ISSN 2279-9362

Bayesian modeling via discrete nonparametric priors

Marta Catalano, Antonio Lijoi, Igor Prunster and Tommaso Rigon

No. 696
July 2023

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Bayesian modeling via discrete nonparametric priors

Marta Catalano [1,4] Antonio Lijoi [2,4], Igor Prünster [2,4] and Tommaso Rigon[3,4]

[1] Department of Statistics, University of Warwick, Coventry, UK

[2] Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University, Milano, Italy

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

[4] Collegio Carlo Alberto, Piazza Arbarello 8, Torino, Italy

Abstract

The availability of complex-structured data has sparked new research directions in Statistics and Machine Learning. Bayesian Nonparametrics is at the forefront of this trend thanks to two crucial features: its coherent probabilistic framework, which naturally leads to principled prediction and uncertainty quantification, and its infinite-dimensionality, which exempts from parametric restrictions and ensures full modeling flexibility. In this paper we provide a concise overview of Bayesian Nonparametrics starting from its foundations and the Dirichlet process, the most popular nonparametric prior. We describe the use of the Dirichlet process in species discovery, density estimation, and clustering problems. Among the many generalizations of the Dirichlet process proposed in the literature we single out the Pitman-Yor process, and compare it to the Dirichlet process. Their different features are showcased with real-data illustrations. Finally, we consider more complex data structures, which require dependent versions of these models. One of the most effective strategies to achieve this goal is represented by hierarchical constructions. We highlight the role of the dependence structure in the borrowing of information and illustrate its effectiveness on unbalanced datasets.

Keywords: Clustering, Density estimation, Dependence, Dirichlet process, Exchangeability, Mixture model, Partial Exchangeability, Pitman-Yor process, Species discovery.

1 Introduction

Statistical learning and predictions are largely based on the tacit assumption of a correspondence between past and future observations. Whether there is a rational justification for this

induction principle has been a long debated question, which can be traced back to the work of the renowned philosopher David Hume in the 18th century. In the 20th century Bruno de Finetti (de Finetti, 1937) reformulates the problem in probabilistic terms: the symmetry between past and future can be postulated with *an appropriate probabilistic framework* through the concept of exchangeability, which requires the distribution of the observations to be invariant with respect to their order. More formally, a sequence $(X_n)_{n\geq 1}$ of observations on a Polish space $\mathbb{X}$ is said to be exchangeable if for every $N \in \mathbb{N}$, and for every permutation π of $\{1, \ldots, N\}$,

$$(X_1, \ldots, X_N) \stackrel{\mathrm{d}}{=} (X_{\pi(1)}, \ldots, X_{\pi(N)}),$$

where $\stackrel{\mathrm{d}}{=}$ denotes the equality in distribution. Crucially, de Finetti proved that exchangeability is equivalent to conditional independence and identity in distribution, through the following representation theorem which goes under his name.

Theorem 1 (de Finetti Representation Theorem)**.** *A sequence $(X_n)_{n\geq 1}$ of random elements on $\mathbb{X}$ is exchangeable if and only if there exists a probability law Q on the space $\mathscr{P}_{\mathbb{X}}$ of probabilities on $\mathbb{X}$ such that, for every $N \in \mathbb{N}$ and any Borel sets $(A_1, \ldots, A_N)$,*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_N \in A_N) = \int_{\mathscr{P}_{\mathbb{X}}} \prod_{i=1}^{N} P(A_i)\, Q(dP).$$

Q is termed the de Finetti measure of $(X_n)_{n\geq 1}$.

The representation theorem was proved in full generality by Hewitt and Savage (Hewitt and Savage, 1955), with compelling consequences in terms of both inference and prediction. First of all, it provides a neat justification of the Bayes-Laplace paradigm, that is, the use of a prior distribution on the model parameters for Bayesian inference. Indeed, one can rewrite the representation theorem in hierarchical form, so that for any exchangeable sequence $(X_n)_{n\geq 1}$ one can define a random probability $\tilde{P}$ on $\mathscr{P}_{\mathbb{X}}$ such that, for every $n \in \mathbb{N}$,

$$\begin{aligned} X_i \mid \tilde{P} &\stackrel{\text{iid}}{\sim} \tilde{P} \qquad i = 1, \ldots, n; \\ \tilde{P} &\sim Q. \end{aligned}$$

From this expression it is apparent that the de Finetti measure Q may act as prior distribution, the latent parameter being the distribution of the data. After observing the values $X^{(n)} = (X_1, \ldots, X_n)$, one can perform Bayesian inference by finding the posterior distribution $Q(\cdot \mid X^{(n)})$.

Moreover, de Finetti's theorem can also be used as a technical tool for assigning explicit laws to exchangeable sequences and deriving predictive inference, *the most important and fundamental kind of inductive inference* according to Carnap (1950). Indeed, under the exchangeability assumption, the one-step predictions are conveniently evaluated as linear functions of the posterior distribution $Q(\cdot \mid X^{(n)})$ as follows

$$\mathbb{P}(X_{n+1} \in A \mid X^{(n)}) = \int_{\mathbb{P}_{\mathbb{X}}} P(A)\, Q(\mathrm{d}P \mid X^{(n)}),$$

and similarly for m-step ahead predictions $\mathbb{P}((X_{n+1}, \ldots, X_{n+m}) \in A_m \mid X^{(n)})$.

Summing up, exchangeability guarantees an elegant and principled framework to perform inference and prediction on homogeneous observations. As for the flexibility of these models, it crucially depends on the support of the de Finetti measure Q on the space of probabilities $\mathscr{P}_{\mathbb{X}}$. Though de Finetti had laid out the Bayesian framework in its full generality during the 1930s, for several decades inference and predictions were confined to parametric models, that is, for priors Q that degenerate on a subclass of $\mathscr{P}_{\mathbb{X}}$ indexed by a finite-dimensional parameter. Still in 1972 Dennis V. Lindley wrote that *it is perhaps worth stopping to remark that the problem is a technical one; the Bayesian method embraces non-parametric problems but cannot solve them because the requisite tool is missing* (Lindley, 1972). However, the times were ripe: in 1973 Thomas S. Ferguson (Ferguson, 1973) made the breakthrough with the definition of the Dirichlet process prior, which paved the way for the development of the field of Bayesian Nonparametrics (BNP). See also Ferguson (1974). The rest of this paper will outline some of the most notable uses of the Dirichlet process and to highlight some effective generalizations. We start by introducing the Dirichlet process through its stick-breaking representation (Sethuraman, 1994), arguably its simplest construction though perhaps not most suitable to gain insight into its distributional properties.

