure G_0. $G = \sum_{h=1}^{\infty} w_h \delta_{\psi_h}$, where w_h are given a stick-breaking construction and $\psi_h \sim G_{00}$; α and κ are two positive concentration parameters.

When $\gamma = 1$, all atoms and weights in F_j are independent and distinct. When $\gamma = 0$, the semi-HDP reduces to the HDP model proposed by Teh et al. (2005), where all F_j share the same atoms, but with different weights. In contrast to Rodriguez et al. (2008), the semi-HDP allows for a positive probability of atom sharing across different F_j's, and atoms can also be shared within the same F_j since all F_j's share the same atoms as G. Moreover, $F_j \neq F_\ell$ with probability one because the weights are different, even when $0 < \gamma < 1$.

Beraha et al. (2021) demonstrate that this prior exhibits full weak support and that the covariance between F_j and F_ℓ, for all j, ℓ in the set $1, 2, \ldots, J$, depends on two parameters: the concentration parameter of the second level, denoted as κ, and the mixing weight of the first level, denoted as γ. As γ approaches 1, F_j and F_ℓ become increasingly uncorrelated. Additionally, Beraha et al. (2021) derive the pEPPF, which is a convex combination of the EPPF corresponding to the fully exchangeable case and the product of the marginal EPPFs for each j.

6. Posterior distributions on the partitions

A model with a prior distribution on the partition ρ_n gives rise to a posterior distribution on the partition. However, estimating the posterior partition in the context of clustering is challenging, since it is typically a high-dimensional problem and it is unlikely that the MCMC algorithm visits a specific partition more than once.

6.1. Stochastic search methods.

Stochastic search methods have been employed to estimate the posterior mode of ρ_n at the end of the MCMC algorithm, as discussed in Brunner and Lo (1999) and Nobile and Fearnside (2007). To approximate the posterior mode of the partition, Dubey et al. (2003), Heller and Ghahramani (2005), and Heard et al. (2006) utilize Bayesian deterministic hierarchical procedures, avoiding the need for MCMC sampling. Another approach, proposed by Medvedovic et al. (2004), involves hierarchical agglomerative clustering, where the distance is based on the posterior similarity matrix estimated through MCMC. While hierarchical clustering methods reduce the involvement of the experimenter in terms of the estimation procedure, they require a method for "cutting the tree", i.e., determining the optimal number of clusters. Fritsch and Ickstadt (2009) propose a method to cut the tree by minimizing the Monte Carlo estimate of a posterior expected loss. This approach is implemented in the `mcclust` package in `R`.

6.2. Decision-theoretic approaches.

6.2.1. Maximum a posteriori.

From a decision-theoretic perspective, it is reasonable to seek the optimal partition that minimizes a posterior expected loss function:

$$\rho_n^* = \arg\min_{\hat{\rho}_n} \mathbb{E}[L(\rho_n, \hat{\rho}_n)|y_1, \ldots, y_n] = \arg\min_{\hat{\rho}_n} \sum_{\rho_n \in \mathcal{P}_n} L(\rho_n, \hat{\rho}_n) p(\rho_n|y_1, \ldots, y_n).$$

where $p(\rho_n|y_1, \ldots, y_n)$ is the posterior distribution of the partition ρ_n. One possible approach is to select the Maximum a Posteriori (MAP) clustering, which corresponds to the optimal Bayesian estimate under the 0-1 loss function $L(\rho_n, \hat{\rho}_n) = \mathbb{I}[\rho_n \neq \hat{\rho}_n]$. However, the estimated MAP obtained from an MCMC output is unlikely to be representative, as a comprehensive exploration of the partition space is infeasible. Additionally, this loss function does not consider any notion of similarity between partitions.

Dahl (2009) proposes an algorithm to derive the MAP partition for a class of univariate product partition models. This algorithm guarantees finding the maximum a posteriori clustering or, at the very least, the maximum likelihood clustering when the partition model can be expressed in terms of a product partition distribution. The algorithm of Dahl (2009) can be applied when two conditions are met: the components in the modal clustering do not overlap (i.e., S_h does not contain integers between the smallest and largest integers in S_k, or vice versa), and the cohesion function $c(S)$ depends only on the number of items contained in S. This algorithm can be highly efficient since it requires only $n(n+1)/2$ evaluations. However, it is restricted to univariate observations and does not provide an estimation error quantification. The algorithm is implemented in the `modalclust` package in `R`.

