Received: 30 November 2023 | Revised: 31 July 2024 | Accepted: 18 December 2024

DOI: 10.1111/sjos.12765

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Enriched Pitman–Yor processes

Tommaso Rigon[1] | **Sonia Petrone**[2] | **Bruno Scarpa**[3]

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

[2]Department of Decision Sciences and Bocconi Institute of Data Science and Analytics, Bocconi University, Italy

[3]Department of Statistical Sciences, Università degli studi di Padova, Italy

Correspondence
Tommaso Rigon, Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy.
Email: tommaso.rigon@unimib.it

Funding information
National Institute of Environmental Health Sciences, Grant/Award Number: R01ES027498; Ministero dell'Università e della Ricerca, Grant/Award Number: 2022CLTYP4

Abstract

Bayesian non-parametrics has evolved into a broad area encompassing flexible methods for Bayesian inference, combinatorial structures, tools for complex data reduction, and more. Discrete prior laws play an important role in these developments, and various choices are available nowadays. However, many existing priors, such as the Dirichlet process, have limitations if data require nested clustering structures. Thus, we introduce a discrete non-parametric prior, termed the enriched Pitman–Yor process, which offers higher flexibility in modeling such elaborate partition structures. We investigate the theoretical properties of this novel prior and establish its formal connection with the enriched Dirichlet process and normalized random measures. Additionally, we present a square-breaking representation and derive closed-form expressions for the posterior law and associated urn schemes. Furthermore, we demonstrate that several established models, including Dirichlet processes with a spike-and-slab base measure and mixture of mixtures models, emerge as special instances of the enriched Pitman–Yor process, which therefore serves as a unified probabilistic framework for various Bayesian non-parametric priors. To illustrate its practical utility, we employ the enriched Pitman–Yor process for a species-sampling ecological problem.

KEYWORDS

Bayesian non-parametrics, enriched Dirichlet process, mixture of mixtures, nested random partitions, species sampling models, spike and slab processes

1 | INTRODUCTION

In the early days, Bayesian non-parametric inference faced theoretical and practical challenges due to the complexities of defining a prior distribution on a space of infinite-dimensional objects, where the realizations are probability measures. The first contributions to the construction of random distributions include Kraft (1964) and Dubins and Freedman (1967), followed by Ferguson's Dirichlet process (Ferguson, 1973) and the nearly contemporary work by Blackwell and MacQueen (1973), Antoniak (1974), Ferguson (1974), which marked a significant turning point. Since then, the field of *Bayesian non-parametrics* experienced substantial growth, covering classical problems like survival analysis (Doksum, 1974; Ghosh et al., 2006; Hjort, 1990; Phadia, 2013; Walker & Muliere, 1997), quantile inference (Hjort & Petrone, 2007; Hjort & Walker, 2009; Kottas & Krnjajić, 2009), non-parametric regression (e.g. Quintana et al., 2022), and frequentist theoretical properties (Ghosal & van der Vaart, 2017), to encompass a wide array of applications in various scientific fields such as ecology, genetics, population dynamics, and biostatistics (Dunson, 2010; Müller et al., 2015), with notable intersections in machine learning (Teh & Jordan, 2010). A recent research direction, based on a Bayesian *predictive* approach, provides uncertainty quantification for classes of predictive algorithms (Fortini & Petrone, 2020) and also builds on Bayesian non-parametric methodologies. In a related stream of work, Fong et al. (2023) interestingly elaborate the Bayesian bootstrap (see, e.g. Hjort, 1985) under a novel perspective based on martingale posteriors.

Several extensions of the Dirichlet process (DP), arguably the most widely studied non-parametric prior, have emerged over the years. These include the Pitman–Yor process (PY) discussed in Perman et al. (1992), Pitman and Yor (1997), the generalized Dirichlet process (Hjort, 2000), the wide class of Gibbs-type priors (De Blasi et al., 2015), normalized random measures with independent increments (Lijoi et al., 2007b; Regazzini et al., 2003), and species sampling models (Pitman, 1996). For a comprehensive overview, we refer to Hjort et al. (2010). More recently, there has been a lively stream of research focusing on discrete non-parametric priors for partially exchangeable data, based on nested (Camerlenghi, Dunson, et al., 2019; Rodríguez et al., 2008), additive (Lijoi et al., 2014), and hierarchical constructions (Camerlenghi, Lijoi, et al., 2019; Teh et al., 2006).