Definition 1. *A random probability $\tilde{P} \sim \mathrm{DP}(\theta, P^*)$ is distributed according to a Dirichlet process prior with concentration $\theta > 0$ and base probability $P^* \in \mathscr{P}_{\mathbb{X}}$ if*

$$\tilde{P} \overset{\mathrm{d}}{=} \sum_{i \geq 1} \tilde{p}_i \delta_{Z_i},$$

where $Z_i \overset{\text{iid}}{\sim} P^*$ are independent of $\omega_i \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \theta)$, and $\tilde{p}_i = \omega_i \prod_{j=1}^{i-1}(1 - \omega_j)$.

For simplicity, we will throughout assume that P^* is non-atomic. Exhaustive accounts of the field can be found in the monographs Hjort et al. (2010), Müller et al. (2015) and Ghosal and van der Vaart (2017).

2 Species discovery

In species sampling problems, a population of individuals is partitioned into different "types" or "species", and each observation is a species label. Given an observed sample of size n, one of the main goals is predicting the number of new distinct species in a future sample of size m. There are two key elements of species sampling problems that make BNP models the natural choice: (i) since the scope of the notion of species is to group individual observations (according to some criterion of similarity), there should be a positive probability of ties among the data; (ii) since each new observation could potentially represent a new species, the model has to assign a positive probability to this event or, in other terms, incorporate a positive discovery probability at each sampling step. This BNP approach to species sampling was first laid out in Lijoi et al. (2007).

Let $X^{(n)} = (X_1, \ldots, X_n)$ be a vector of observed species labels. If one believes that the order with which the species are observed is irrelevant, assuming exchangeability is the natural choice leading to

$$X_1, \ldots, X_n \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}. \tag{1}$$

In this modeling framework it is quite natural to assume that $\tilde{P}$ is discrete and defined as $\tilde{P} = \sum_{i \geq 1} \tilde{p}_i \delta_{Z_i}$, where the $\tilde{p}_j$'s are the random species proportions in the population and the Z_j's are the corresponding species labels such that $Z_j \overset{\text{iid}}{\sim} P^*$ for some non–atomic base measure P^*. These assumptions imply that ties may be recorded in $X^{(n)}$ with positive probability, namely $\mathbb{P}(X_i = X_j) > 0$ for any $i \neq j$. Henceforth, we will denote by K_n the total number of distinct species in the sample $X^{(n)}$, by $X_1^*, \ldots, X_{K_n}^*$ the unique values for the species, and by N_i the frequency of the i-th distinct species in order of appearance.

2.1 Dirichlet process

The task of learning the next species may be rephrased into finding the predictive distribution $\mathbb{P}(X_{n+1} \mid X^{(n)})$. The species corresponding to the next observation X_{n+1} should have a positive probability of coinciding with the already discovered species but also a positive probability of being new. A natural probabilistic structure achieving these *desiderata* is obtained by taking a linear combination of one's prior guess at the (marginal) distribution of the species labels P^* and the empirical measure. This results in a learning mechanism of the

form

$$\mathbb{P}(X_{n+1} \in \cdot \mid X^{(n)}) = \mathbb{P}(X_{n+1} = \text{“new”} \mid X^{(n)})\ P^*(\cdot) + \\ + \mathbb{P}(X_{n+1} = \text{“old”} \mid X^{(n)})\ \frac{1}{n}\sum_{i=1}^{K_n} N_i \delta_{X_i^*}(\cdot). \quad (2)$$

Now, if $\tilde{P}$ is distributed according to a Dirichlet process prior (Definition 1), the corresponding predictive distributions replicate exactly (2) with

$$\mathbb{P}(X_{n+1} = \text{“new”} \mid X^{(n)}) = \frac{\theta}{\theta + n}; \quad \mathbb{P}(X_{n+1} = \text{“old”} \mid X^{(n)}) = \frac{n}{\theta + n}. \quad (3)$$

Moreover, it can be proved (Regazzini, 1978; Lo, 1991) that the predictive distributions associated to an exchangeable sequence are a linear combination of P^* and the empirical measure $\frac{1}{n}\sum_{i=1}^{K_n} N_i \delta_{X_i^*}$ if and only if the underlying $\tilde{P}$ has a Dirichlet process distribution.

This predictive scheme, framed as a generative model, is known as the Blackwell-MacQueen generalized Pólya urn (Blackwell and MacQueen, 1973). Moreover, when focusing on the induced partition distribution, it reduces to the *Chinese restaurant process*, whose name originated from the following metaphor, attributed to L. Dubins and J. Pitman by D. Aldous. A sequence of customers arrives at a restaurant with X_i denoting the table of customer i. The first customer sits at one of the empty tables. The second customer sits at the same table ($X_2 = X_1$) with probability $1/(\theta+1)$ and at a different table ($X_2 =$ “new”) with probability $1/(1+\theta)$, and so on. Thus, tables and species may be treated in the same way: the key feature they share is that they both determine a random partition of observations into clusters.

Let $\Pi_k^n(n_1, \ldots, n_k)$ be the probability that a given sample $X^{(n)}$ displays $k \leq n$ species with cardinalities $n_1, \ldots, n_k$. Whenever $\tilde{P}$ is an almost surely discrete distribution this can be characterized in terms of the de Finetti measure as

$$\Pi_k^n(n_1, \ldots, n_k) = \mathbb{E}\bigg(\int_{\mathbb{X}_*^k} \prod_{i=1}^{k} \tilde{P}^{n_i}(dx_i) \bigg), \quad (4)$$

where $\mathbb{X}_*^k = \{(x_1, \ldots, x_k) \in \mathbb{X}^k : x_i \neq x_j \text{ for } i \neq j\}$, and it is referred to as the exchangeable partition probability function, a notion introduced by J. Pitman (Pitman, 1995). If $\tilde{P}$ is sampled from a Dirichlet process this coincides with a variation of the popular Ewens sampling formula (Antoniak, 1974; Ewens, 1972), which plays a major role in Population Genetics and is given by

$$\Pi_k^n(n_1, \ldots, n_k) = \frac{\theta^k}{(\theta)_n} \prod_{i=1}^{k} (n_i - 1)!,$$

where $(\theta)_n = \theta(\theta+1)\cdots(\theta+n-1)$ denotes the n–th ascending factorial of θ. By summing (4) over all partitions of n elements into k groups, one obtains the probability of observing k distinct species in a sample of size n. Indeed,

$$\mathbb{P}(K_n = k) = \frac{\theta^k}{(\theta)_n} |s(n,k)|,$$

with $|s(n,k)|$ the signless Stirling number of the first kind. From a modeling perspective, an important aspect is the growth rate of the number of distinct species K_n as n increases: under a Dirichlet process prior, K_n diverges with a logarithmic behavior. More specifically, as shown in Korwar and Hollander (1973), for $n \to \infty$,

$$\frac{K_n}{\log n} \xrightarrow{\text{a.s.}} \theta.$$

We refer to Pitman (2006), Mano (2018) and Yamato (2020) for further stimulating accounts.