6.2.2. Methods based on the Binder loss.

Posterior modes can become increasingly unrepresentative of the posterior distribution of the partition as the number of items increases. To address this issue, Lau and Green (2007) propose using the Binder loss (Binder, 1978) instead of the 0-1 loss, aiming to respect the exchangeability in the labeling of clusters and items. The Binder loss penalizes pairs of items that are assigned to different clusters when they should be clustered together, and vice versa. This loss function is commonly used in Bayesian clustering because it can be expressed in terms of the posterior similarity matrix, which is an $n \times n$ matrix where the (i, j)-th element represents the posterior probability that observation i is allocated together with observation j. The posterior similarity matrix can be easily estimated using MCMC.

In terms of allocation variables C_i for $i \in 1, 2, \ldots, n$, the Binder loss is defined as:

$$L(\rho_n, \hat{\rho}_n) = \sum_{(i,j) \in [n]} a\mathbb{I}_{[C_i = C_j, \hat{C}_i \neq \hat{C}_j]} + b\mathbb{I}_{[C_i \neq C_j, \hat{C}_i = \hat{C}_j]}$$

where a and b are non-negative constants that represent the costs of pairwise misclassification. Specifically, a represents the cost of not clustering observations that should be together, while b represents the cost of clustering observations that should not be together. Lau and Green (2007) propose an algorithm where optimizing the expected posterior Binder loss is formulated as a binary integer programming problem. In this formulation, the binary variable $X_{ij} = \mathbb{I}_{[\hat{C}_i = \hat{C}_j]}$ is used, and the objective function (the posterior expected Binder loss) is a linear combination of variables X_{ij} with weights $\Pr(C_i = C_j|y_1, \ldots, y_n)$. The algorithm iteratively targets the

objective function for each item i, reassigning it optimally by either assigning it to an existing cluster or creating a new cluster. The algorithm leverages the fact that minimizing the posterior expectation of the Binder loss is equivalent to minimizing:

$$\sum_{i \leq j} \mathbb{I}[\hat{C}_i = \hat{C}_j] \left(p_{ij} - \frac{b}{a+b} \right)$$

where $p_{ij} = \Pr(C_i = C_j | y_1, \ldots, y_n)$ is the (i,j)-th element of the posterior similarity matrix. However, this algorithm may suffer from scalability issues and is only applied to the Binder loss, without generalisations to other loss functions.

The Binder loss exhibits some asymmetry, preferring to split clusters rather than merge them, while the VI loss is more symmetric. In practice, this asymmetry in the Binder loss may lead to the identification of extra-small clusters in the optimal partition, particularly at the boundary between clusters. This is typical when $a = b$, but it can be mitigated by imposing $a > b$, meaning a higher penalty for allocating observations to different clusters when they should be clustered together.

Dahl (2006) suggests a least squares clustering criterion which seeks the clustering that minimizes:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbb{I}[\hat{C}_i = \hat{C}_j] - \hat{p}_{ij})^2.$$

Minimizing this criterion is equivalent to minimizing the Monte Carlo estimate of the posterior expectation of the Binder loss when $a = b$. However, the method of Dahl (2006) is limited to searching among the partitions visited during the MCMC algorithm.

6.2.3. Methods based on the Variation of Information.

Wade and Ghahramani (2018) propose an approach to summarize the posterior distribution of the clustering structure through both point estimates and credible sets. Unlike Lau and Green (2007), Wade and Ghahramani (2018) propose using the Variation of Information (VI) loss function, as developed by Meilă (2007), and demonstrate through extensive simulation that the Binder loss and the VI loss can yield very different optimal partitions. The VI loss compares the information in two clusterings with the information shared between them:

$$VI(\rho_n, \hat{\rho}_n) = H(\rho_n) + H(\hat{\rho}_n) - 2I(\rho_n, \hat{\rho}_n).$$

where $H(\rho_n)$ and $H(\hat{\rho}_n)$ represent the entropy of each of the two partitions, which measures the uncertainty in cluster allocation. $I(\rho_n, \hat{\rho}_n)$ represents the mutual information between the two partitions. Using the VI loss avoids the choice of a and b as for the Binder loss, however the VI loss does not have a representation in terms of a posterior similarity matrix, making it computationally more expensive to evaluate.

Wade and Ghahramani (2018) propose a greedy search algorithm to explore the partition space, which utilizes an approximation. Minimizing the posterior expectation of the VI loss is equivalent to finding the optimum:

$$\rho_n^* = \arg\min_{\hat{\rho}_n} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{n} \mathbb{I}[\hat{C}_i = \hat{C}_j] \right) - 2 \sum_{i=1}^{n} \mathbb{E} \left[\log \left(\sum_{j=1}^{n} \mathbb{I}[C_i = C_j, \hat{C}_i = \hat{C}_j] \right) | y_1, \ldots, y_n \right].$$

The expectation in the second term can be approximated using an MCMC output. However, evaluating this approximation is computationally costly, as it scales as $O(Tn^2)$, where T is the number of MCMC simulations.