Despite these numerous advances, non-parametric priors specifically designed for multivariate random measures, say on $\mathbb{R}^d$, seem less developed. The DP and its extensions are mathematically well-defined on general spaces, but they might be too rigid to capture the complexity of multivariate distributions. This rigidity stems from the Dirichlet finite-dimensional distributions inherent in the DP. Indeed, Doksum (1974) used the generalized Dirichlet distribution (Connor & Mosimman, 1969) in defining neutral to the right processes for *univariate* random distributions. Extending Doksum's construction to random distributions on $\mathbb{R}^d$ is not obvious, since the generalized Dirichlet distribution loses the property of complete neutrality, thus an ordering has to be chosen in the sample space; and while there is a natural ordering in the real line, this is not the case for $\mathbb{R}^d$. However, an ordering is often suggested by the applied context, whenever it is natural to regard the sample space as a product space, say $\mathbb{R}^k \times \mathbb{R}^{d-k}$. In such settings, Wade et al. (2011) have proposed to *enrich* the DP by using novel enriched Dirichlet distributions as the finite-dimensional laws of the process. The resulting enriched Dirichlet process (EDP) allows for more freedom in the parametrization while preserving conjugacy. However, it still suffers from some rigidity due to the Dirichlet components in its construction. One could address these limitations by further generalizing the finite-dimensional distributions of the process; however, this direction seems to become unnecessarily complex. In this paper, we take a different approach.

Moving from a predictive perspective (de Finetti, 1937), we rather reason on the predictive rule that characterizes the prior law and on the implied clustering structure.

Let $\mathbb{X}$ and $\mathbb{Y}$ be two complete and separable Polish spaces and let $((X_n, Y_n))_{n\geq 1}$ be an infinite exchangeable sequence with elements taking values in the product space $\mathbb{X} \times \mathbb{Y}$. By de Finetti representation theorem (de Finetti, 1937), there exists a unique random probability measure $\tilde{p}$ such that the (X_n, Y_n) are independent and identically distributed (iid) conditionally on $\tilde{p}$

$$\begin{aligned} (X_n, Y_n) \mid \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \qquad n \geq 1, \\ \tilde{p} &\sim Q, \end{aligned} \tag{1}$$

where Q is the probability law of the random $\tilde{p}$, serving as the prior law in Bayesian inference. In our motivating application, the random variables X_n and Y_n represent families and species of trees sampled in the Amazonian basin, respectively. The interest lies in predicting the number of novel families and species one would get in a future sample. Thus, neither the DP, the PY, or other species sampling models are suitable priors Q for this specific application. These choices would associate the discovery of a new species with that of a new family, which is inappropriate. Instead, we seek priors that induce a *nested clustering* mechanism so that the discovery of a new family of trees corresponds to that of a new species, but not vice versa. Note that we are considering exchangeable data and not, for example, data from stratified sampling, where observations would be exchangeable within each stratum but not across them. In such cases, an assumption of partial exchangeability would be more appropriate. Still, hierarchical priors for partially exchangeable data, such as the hierarchical Dirichlet process of Teh et al. (2006), would not provide the desired nested clustering within the strata. Here, we focus on the exchangeable setting, which is the basis for extensions to more structured sampling schemes.

Whenever observations suggest a nested structure, as in our motivating application, the EDP is more suitable prior for the joint random probability measure $\tilde{p}$. The EDP allows for finer control of the dependence structure between the X_n and Y_n, leading to the desired *nested clustering*. These appealing features have been leveraged among others in Wade et al. (2014); Gadd et al. (2020) for Bayesian non-parametric regression models, in Roy et al. (2018) for causal inference with missing covariates, and in Zeldow et al. (2021) for functional clustering of longitudinal data; an interesting recent developement is Franzolini et al. (2023). Despite its advantages, the EDP inherits some drawbacks shared by all Dirichlet-based priors, motivating our extension. In particular, in the predictive rule of the DP (Blackwell & MacQueen, 1973), the probability of observing a new species depends solely on the sample size. As remarked by Lijoi et al. (2007a); De Blasi et al. (2015), this feature of the predictive rule is problematic in species sampling problems because the predictive probability of discovering a new species does not depend on the data. Moreover, it leads to a logarithmic growth of the number of clusters and to a lack of robustness to miscalibrated prior choices, which might be undesirable in several applied contexts; see, for example, Lijoi et al. (2007b) and De Blasi et al. (2015). To address these issues, we propose a novel discrete prior law Q building upon the EDP and the PY process. By combining their appealing properties, the proposed *enriched Pitman–Yor* process (EPY) leads to different rates for the discovery of new species, a key aspect in species sampling models. Importantly, improved flexibility is achieved while preserving analytical and computational tractability. We obtain a simple urn scheme, a tractable posterior characterization, and a square-breaking representation for the EPY. In addition, we show that the EPY can be defined by normalizing a suitable random measure. This alternative definition parallels the construction of Regazzini et al. (2003), which bears important theoretical implications.