2.2 Beyond the Dirichlet process

When modeling the probability of discovering a new species, one would expect this probability to depend explicitly on the number K_n of distinct species in the sample. More specifically, it is often desirable for the probability of discovering a new species to be monotonically increasing in K_n, so that the probability of discovering a new species is higher if mostly distinct species have been recorded in the past, and viceversa. The Dirichlet process prior does not accommodate for this feature, as it is apparent from (3). However, other choices of the de Finetti measure allow for this modeling behavior: a popular instance is given by the two–parameter Poisson–Dirichlet process (Perman et al., 1992; Pitman, 1995; Pitman and Yor, 1997), also known as Pitman-Yor process. See Lijoi and Prünster (2010) and Müller et al. (2018) for reviews of the various classes of discrete random probability measures that share this property.

Definition 2. *A random probability $\tilde{P} \sim \text{PY}(\sigma, \theta, P^*)$ is distributed according to a Pitman-Yor process prior with discount $\sigma \in [0,1)$, concentration $\theta > -\sigma$, and diffuse base probability $P^* \in \mathscr{P}_{\mathbb{X}}$ if*

$$\tilde{P} \overset{\text{d}}{=} \sum_{i \geq 1} \tilde{p}_i \delta_{Z_i},$$

where $Z_i \overset{\text{iid}}{\sim} P^*$ are independent of $\omega_i \overset{\text{ind}}{\sim} \text{Beta}(1-\sigma, \theta+i\sigma)$, and $\tilde{p}_i = \omega_i \prod_{j=1}^{i-1}(1-\omega_j)$.

Clearly, the Dirichlet process represents a special case, corresponding to $\sigma = 0$. When $\tilde{P} \sim \text{PY}(\theta, \sigma, P^*)$ is sampled from a Pitman-Yor process, the exchangeable partition probability function becomes

$$\Pi_k^n(n_1, \ldots, n_k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{i=1}^{k} (1-\sigma)_{n_i - 1}.$$

Moreover, the predictive distribution of a Pitman-Yor process satisfies

$$\mathbb{P}(X_{n+1} = \text{“new”} \mid X^{(n)}) = \frac{\theta + \sigma k}{\theta + n},$$

where k is the number of distinct species in the sample $X^{(n)}$. The complete prediction scheme is of the form

$$\mathbb{P}(X_{n+1} \in \cdot \mid X^{(n)}) = \frac{\theta + \sigma k}{\theta + n} \; P^*(\cdot) + \frac{n}{\theta + n} \; \frac{1}{n} \sum_{i=1}^{k} (N_i - \sigma) \delta_{X_i^*}(\cdot),$$

which notably features a suitably weighted empirical measure. Depending on the value of σ, this leads to a markedly different learning scheme, as highlighted in Figure 1. Here, we compare the prediction of the number of new unique values $K_m^{(n)}$ in an additional sample of size m, conditionally on $X^{(n)}$. In particular, we consider the Bayesian nonparametric estimator for $K_m^{(n)}$ derived in Favaro et al. (2009), that is

$$\mathbb{E}(K_m^{(n)} \mid X^{(n)}) = \left(k + \frac{\theta}{\sigma}\right) \left\{ \frac{(\theta + n + \sigma)_m}{(\theta + n)_m} - 1 \right\}. \tag{5}$$

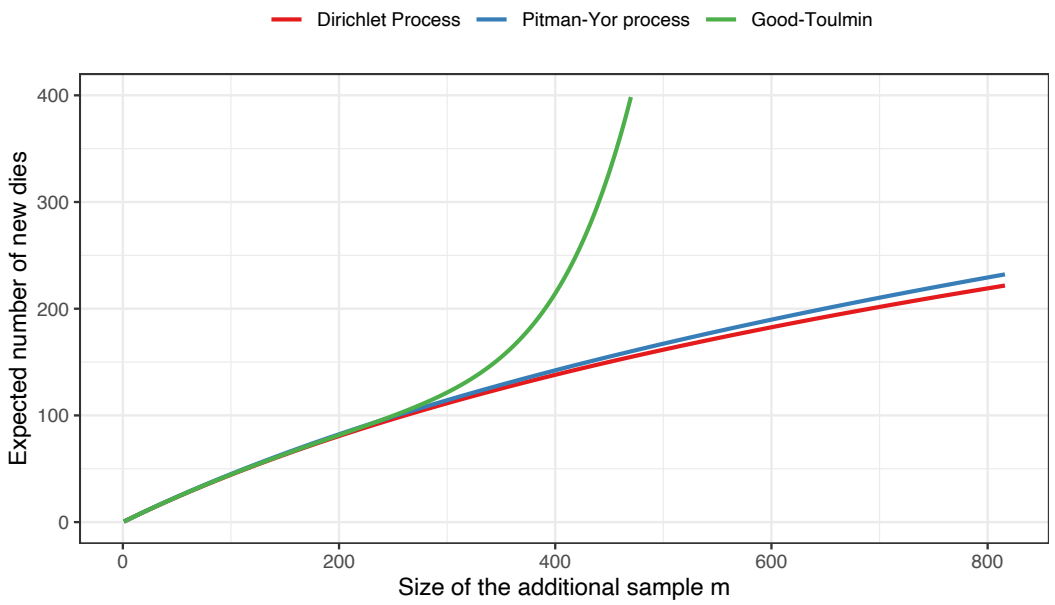
We refer to Lijoi et al. (2007) and Favaro et al. (2009) details. In the Dirichlet process case, as shown in Favaro et al. (2011), (5) reduces to

$$\mathbb{E}(K_m^{(n)} \mid X^{(n)}) = \sum_{i=1}^{m} \frac{\theta}{\theta + n + i - 1}.$$