Considering many candidate $\hat{\rho}_n$ can be computationally prohibitive. Therefore, Wade and Ghahramani (2018) propose using Jensen's inequality to swap the logarithm and the expectation, obtaining a lower bound on the expected loss that is more efficient to evaluate, reducing the complexity of the algorithm to $O(n^2)$ for a given $\hat{\rho}_n$. While this approximation reduces computational complexity, its impact on the estimated optimal partition is not clear. Specifically, the properties of the VI loss function found in Meilă (2007) are not guaranteed to hold when applying Jensen's inequality. Another drawback of the algorithm is its dependence on initialization, so it is advisable to run it multiple times starting from different initial partitions. The complexity of this algorithm is $O(\ell n^2)$, where 2ℓ defines the number of partitions to consider at each iteration. Wade and Ghahramani (2018) suggest $\ell = n$. This method is implemented in the `R` package `mcclust.ext`.

6.2.4. Methods based on a generic loss.

Rastelli and Friel (2018) also rely on a decision-theoretic framework to derive the optimal partition. Unlike Wade and Ghahramani (2018), they propose an approach that does not depend on the posterior similarity matrix and does not involve any approximation. This method can be used with any loss function $L(\rho_n, \hat{\rho}n)$ that considers the two partitions through the counts n_{hk}, which denote the number of data points allocated to group h in partition ρ_n and to group k in partition $\hat{\rho}_n$. Since the approach does not require the posterior similarity matrix, its computational complexity in terms of n is reduced to linear order.

The method begins by randomly selecting a partition with small clusters and iteratively reassigning one item at a time to either an existing cluster or a new cluster, depending on the minimal Monte Carlo estimate of the expected loss. Similar to Wade and Ghahramani (2018), the algorithm requires multiple runs to obtain a partition that is closer to the optimal solution. The approach also requires defining a maximum number of clusters K_d, which Rastelli and Friel (2018) suggest setting equal to n, although this choice can increase complexity. The complexity of this algorithm is $O(T \cdot K_d^2 \cdot n)$; if $K_d = n$, the complexity becomes $O(Tn^3)$. This method is implemented in the `R` package `GreedyEPL`, which provides support for various loss functions, including the Binder loss and the VI loss.

6.2.5. Methods based on the generalised VI loss.

Dahl et al. (2022) provide a generalization of the original Variation of Information (VI) loss, similar to the original Binder loss, where weights a and b represent the cost of failing to cluster two observations that should be clustered together and clustering two observations that should not be clustered together, respectively. This generalization maintains the properties of the original VI loss and can be evaluated without incurring higher computational costs. The generalized VI loss with positive weights a and b is given by:

$$L(\rho_n, \hat{\rho}_n) = a \sum_{S \in \rho_n} \frac{|S|}{2} \log_2 \frac{|S|}{2} + b \sum_{S' \in \hat{\rho}_n} \frac{|S'|}{2} \log_2 \frac{|S'|}{2} - (a+b) \sum_{S \in \rho_n} \sum_{S' \in \hat{\rho}_n} \frac{|S \cap S'|}{n} \log_2 \frac{|S \cap S'|}{n}.$$

Here, $S = (S_1, \ldots, S_K)$ and $S' = (S'_1, \ldots, S'_{K'})$ represent two partitions. This loss function can be targeted by a greedy stochastic search algorithm, which can result in the algorithm of Rastelli and Friel (2018) as a special case.

The algorithm begins with an initialization step, either random or sequential. Then, one-at-a-time reallocation of individual observations is performed in a random order. Each observation is removed from its cluster and reallocated to either existing clusters or a new cluster, based on the choice that maximizes the Monte Carlo estimate of the posterior expected loss. This process is repeated until there is no change after a complete run

on all n observations. To avoid getting stuck in a local minimum, occasionally a cluster is "killed" by removing all observations from it and reallocating them sequentially to other clusters. If the Monte Carlo estimate of the expected loss at the end of this reallocation is not lower than the one obtained before the cluster was destroyed, the step is forgotten. This algorithm reduces to that of Rastelli and Friel (2018) when the initialization is not sequential and the number of times clusters are destroyed is set to zero.

The complexity of this algorithm is $O(T \cdot K_d \cdot K_{MCMC} \cdot n)$, where K_{MCMC} is the maximum number of clusters observed among the MCMC samples. This algorithm is implemented in the `R` package `salso`.