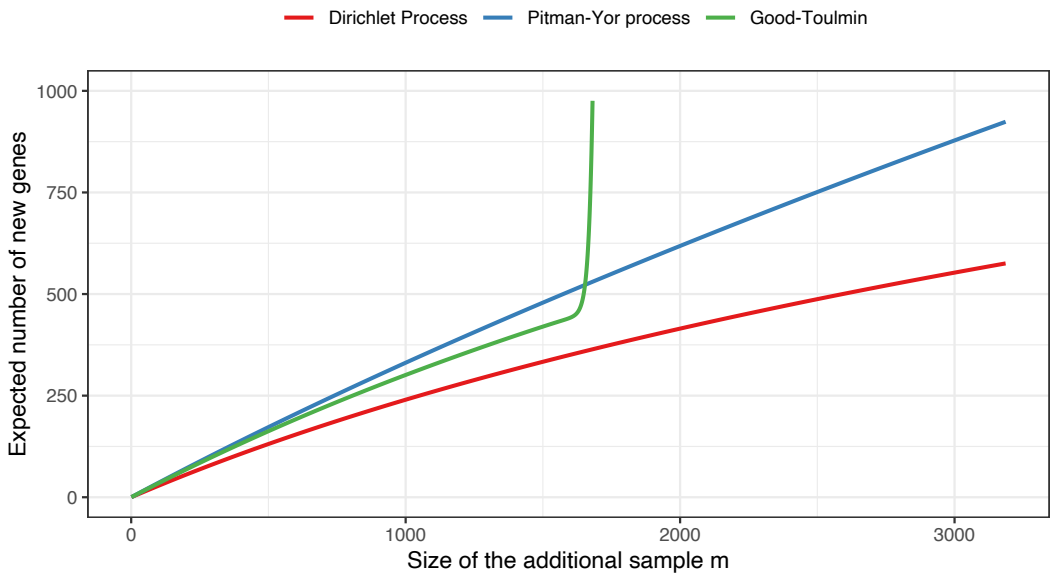
A practical issue to face is the specification of the parameters (σ, θ). A first convenient approach proposed in Lijoi et al. (2007) is based on empirical Bayes ideas. It consists in fixing (σ, θ) to maximize the exchangeable partition probability function, so that in the Pitman-Yor case we have

$$(\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma,\theta)} \left\{ \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \prod_{i=1}^{k} (1-\sigma)_{n_i - 1} \right\}.$$

An alternative way of specifying (σ, θ) is by placing a prior on it. However, as showcased in Lijoi et al. (2008), this often results in negligible differences for the estimates of $K_m^{(n)}$ compared to the empirical Bayes approach, since in many practical scenarios the posterior distribution for (σ, θ) is highly concentrated.



(a) Coins dataset: $n = 204$ coins made with $k = 141$ distinct dies.



(b) Citrus clementina dataset: $n = 1593$ expressed sequence tags corresponding to $k = 806$ distinct genes.

Figure 1: Species discovery: Bayesian nonparametric estimators $\mathbb{E}(K_m^{(n)} \mid X^{(n)})$ of the number of new distinct values $K_m^{(n)}$ in an additional sample of size m, conditionally on the data $X^{(n)}$, for a Dirichlet process (red line) and a Pitman-Yor process (blue line). The Good-Toulmin estimator (green line) is also displayed for comparison.

In order to illustrate the BNP methodology for species sampling problems we consider two real-world applications. First we analyze a dataset of Holst (1981), which comprises $n = 204$ coins (obverse side) found in a hoard of ancient coins. They were classified according to different die types, which are the tools used to produce them and represent the "species" in this context: $k = 141$ distinct dies were recorded. The frequencies are conveniently summarized in terms of the number of "species" of a given size, i.e., r_i represents the number of species with frequency i, and we have $(r_1, r_2, r_3, r_4, r_5, r_6, r_7) = (102, 26, 8, 2, 1, 1, 1)$. The goal is to predict the number of new dies one may observe in a future hoard of coins. We adopt the BNP estimators $\mathbb{E}(K_m^{(n)} \mid X^{(n)})$ with empirical Bayes elicitation of (σ, θ), which yields the parameter specifications $\hat{\theta} = 200.65$ for the Dirichlet process and $(\hat{\sigma}, \hat{\theta}) = (0.11, 172.48)$ for the Pitman-Yor process. The two BNP estimators are compared to the Good-Toulmin estimator (Good and Toulmin, 1956; Mao, 2004), which is one of the most popular frequentist estimators. It is well-known that the Good-Toulmin estimator usually provides reliable predictions for $m < n$, that is, if the additional sample over which predictions are made is not larger than the observed sample, whereas for values $m > n$ it often shows an erratic behavior. All three estimators are depicted in Figure 1a. The Dirichlet process and the Pitman-Yor process behave similarly. This is not surprising since the empirical Bayes estimate favors a low value for the discount parameter ($\hat{\sigma} = 0.11$), and we recall that the Pitman-Yor process reduces to the Dirichlet process if $\sigma = 0$. Moreover, we highlight that they both resemble the behavior of the Good-Toulmin estimator when m is smaller than the observed sample size. For larger values of m, the Good-Toulmin estimator becomes erratic, whereas this is not the case for the BNP estimators since they rely on principled probabilistic models.

Our second application concerns the genomic dataset FlavFr1: it comprises 1593 expressed sequence tags (ESTs) of a cDNA library obtained from the fruits of citrus clementina. These are categorized into 806 different genes with $r_i = 561, 148, 37, 18, 6, 5, 12, 1, 1, 3, 1, 2$ for $i \in \{1, \ldots, 12\}$, and $r_i = 3, 2, 1, 1, 1, 1, 1, 1$, for $i \in \{14, 15, 16, 19, 22, 23, 58, 117\}$. As before, we aim to predict the number of new genes in a future sample; the empirical Bayes specification yields $\hat{\theta} = 651.05$ and $(\hat{\sigma}, \hat{\theta}) = (0.63, 110.24)$ for the Dirichlet and the Pitman-Yor processes, respectively. The corresponding estimators together with the Good-Toulmin estimator are depicted in Figure 1b. In this case there are striking differences between the estimators obtained from the Dirichlet and the Pitman-Yor processes. This is due to the data favoring a large value for the discount parameter ($\hat{\sigma} = 0.63$), setting the Pitman-Yor case far apart from the Dirichlet process case ($\hat{\sigma} = 0$). This clearly showcases the usefulness of the additional

flexibility of the Pitman-Yor process, which allows for polynomial growth rates controlled by the parameter σ. Once again, the Good-Toulmin estimator diverges for large values of m. However, for moderately large values of m it nicely resembles the predictions of the Pitman-Yor estimator. This further underpins the need to go beyond the Dirichlet process in species sampling contexts. Further instances of applications requiring nonparametric priors with polynomial growth rate, or equivalently power law tail behavior, can be found in Hoshino (2001), Teh (2006), Caron (2012) and Caron and Fox (2017).

3 Mixture models

The Dirichlet and Pitman-Yor process priors are laws for almost surely discrete random probabilities. This implies that, when used as de Finetti measures, the induced exchangeable observations (1) will have a positive probability of being equal, i.e., $\mathbb{P}(X_i = X_j) > 0$ for every i, j. Thus, different models should be considered whenever the data do not display ties. A popular strategy is to model a random density function through a kernel mixture. Let f be a probability density kernel and $\tilde{P}$ be an almost surely discrete random probability. The random probability density is then defined as

$$\tilde{f}(y) = \int_{\mathbb{X}} f(y \mid x)\tilde{P}(\mathrm{d}x).$$

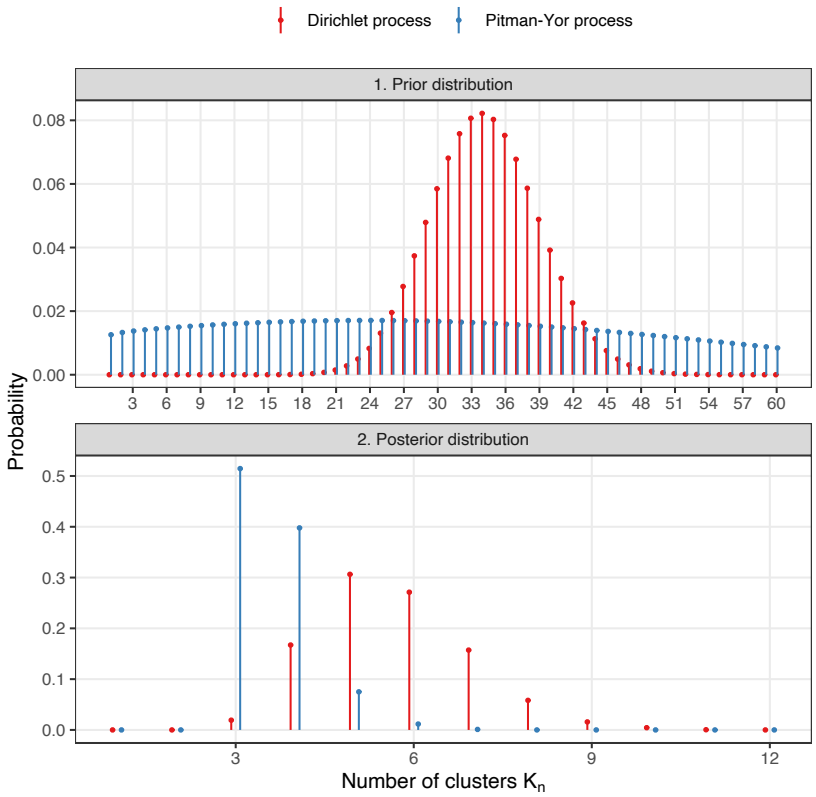
Using the law Q_f of $\tilde{f}$ as de Finetti measure, the law of induced exchangeable observations $(Y_i)_{i\geq 1}$ may be equivalently described in hierarchical form as

$$\begin{aligned} Y_i \mid X_i, \tilde{P} &\overset{\text{ind}}{\sim} f(\cdot \mid X_i) && i = 1, \ldots, n; \\ X_i \mid \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} && i = 1, \ldots, n; \\ \tilde{P} &\sim Q, \end{aligned}$$

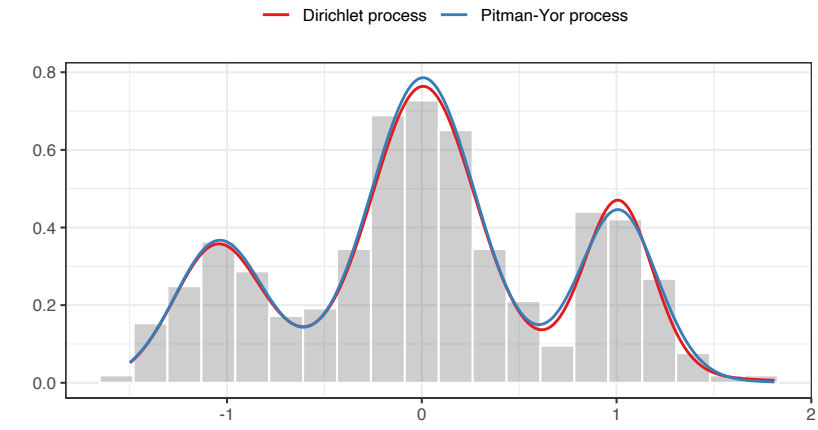
where the latent variable X_i represents the parameter of the probability density from which Y_i is sampled. Kernel mixtures over a Dirichlet process have been introduced in Lo (1984) and later extended to more general mixing measures, including the Pitman-Yor process (Ishwaran and James, 2001). See Müller and Quintana (2004), Barrios et al. (2013) and De Blasi et al. (2015) for reviews. Arguably, kernel mixtures over discrete random probabilities are the most popular Bayesian nonparametric models. This is because they allow one to perform two important statistical tasks at once: (i) flexible density estimation, which avoids parametric constraints and adapts to any data generating distribution; (ii) model-based clustering,

which exempts from fixing the number of clusters a priori. Indeed, due to the discreteness of $\tilde{P}$, any sample from the posterior distribution of the latent variables X_i given $Y_{(n)}$ has $K_n \in \{1, \dots, n\}$ distinct values. This automatically partitions the n observed data into K_n clusters, with two observations Y_i and Y_j belonging to the same cluster if they are sampled from the same mixture component, i.e., if $f(\cdot \mid X_i) = f(\cdot \mid X_j)$ a posteriori.
Different discrete nonparametric priors induce different distributions for the number of clusters K_n. Given the important role played by K_n for the probabilistic clustering, when eliciting the prior this aspect should be taken into account. In the same way, the posterior distribution of $\tilde{P}$ given the data $Y^{(n)}$ will induce a posterior distribution $\mathscr{L}(K_n \mid Y^{(n)})$, providing both point estimates and uncertainty quantification on the number of clusters in the data. In practice, posterior computations are customarily carried out using Markov Chain Monte Carlo, an avenue first explored in Escobar and West (1995). This is a highly non-trivial task, since it entails exploring the partition space. To address this issue, several computational methods have been proposed over the years. Another practical concern is the choice of the baseline distribution P^*, which is known to have a significant impact on $\mathscr{L}(K_n \mid Y^{(n)})$. See Richardson and Green (1997). However, this equally affects parametric and nonparametric specifications.