6.2.6. Credible balls for the partition.

In addition to estimating the optimal partition, Wade and Ghahramani (2018) also define credible balls:

$$B_{\varepsilon^*}(\rho_n^*) = \{\rho_n : d(\rho_n^*, \rho) \geq \varepsilon^*\}$$

where ε^* is the smallest $\varepsilon > 0$ such that $\Pr(B_\varepsilon(\rho_n^*|y_1, \ldots, y_n)) \geq 1-\alpha$. The bounds of these balls are represented by partitions such that:

$$\Pr(B_{\varepsilon^*}(\rho_n^*)|y_1, \ldots, y_n) = \mathbb{E}[\mathbb{I}[d(\rho_n^*, \rho_n) \leq \varepsilon]|y_1, \ldots, y_n],$$

where the expected value can be estimated based on the partitions visited by the MCMC algorithm with positive probability. However, this definition of credible balls does not guarantee that all partitions inside the credible balls have a higher posterior probability than partitions outside the credible balls. Alternatively, one can consider the highest posterior regions and list all the partitions that have a posterior probability above a certain threshold.

7. Conclusions

Clustering is a fundamental problem in statistics. Model-based clustering offers the advantage of introducing a probabilistic allocation of each observation to possible clusters, as well as a probabilistic definition of the number of clusters. In a Bayesian framework, a popular approach for model-based clustering is to impose a mixture model with an unknown number of clusters, using either a finite or infinite number of components. This leads to a random model on the partition of observations. Various models have been introduced, but the Dirichlet process mixture model and its extensions have been shown to be inconsistent for estimating the number of clusters. Additionally, the prior distribution chosen for the number of components in a finite mixture model strongly affects the estimation of the number of clusters.

An alternative approach is to directly choose a model for the partition, such as through product partition models. While these methods aim directly at modeling clustering, they rely on assumptions about the partitions that may not always hold and can be difficult to verify.

In recent years, there has been attention towards an interesting extension of clustering, which is the problem of clustering populations. In this case, we have reviewed methods based on Gibbs-type priors that extend mixture models.

Finally, an interesting development in recent years is focused on the best ways to summarise the posterior distribution of the partition, and several decision-theoretic approaches have been compared.

This work provides a review of the results and models proposed for Bayesian clustering. Some of the reviewed models were not necessarily introduced for clustering, particularly models that include covariates. However, they are frequently used in applied settings for clustering purposes. This work contributes to the existing literature by comparing the properties and limitations of available models for Bayesian clustering, taking into account

the inconsistency result reported by Miller and Harrison (2014). The proposed comparison aims to highlight the advantages and disadvantages of different methodologies, with the hope of inspiring new avenues for future research.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). "Mixed membership stochastic blockmodels." *Journal of Machine Learning Research*, 9(Sep): 1981–2014.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London, UK, UK: Chapman & Hall, Ltd.

Aldous, D. J. (1985). "Exchangeability and related topics." In *École d'Été de Probabilités de Saint-Flour XIII—1983*, 1–198. Springer.

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2(6): 1152–1174.

Argiento, R. and De Iorio, M. (2022). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." *The Annals of Statistics*, 50(5): 2641–2663.

Argiento, R., Guglielmi, A., and Pievatolo, A. (2009). "A comparison of nonparametric priors in hierarchical mixture modelling for AFT regression." *Journal of Statistical Planning and Inference*, 139(12): 3989–4005.

Barry, D. and Hartigan, J. A. (1992). "Product partition models for change point problems." *The Annals of Statistics*, 260–279.

Beraha, M., Guglielmi, A., and Quintana, F. A. (2021). "The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions." *Bayesian Analysis*, 1(1): 1–33.

Biernacki, C., Celeux, G., and Govaert, G. (2000). "Assessing a mixture model for clustering with the integrated completed likelihood." *IEEE transactions on pattern analysis and machine intelligence*, 22(7): 719–725.

Billio, M., Casarin, R., and Rossini, L. (2019). "Bayesian nonparametric sparse VAR models." *Journal of Econometrics*, 212(1): 97–115.

Binder, D. A. (1978). "Bayesian cluster analysis." *Biometrika*, 65(1): 31–38.

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 353–355.

Blei, D. M. and Frazier, P. I. (2011). "Distance dependent Chinese restaurant processes." *Journal of Machine Learning Research*, 12(8).

Brunner, L. J. and Lo, A. Y. (1999). "Bayesian classifications." *Preprint, University of Toronto, Canada*.