In Figure 2 we consider synthetic data generated from a finite mixture of normal distributions with three components, and the estimates resulting from a normal mixture model with Dirichlet and Pitman-Yor process as mixing measures. We consider a highly miss-specified scenario, where the expected number of clusters a priori is about 34, significantly larger than the *true* number of mixing components (three). This leads to consider a Dirichlet process with $\theta = 10$, whereas for the Pitman-Yor process infinite combinations of σ and θ are possible: we set $\sigma = 0.6$ to highlight the role of σ, and this results in $\theta = 0.026$. In particular, the plot displays the prior and posterior distributions of the number of clusters in Figure 2a, and the posterior densities in Figure 2b. For both models the posterior distribution moves away from the prior miss-specification. However, in the Pitman-Yor case this correction is stronger leading to more accurate posterior inference on the number of clusters. The robustness of the clustering with respect to prior miss-specifications represents a highly appealing feature of Pitman-Yor process mixtures. In contrast, both models lead to highly accurate but essentially indistinguishable posterior densities. This is due to the lack of identifiability of the number of mixture components in a mixture density: one can always fit a mixture density with more components than needed.



(a) Distribution of the number of clusters K_{300}: prior distribution and posterior update given $Y^{(300)}$. Probability masses corresponding to the same number of clusters are slightly shifted for visualization purposes.



(b) Histogram of $Y^{(300)}$, together with the Monte Carlo approximations of the posterior means $\mathbb{E}(\tilde{f}(y) \mid Y^{(300)})$, for a grid of values of y.

Figure 2: Mixture models. Nonparametric density estimation using a synthetic dataset of $n = 300$ observations from a mixture of Gaussians $1/4\,\mathrm{N}(-1, 1/4^2) + 1/2\,\mathrm{N}(0, 1/4^2) + 1/4\,\mathrm{N}(1, 1/4^2)$, under the Dirichlet process (red lines) and the Pitman-Yor process (blue lines).

4 Borrowing of information

In the previous sections we focused on exchangeable models, which translate a notion of homogeneity in the data. Yet, there are many situations where this can be seen as a restrictive assumption. In de Finetti's words (de Finetti, 1938) *'But the case of exchangeability can only be considered as a limiting case: the case in which this "analogy" is, in a certain sense, absolute for all events under consideration. '*. More specifically, one may wish to generalize exchangeability to the case where data is collected under different experimental conditions, such that one retains homogeneity within each experiment though allowing for heterogeneity across different experiments. Typical examples include topic modeling, meta-analysis, two-sample problems, nonparametric regression, time-dependent data, and change point problems, to mention a few. A natural probabilistic framework that achieves this is partial exchangeability, as defined below. To simplify the notation, we focus on two groups of observations, though all the contents of this section may be easily extended to an arbitrary number of groups.

Definition 3. *An array* $(\boldsymbol{X}_1, \boldsymbol{X}_2) = (X_{1,j}, X_{2,j})_{j\geq 1}$ *is partially exchangeable if for any* $N_1, N_2 \geq 1$*,* π *permutation of* $\{1, \ldots, N_1\}$*, and* ϕ *permutation of* $\{1, \ldots, N_2\}$*,*

$$(X_{1,1}, \ldots, X_{1,N_1}, X_{2,1}, \ldots, X_{2,N_2}) \stackrel{\mathrm{d}}{=} (X_{1,\pi(1)}, \ldots, X_{1,\pi(N_1)}, X_{2,\phi(1)}, \ldots, X_{2,\phi(N_2)}).$$

Partially exchangeable sequences may be characterized as conditionally independent through an extension of de Finetti's representation theorem. Specifically, $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is partially exchangeable if and only if there exists a probability distribution Q on $\mathscr{P}_{\mathbb{X}} \times \mathscr{P}_{\mathbb{X}}$ such that

$$\mathbb{P}(X_{1,1} \in A_1, \ldots, X_{1,N_1} \in A_{N_1}, X_{2,1} \in B_1, \ldots, X_{2,N_2} \in B_{N_2}) = \\ = \int_{P_{\mathbb{X}} \times P_{\mathbb{X}}} \prod_{i=1}^{N_1} P_1(A_i) \prod_{j=1}^{N_2} P_2(B_j)\, Q(\mathrm{d}P_1, \mathrm{d}P_2).$$

Equivalently, there exists a vector of dependent random probabilities $(\tilde{P}_1, \tilde{P}_2)$ such that

$$\begin{aligned} (X_{1,i}, X_{2,j}) \mid (\tilde{P}_1, \tilde{P}_2) &\stackrel{\text{iid}}{\sim} \tilde{P}_1 \times \tilde{P}_2 \qquad && \forall i, j \geq 1; \\ (\tilde{P}_1, \tilde{P}_2) &\sim Q. \end{aligned}$$

We observe that the dependence between random probabilities induces dependence between the sequences $\boldsymbol{X}_1, \boldsymbol{X}_2$ of (exchangeable) observations, with two extreme scenarios:

when $\tilde{P}_1$ and $\tilde{P}_2$ are independent, so are $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$; when $\tilde{P}_1 = \tilde{P}_2 = \tilde{P}$ almost surely, $\boldsymbol{X}_1, \boldsymbol{X}_2$ are (fully) exchangeable, as it is clear by comparing the "degenerate" form of de Finetti's representation theorem to the one for exchangeable sequences (1). Thus, we can interpret partial exchangeability as a dependence assumption on the observables that ranges from independence to exchangeability.

The use of partially exchangeable sequences for statistical purposes has been pioneered by Cifarelli and Regazzini (1978), MacEachern (1999) and MacEachern (2000). To this end, one needs to build dependent random probability measures (see Quintana et al. (2022) for a recent review) with two key features in mind: (i) mathematical tractability, which corresponds to obtaining manageable representations for the posterior and/or the marginal structure, i.e., the partition distribution or the prediction rule; (ii) the ability to control the amount of dependence, since this is directly linked to the borrowing of information between the two groups: the more the dependence, the more information will be shared across the two groups. This is usually done by expressing the linear correlation between pairwise set-wise evaluations, $\text{Cor}(\tilde{P}_1(A), \tilde{P}_2(A))$ for any Borel set A, and has been recently extended to an arbitrary number of groups by relying on the Wasserstein distance (Catalano et al., 2021b,a).

There are many proposals in the literature that share both these features. We divide them into three categories: hierarchical structures (e.g., Teh et al. (2006); Teh and Jordan (2010); Camerlenghi et al. (2017, 2019b, 2021)), nested structures (e.g.,Rodríguez et al. (2008); Camerlenghi et al. (2019a); Lijoi et al. (2023)), and multivariate Lévy structures (e.g., Epifani and Lijoi (2010); Lijoi et al. (2014a,b); Griffin and Leisen (2017); Lau and Cripps (2022)).