Cai, B., Lawson, A. B., Hossain, M. M., Choi, J., Kirby, R. S., and Liu, J. (2013). "Bayesian semiparametric model with spatially–temporally varying coefficients selection." *Statistics in Medicine*, 32(21): 3670–3685.

Cai, D., Campbell, T., and Broderick, T. (2021). "Finite mixture models do not reliably learn the number of components." In *International Conference on Machine Learning*, 1158–1169. PMLR.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019a). "Latent nested nonparametric priors (with discussion)." *Bayesian Analysis*, 14(4): 1303–1356.

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

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2007). "Bayesian inference for linear dynamic models with Dirichlet process mixtures." *IEEE Transactions on Signal Processing*, 56(1): 71–84.

Caron, F., Teh, Y. W., and Murphy, T. B. (2014). "Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes." *The Annals of Applied Statistics*, 8(2): 1145–1181.

Casella, G., Moreno, E., and Girón, F. J. (2014). "Cluster analysis, model selection, and prior distributions on models." *Bayesian Analysis*, 9(3): 613–658.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). "Deviance information criteria for missing data models." *Bayesian Analysis*, 1(4): 651–673.

Cerquetti, A. (2013). "Marginals of multivariate Gibbs distributions with applications in Bayesian species sampling." *Electronic Journal of Statistics*, 7: 697–716.

Christensen, J. and 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.

Chung, Y. and Dunson, D. B. (2009). "Nonparametric Bayes conditional distribution modeling with variable selection." *Journal of the American Statistical Association*, 104(488): 1646–1660.

Cron, A., Gouttefangeas, C., Frelinger, J., Lin, L., Singh, S. K., Britten, C. M., Welters, M. J., van der Burg, S. H., West, M., and Chan, C. (2013). "Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples." *PLoS Computional Biology*, 9(7): e1003130.

Dahl, D. B. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model." *Bayesian Inference for Gene Expression and Proteomics*, 4: 201–218.

— (2009). "Modal clustering in a class of product partition models." *Bayesian Analysis*, 4(2): 243–264.

Dahl, D. B., Day, R., and Tsai, J. W. (2017). "Random partition distribution indexed by pairwise information." *Journal of the American Statistical Association*, 112(518): 721–732.

Dahl, D. B., Johnson, D. J., and Müller, P. (2022). "Search algorithms and loss functions for Bayesian clustering." *Journal of Computational and Graphical Statistics*, 31(4): 1189–1201.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE transactions on pattern analysis and machine intelligence*, 37(2): 212–229.

Dorazio, R. M. (2009). "On selecting a prior for the precision parameter of Dirichlet process mixture models." *Journal of Statistical Planning and Inference*, 139(9): 3384–3390.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). "Generalized spatial Dirichlet process models." *Biometrika*, 94(4): 809–825.

Dubey, A., Hwang, S., Rangel, C., Rasmussen, C. E., Ghahramani, Z., and Wild, D. L. (2003). "Clustering protein sequence and structure space with infinite Gaussian mixture models." In *Biocomputing 2004*, 399–410. World Scientific.

Dunson, D. B. (2006). "Bayesian dynamic modeling of latent trait distributions." *Biostatistics*, 7(4): 551–568.

Dunson, D. B. and Park, J.-H. (2008). "Kernel stick-breaking processes." *Biometrika*, 95(2): 307–323.

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 West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230.

Fernández, C. and Green, P. J. (2002). "Modelling spatially correlated data via mixtures: a Bayesian approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 805–826.

Fritsch, A. and Ickstadt, K. (2009). "Improved criteria for clustering based on the posterior similarity matrix." *Bayesian Analysis*, 4(2): 367–391.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2019). *Handbook of mixture analysis*. CRC press.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Generalized mixtures of finite mixtures and telescoping sampling." *Bayesian Analysis*, 16(4): 1279–1307.

Fuentes, M. and Reich, B. (2013). "Multivariate spatial nonparametric modelling via kernel processes mixing." *Statistica Sinica*, 23(1).

Fuentes-García, R., Mena, R. H., and Walker, S. G. (2009). "A nonparametric dependent process for Bayesian regression." *Statistics & Probability Letters*, 79(8): 1112–1119.

Fúquene, J., Steel, M., and Rossell, D. (2019). "On choosing mixture components via non-local priors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5): 809–837.

Gaba, D. and Mittal, N. (2020). "Implementation and classification of machine learning algorithms in healthcare informatics: approaches, challenges, and future scope." *Computational Intelligence for Machine Learning and Healthcare Informatics*, 1: 21.