Partial exchangeability represents the ideal probabilistic framework in many contexts. We showcase this by means of the hierarchical Dirichlet process mixture model (Teh et al., 2006), which is among the most popular and intuitive ways of introducing dependence between random probabilities and also enjoys attractive frequentist asymptotic properties (Catalano et al., 2022). We define $(\tilde{P}_1, \tilde{P}_2)$ to be distributed according to a hierarchical Dirichlet process prior if for $\alpha, \alpha^* > 0$ and P^* a diffuse probability measure on $\mathbb{X}$ such that

$$
\begin{aligned}
(\tilde{P}_1, \tilde{P}_2) \mid \tilde{P}_0 &\overset{\text{iid}}{\sim} \text{DP}(\alpha, \tilde{P}_0); \\
\tilde{P}_0 &\sim \text{DP}(\alpha^*, P^*).
\end{aligned}
$$

We will use the notation $(\tilde{P}_1, \tilde{P}_2) \sim \text{HDP}(\alpha, \alpha^*, P^*)$.

Different levels of dependence may be achieved by tuning the concentration hyperparam-

eters $\alpha, \alpha^* > 0$ of the HDP. The corresponding correlation structure is

$$\text{Cor}(\tilde{P}_1(A), \tilde{P}_2(A)) = \frac{\alpha + 1}{\alpha + 1 + \alpha^*},$$

which remarkably does not depend on the set A nor on the base probability P^*. Starting from a hierarchical Dirichlet process, one can use it to induce dependence between mixture densities in a natural way, that is,

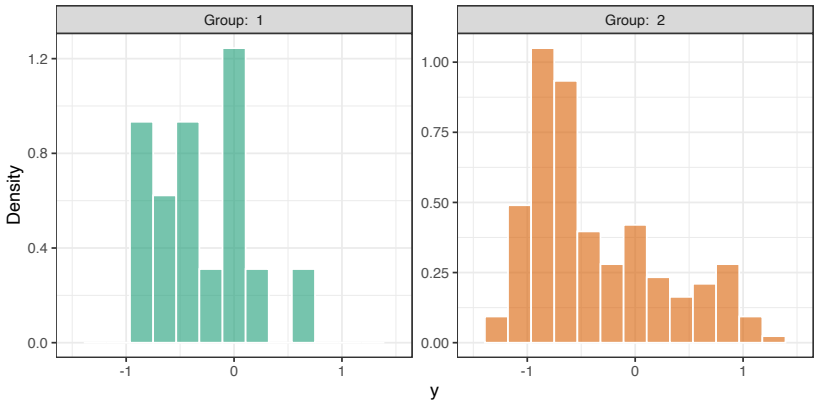
$$\begin{aligned}
(Y_{1,i}, Y_{2,j}) \mid (X_{1,i}, X_{2,j}, \tilde{P}_1, \tilde{P}_2) &\overset{\text{ind}}{\sim} f(\cdot \mid X_{1,i}) \times f(\cdot \mid X_{2,j}) && \forall i, j \geq 1; \\
(X_{1,i}, X_{2,j}) \mid (\tilde{P}_1, \tilde{P}_2) &\overset{\text{iid}}{\sim} \tilde{P}_1 \times \tilde{P}_2 && \forall i, j \geq 1; \\
(\tilde{P}_1, \tilde{P}_2) &\sim \text{HDP}(\alpha, \alpha^*, P^*).
\end{aligned}$$

To test the performance of the HDP in terms of borrowing of information we consider synthetic data generated from mixtures of three normal distributions. More specifically, the first group of $n_1 = 15$ synthetic observations are iid from the mixture $0.6\,\text{N}(-0.8, 0.2^2) + 0.3\,\text{N}(0, 0.2^2) + 0.1\,\text{N}(0.8, 0.2^2)$, whereas the second group of $n_2 = 200$ observations are iid from the mixture $0.575\,\text{N}(-0.8, 0.2^2) + 0.275\,\text{N}(0, 0.2^2) + 0.15\,\text{N}(0.8, 0.2^2)$. In other words, the two mixtures have the same scale and location parameters, but different probability weights. Moreover, the two groups of observations have markedly different sample sizes. As for the HDP model, we compare two different scenarios. We tune the parameters to achieve different values for the correlation, namely 0.09 and 0.91, corresponding to, respectively, weakly and highly dependent specifications. Among the infinite choices for α and α^* that induce the aforementioned values for the correlation, in both cases we select those that lead to an overall prior expected number of clusters among the two groups approximately equal to 7.5. This allows for a fair comparison of the two models. The resulting parameters are $\alpha = 1$ and $\alpha^* = 20$ for the weakly and $\alpha = 19$ and $\alpha^* = 2$ highly dependent HPD. In Figure 3 we display the corresponding mean posterior densities

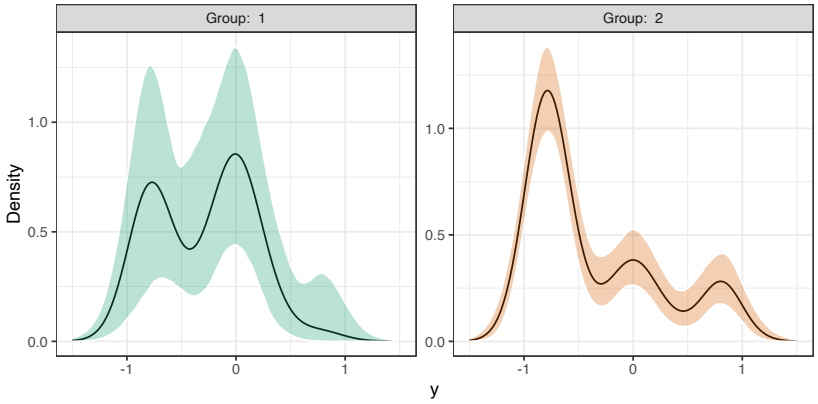
$$\mathbb{E}\left(\int_{\mathbb{X}} f(y \mid x) \tilde{P}_j(\mathrm{d}x) \,\middle|\, Y_{1,1}, \ldots, Y_{1,15}, Y_{2,1}, \ldots, Y_{2,200} \right).$$

As in the exchangeable case, posterior computations are based on Markov Chain Monte Carlo, leveraging specialized algorithms such as Lijoi et al. (2020).