Gelfand, A., Guindani, M., and Petrone, S. (2007). "Bayesian nonparametric modelling for spatial data using Dirichlet processes." In *Bayesian statistics 8: proceedings of the eighth Valencia International Meeting, June 2-6, 2006*, volume 8, 175. Oxford University Press, USA.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100(471): 1021–1035.

Gerlach, M., Peixoto, T. P., and Altmann, E. G. (2018). "A network approach to topic models." *Science Advances*, 4(7): eaaq1360.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *The Annals of Statistics*, 27(1): 143–158.

Ghosal, S. and Van Der Vaart, A. (2007). "Convergence rates of posterior distributions for noniid observations." *The Annals of Statistics*, 35(1): 192–223.

Ghosal, S. and Van Der Vaart, A. W. (2001). "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities." *The Annals of Statistics*, 29(5): 1233–1263.

Ghosh, S., Ungureanu, A., Sudderth, E., and Blei, D. (2011). "Spatial distance dependent Chinese restaurant processes for image segmentation." *Advances in Neural Information Processing Systems*, 24: 1476–1484.

Gnedin, A. (2010). "A species sampling model with finitely many types." *Electronic Communications in Probability*, 15: 79–88.

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences*, 138(3): 5674–5685.

Goldwater, S., Johnson, M., and Griffiths, T. L. (2006). "Interpolating between types and tokens by estimating power-law generators." In *Advances in Neural Information Processing Systems*, 459–466.

Grazian, C. (2023). "Spatio-temporal stick-breaking process." *arXiv preprint arXiv:2303.17177*.

Grazian, C. and Robert, C. P. (2018). "Jeffreys priors for mixture estimation: Properties and alternatives." *Computational Statistics & Data Analysis*, 121: 149–163.

Grazian, C., Villa, C., and Liseo, B. (2020). "On a loss-based prior for the number of components in mixture models." *Statistics & Probability Letters*, 158: 108656.

Green, P. J. and Richardson, S. (2001). "Modelling heterogeneity with and without the Dirichlet process." *Scandinavian Journal of Statistics*, 28(2): 355–375.

Greve, J., Grün, B., Malsiner-Walli, G., Frühwirth-Schnatter, S., and Greve, M. J. (2021). "Package 'fipp'."

Griffin, J. E. and Steel, M. J. (2006). "Order-based dependent Dirichlet processes." *Journal of the American Statistical Association*, 101(473): 179–194.

Gutiérrez, L., Barrientos, A. F., González, J., and Taylor-Rodríguez, D. (2019). "A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control." *Bayesian Analysis*, 14(2): 649–675.

Hartigan, J. A. (1990). "Partition models." *Communications in Statistics-Theory and methods*, 19(8): 2745–2756.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). "A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves." *Journal of the American Statistical Association*, 101(473): 18–29.

Heller, K. A. and Ghahramani, Z. (2005). "Bayesian hierarchical clustering." In *Proceedings of the 22nd international conference on Machine learning*, 297–304.

Ho, M.-W., James, L. F., and Lau, J. W. (2007). "Gibbs partitions (EPPF's) derived from a stable subordinator are Fox H and Meijer G transforms." *arXiv preprint arXiv:0708.0619*.

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

Jain, S. and Neal, R. M. (2004). "A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *Journal of Computational and Graphical Statistics*, 13(1): 158–182.

Jara, A., Garcia-Zattera, M. J., and Lesaffre, E. (2007). "A Dirichlet process mixture model for the analysis of correlated binary responses." *Computational Statistics & Data Analysis*, 51(11): 5402–5415.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." *Statistical Science*, 20(1): 50–67.

Jbabdi, S., Woolrich, M. W., and Behrens, T. E. J. (2009). "Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models." *NeuroImage*, 44(2): 373–384.

Jensen, S. T. and Liu, J. S. (2008). "Bayesian clustering of transcription factor binding motifs." *Journal of the American Statistical Association*, 103(481): 188–200.

Kivinen, J. J., Sudderth, E. B., and Jordan, M. I. (2007). "Learning multiscale representations of natural scenes using Dirichlet processes." In *2007 IEEE 11th International Conference on Computer Vision*, 1–8.

Kolossiatis, M., Griffin, J. E., and Steel, M. F. (2013). "On Bayesian nonparametric modelling of two correlated distributions." *Statistics and Computing*, 23(1): 1–15.

Kruijer, W., Rousseau, J., and Van Der Vaart, A. (2010). "Adaptive Bayesian density estimation with location-scale mixtures." *Electronic Journal of Statistics*, 4: 1225–1257.

Lau, J. W. and Green, P. J. (2007). "Bayesian model-based clustering procedures." *Journal of Computational and Graphical Statistics*, 16(3): 526–558.

Lee, C. J. and Sang, H. (2022). "Why the Rich Get Richer? On the Balancedness of Random Partition Models." In *International Conference on Machine Learning*, 12521–12541. PMLR.

Lee, J., Quintana, F. A., Müller, P., and Trippa, L. (2013). "Defining predictive probability functions for species sampling models." *Statistical Science*, 28(2): 209.

Li, P., Banerjee, S., Hanson, T. E., and McBean, A. M. (2010). "Nonparametric hierarchical modeling for detecting boundaries in areally referenced spatial datasets." Technical report.

Lijoi, A., Nipoti, B., and Prünster, I. (2014). "Bayesian inference with dependent normalized completely random measures." *Bernoulli*, 20(3): 1260–1291.

Lijoi, A., Prünster, I., and Walker, S. G. (2005). "On consistency of nonparametric normal mixtures for Bayesian density estimation." *Journal of the American Statistical Association*, 100(472): 1292–1296.

— (2008). "Bayesian nonparametric estimators derived from conditional Gibbs structures." *The Annals of Applied Probability*, 18(4): 1519–1547.

Liu, V. and Müller, P. (2019). "Invited discussion on"Latent nested nonparametric priors"by Camerlenghi F., Dunson DB, Lijoi A., Prunster I. and Rodriguez A., Bayesian Analysis."

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics-Simulation and Computation*, 23(3): 727–741.

— (1998). "Computational methods for mixture of Dirichlet process models." In *Practical nonparametric and semiparametric Bayesian statistics*, 23–43. Springer.

— (2000). "Dependent dirichlet processes." *Unpublished manuscript, Department of Statistics, The Ohio State University*, 1–40.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). "Identifying mixtures of mixtures using Bayesian estimation." *Journal of Computational and Graphical Statistics*, 26(2): 285–295.

Mastrantonio, G., Grazian, C., Mancinelli, S., and Bibbona, E. (2019). "New formulation of the logistic-Gaussian process to analyze trajectory tracking data." *The Annals of Applied Statistics*, 13(4): 2483–2508.

McCullagh, P. and Yang, J. (2008). "How many clusters?" *Bayesian Analysis*, 3(1): 101–120.

Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004). "Bayesian mixture model based clustering of replicated microarray data." *Bioinformatics*, 20(8): 1222–1232.

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

Miller, J. W. and Harrison, M. T. (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components." *The Journal of Machine Learning Research*, 15(1): 3333–3370.

— (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113(521): 340–356.

Müller, P., Quintana, F., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of Royal Statistical Society, Series B*, 66: 735–749.

Müller, P., Quintana, F., and Rosner, G. L. (2011). "A product partition model with regression on covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–278.

Neelon, B., Gelfand, A. E., and Miranda, M. L. (2014). "A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(5): 737–761.

Nguyen, X. (2013). "Convergence of latent mixing measures in finite and infinite mixture models." *The Annals of Statistics*, 41(1): 370–400.

Nobile, A. (1994). *Bayesian analysis of finite mixture distributions*. Carnegie Mellon University.

— (2004). "On the posterior distribution of the number of components in a finite mixture." *The Annals of Statistics*, 32(5): 2044–2073.

Nobile, A. and Fearnside, A. T. (2007). "Bayesian finite mixtures with an unknown number of components: The allocation sampler." *Statistics and Computing*, 17(2): 147–162.

Ogata, Y. and Tanemura, M. (1985). "Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method." *Biometrics*, 421–433.

Ongaro, A. and Cattaneo, C. (2004). "Discrete random probability measures: a general framework for nonparametric Bayesian inference." *Statistics & Probability Letters*, 67(1): 33–45.

Onogi, A., Nurimoto, M., and Morita, M. (2011). "Characterization of a Bayesian genetic clustering algorithm based on a Dirichlet process prior and comparison among Bayesian clustering methods." *BMC Bioinformatics*, 12(1): 1–16.

Paci, L. and Finazzi, F. (2018). "Dynamic model-based clustering for spatio-temporal data." *Statistics and Computing*, 28(2): 359–374.

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2021). "Centered partition processes: Informative priors for clustering (with discussion)." *Bayesian Analysis*, 16(1): 301–370.

Page, G. L. and Quintana, F. A. (2016). "Spatial product partition models." *Bayesian Analysis*, 11(1): 265–298.

Papageorgiou, G., Richardson, S., and Best, N. (2015). "Bayesian non-parametric models for spatially indexed data of mixed type." *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 973–999.

Park, J.-H. and Dunson, D. B. (2010). "Bayesian generalized product partition model." *Statistica Sinica*, 1203–1226.