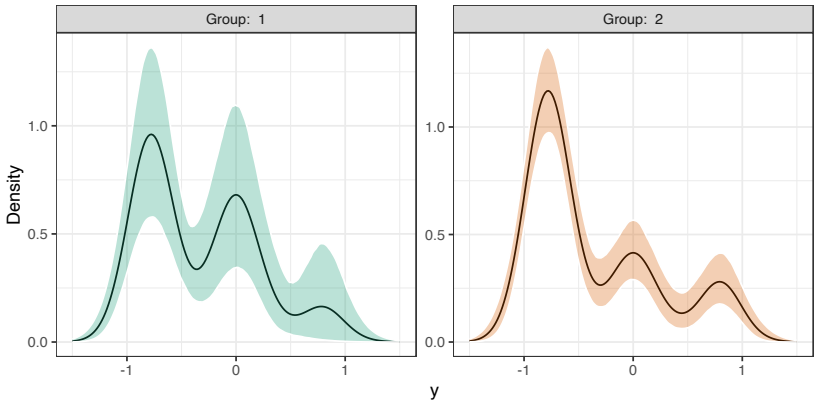
The plots in Figure 3 highlight the crucial role of the specification of the dependence structure. In the case of an HDP with weak dependence (Figure 3b), the data of the first group (left panel) are not sufficiently informative to discover the existence of a third mixture component. Conversely, when we impose a stronger dependence (Figure 3c), the data of



(a) Histograms of the two groups of data $Y_{1,1}, \ldots, Y_{1,15}$ (left panel) and $Y_{2,1}, \ldots, Y_{2,200}$ (right panel).



(b) Low correlation structure (0.09): solid lines are Monte Carlo approximations of the mean posterior densities, shaded areas are 95% pointwise credible intervals.



(c) High correlation structure (0.91): solid lines are Monte Carlo approximations of the mean posterior densities, shaded areas are 95% pointwise credible intervals.

Figure 3: Borrowing of information: nonparametric density estimation using a Hierarchical Dirichlet process for two groups of synthetic observations with 15 and 200 data points, respectively.

the first group can effectively borrow strength from the second, whose sample size is much larger and therefore contains more information. As a result, the correct number of mixture components is recovered, despite the limited sample size of the first group. Moreover, the borrowing of information is also apparent from the credible intervals: a larger borrowing of information greatly reduces the uncertainty around the point estimator for the first group.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2:1152 – 1174.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science*, 28:313–334.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355.

Camerlenghi, F., Dunson, D., Lijoi, A., Prünster, I., and Rodriguez, A. (2019a). Latent nested nonparametric priors. (With discussion). *Bayesian Analysis*, 15:1303–1356.

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

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis*, 156:18–28.

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

Carnap, R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*, volume 25.

Caron, F. and Fox, E. B. (2017). Sparse Graphs Using Exchangeable Random Measures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79:1295–1366.

Catalano, M., De Blasi, P., Lijoi, A., and Prünster, I. (2022). Posterior asymptotics for boosted hierarchical Dirichlet process mixtures. *Journal of Machine Learning Research*, 23(80):1–23.

Catalano, M., Lavenant, H., Lijoi, A., and Prünster, I. (2021a). A Wasserstein index of dependence for random measures. *arXiv 2109.06646*.

Catalano, M., Lijoi, A., and Prünster, I. (2021b). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics*, 49:2916–2947.

Cifarelli, D. M. and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. *Quaderni Istituto Matematica Finanziaria dell'Università di Torino Serie III*, 12:1–36.

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*, 37:212–229.

de Finetti, B. (1937). La prévision, ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincaré*, 7:1–68.

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

Epifani, I. and Lijoi, A. (2010). Nonparametric priors for vectors of survival functions. *Statistica Sinica*, 20:1455–1484.

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

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112.

Favaro, S., Lijoi, A., Mena, R. H., and Prünster, I. (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71:993–1008.

Favaro, S., Prünster, I., and Walker, S. G. (2011). On a class of random probability measures with general predictive structure. *Scandinavian Journal of Statistics*, 38:359–376.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1:209 – 230.

Ferguson, T. S. (1974). Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, 2:615 – 629.

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

Good, I. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43:45–63.

Griffin, J. E. and Leisen, F. (2017). Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79:525–545.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80:470–501.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors (2010). *Bayesian nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Holst, L. (1981). Some asymptotic results for incomplete multinomial or Poisson sample. *Scandinavian Journal of Statistics*, 8:243–246.

Hoshino, N. (2001). Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, 17:499–520.

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

Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *The Annals of Probability*, 1:705–711.

Lau, J. W. and Cripps, E. (2022). Thinned completely random measures with applications in competing risks models. *Bernoulli*, 28:638 – 662.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94:769–786.

Lijoi, A., Mena, R. H., and Prünster, I. (2008). A Bayesian nonparametric approach for comparing clustering structures in est libraries. *Journal of Computational Biology*, 15:1315–1327.

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

Lijoi, A., Nipoti, B., and Prünster, I. (2014b). Dependent mixture models: clustering and borrowing information. *Computational Statistics and Data Analysis*, 71:417–433.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In Hjort, N. L., Holmes, C. C., Muller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge University Press, Cambridge.

Lijoi, A., Prünster, I., and Rebaudo, G. (2023). Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics*, 50:213–234.

Lijoi, A., Prünster, I., and Rigon, T. (2020). Sampling hierarchies of discrete random structures. *Statistics and Computing*, 30:1591–1607.

Lindley, D. V. (1972). *Bayesian statistics: A review*. Society for Industrial and Applied Mathematics, Philadelphia.

Lo, A. (1984). On a class of Bayesian nonparametric estimates. *The Annals of Statistics*, 12:351–57.

Lo, A. Y. (1991). A characterization of the Dirichlet process. *Statistics & Probability Letters*, 12:185–187.

MacEachern, S. N. (1999). Dependent nonparametric processes. *In ASA Proceedings of the Section on Bayesian Statistical Science, American Statistical Association*.

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

Mano, S. (2018). *Partitions, Hypergeometric Systems, and Dirichlet Processes in Statistics*. Springer, Tokyo.

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *Journal of the American Statistical Association*, 99:1108–1118.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, 19:95–110.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer Series in Statistics. Springer, Cham.

Müller, P., Quintana, F. A., and Page, G. (2018). Nonparametric Bayesian inference in applications. *Statistical Methods & Applications*, 27:175–206.

Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92:21–39.

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

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Math. 1875, Springer, Berlin.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25:855 – 900.

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

Regazzini, E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giornale Istituto Italiano Attuari*, 41:77–89.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 59:768–769.

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

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

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 985–992, USA. Association for Computational Linguistics.

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

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

Yamato, H. (2020). *Statistics Based on Dirichlet Processes and Related Topics*. Springer, Singapore.