Pawlowsky, V. and Burger, H. (1992). "Spatial structure analysis of regionalized compositions." *Mathematical Geology*, 24(6): 675–691.

Petralia, F., Rao, V., and Dunson, D. (2012). "Repulsive mixtures." *Advances in Neural Information Processing Systems*, 25.

Pitman, J. (1995). "Exchangeable and partially exchangeable random partitions." *Probability Theory and Related Fields*, 102(2): 145–158.

— (2002). "Combinatorial stochastic processes." *Technical Report*.

Quinlan, J. J., Page, G. L., and Quintana, F. A. (2018). "Density regression using repulsive distributions." *Journal of Statistical Computation and Simulation*, 88(15): 2931–2947.

Quinlan, J. J., Quintana, F. A., and Page, G. L. (2021). "On a class of repulsive mixture models." *TEST*, 30(2): 445–461.

Quintana, F. A. and Iglesias, P. L. (2003). "Bayesian clustering and product partition models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 557–574.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). "The dependent Dirichlet process and related models." *Statistical Science*, 37(1): 24–41.

Rastelli, R. and Friel, N. (2018). "Optimal Bayesian estimators for latent variable cluster models." *Statistics and Computing*, 28(6): 1169–1186.

Reich, B. J. and Bondell, H. D. (2011). "A spatial Dirichlet process mixture model for clustering population genetics data." *Biometrics*, 67(2): 381–390.

Reich, B. J. and Fuentes, M. (2007). "A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields." *The Annals of Applied Statistics*, 1(1): 249–264.

Reich, B. J., Kalendra, E., Storlie, C. B., Bondell, H. D., and Fuentes, M. (2012). "Variable selection for high dimensional Bayesian density estimation: application to human exposure simulation." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(1): 47–66.

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011). "Logistic stick-breaking process." *Journal of Machine Learning Research*, 12(1).

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4): 731–792.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1154.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2010). "Latent stick-breaking processes." *Journal of the American Statistical Association*, 105(490): 647–659.

Rodríguez, C. E. and Walker, S. G. (2014). "Univariate Bayesian nonparametric mixture modeling with unimodal kernels." *Statistics and Computing*, 24(1): 35–49.

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5): 689–710.

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

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. (2005). "Discovering objects and their location in images." In *Tenth IEEE International Conference on Computer Vision Volume 1*, volume 1, 370–377. IEEE.

Sohn, K.-A., Xing, E. P., et al. (2009). "A hierarchical Dirichlet process mixture model for haplotype reconstruction from multi-population data." *The Annals of Applied Statistics*, 3(2): 791–821.

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

Stephens, M. (2000). "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods." *The Annals of Statistics*, 40–74.

Sudderth, E. B. and Jordan, M. I. (2008). "Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes." 1585–1592.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). "Sharing clusters among related groups: Hierarchical Dirichlet processes." In *Advances in Neural Information Processing Systems*, 1385–1392.

Tjelmeland, H. and Lund, K. V. (2003). "Bayesian modelling of spatial compositional data." *Journal of Applied Statistics*, 30(1): 87–100.

Tokdar, S. T. (2006). "Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression." *Sankhyā: The Indian Journal of Statistics*, 90–110.

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

Walker, S. G., Lijoi, A., and Prünster, I. (2007). "On rates of convergence for posterior distributions in infinite-dimensional models." *The Annals of Statistics*, 35(2): 738–746.

Wall, M. M. (2004). "A close look at the spatial structure implied by the CAR and SAR models." *Journal of Statistical Planning and Inference*, 121(2): 311–324.

Wang, X., Grimson, W. E. L., and Westin, C.-F. (2011). "Tractography segmentation using a hierarchical Dirichlet processes mixture model." *NeuroImage*, 54(1): 290–302.

Warren, J., Fuentes, M., Herring, A., and Langlois, P. (2012). "Bayesian spatial–temporal model for cardiac congenital anomalies and ambient air pollution risk assessment." *Environmetrics*, 23(8): 673–684.

Woo, M.-J. and Sriram, T. (2006). "Robust estimation of mixture complexity." *Journal of the American Statistical Association*, 101(476): 1475–1486.

— (2007). "Robust estimation of mixture complexity for count data." *Computational Statistics & Data Analysis*, 51(9): 4379–4392.

Wu, Y. and Ghosal, S. (2010). "The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation." *Journal of Multivariate Analysis*, 101(10): 2411–2419.

Xie, F. and Xu, Y. (2020). "Bayesian repulsive gaussian mixture model." *Journal of the American Statistical Association*, 115(529): 187–203.