The EDP is designed for observations in product spaces $\mathbb{X} \times \mathbb{Y}$; yet, interestingly, the X_n component does not need to be observable. Given a random distribution $\tilde{p}$ on $\mathbb{X} \times \mathbb{Y}$ with an EPY probability law, we define a *marginal* EPY process as

$$\tilde{p}_Y(B) = \tilde{p}(\mathbb{X} \times B), \qquad \tilde{p} \sim Q, \tag{2}$$

for any measurable set B of $\mathbb{Y}$. The space $\mathbb{X}$ can be interpreted as a latent dimension that induces enriched specifications. Thus, the marginal EPY can be used as the prior law Q_Y for exchangeable sequences $(Y_n)_{n\geq 1}$ taking values in $\mathbb{Y}$

$$\begin{aligned} Y_n \mid \tilde{p}_Y &\overset{\text{iid}}{\sim} \tilde{p}_Y, \qquad n \geq 1, \\ \tilde{p}_Y &\sim Q_Y. \end{aligned} \tag{3}$$

We show that the marginal EPY $\tilde{p}_Y$ is an infinite mixture of PY processes, which is arguably much more flexible than a single PY process. It encompasses a wide variety of prior proposals in the literature as a special case. To the best of our knowledge, their connection with enriched processes has not been previously emphasized. For example, a specific marginal EPY process has been implicitly studied in Scarpa and Dunson (2014) and Rigon (2023) for the analysis of functional data. Dirichlet processes with spike and slab base measures (e.g. Dunson et al., 2008; Guindani et al., 2009; MacLehose et al., 2007), or general atomic contaminations (Scarpa & Dunson, 2009), are in fact special cases of a marginal EDP. Similarly, a mixture of finite-dimensional DP's has been employed in Malsiner-Walli et al. (2017) and Rigon (2023) to perform model-based clustering, while convex combinations of DP's have been considered by Müller et al. (2004), Lijoi et al. (2014) to induce dependence across groups of random variables. These models are closely linked to the marginal EPY. Thus, as a further contribution in this paper, we highlight the connections between the EPY and the aforementioned methods, providing a unified probabilistic framework for these classes of processes that allows us to naturally suggest several extensions.

The paper is organized as follows. In Section 2 we introduce the EPY process and we discuss its fundamental probabilistic characterizations, including the square-breaking construction. In Section 3, we discuss its predictive rule, providing an enriched urn scheme, and posterior representations. In Section 4, we illustrate the marginal EPY process and its connection with several existing approaches, and we propose numerous extensions. In Section 5, we employ the EPY to estimate the number of unobserved species in the Amazonian tree flora. Final remarks are provided in Section 6. All the proofs are collected in the Appendix.

2 | THE ENRICHED PITMAN–YOR PROCESS

2.1 | Background material

Let P be a probability measure on $\mathbb{X} \times \mathbb{Y}$ and let $(\nu_j)_{j\geq 1}$ be a sequence of independent Beta random variables such that $\nu_j \sim \text{BETA}(1-\sigma, \alpha + j\sigma)$, where $\sigma \in [0, 1)$ and $\alpha > -\sigma$, or $\sigma = -\alpha/H$ and $\alpha > 0$ for some integer $H \in \{2, 3, \dots\}$. A discrete random probability measure $\tilde{p}$ follows a Pitman–Yor (PY) process with parameters $(\sigma, \alpha P)$, denoted as $\tilde{p} \sim \text{PY}(\sigma, \alpha P)$, if

$$\tilde{p}(\cdot) = \sum_{h=1}^{\infty} \xi_h \, \delta_{(\phi_h, \theta_h)}(\cdot), \qquad (\phi_h, \theta_h) \overset{\text{iid}}{\sim} P, \tag{4}$$

where $\delta_{(x,y)}(\cdot)$ is the Dirac point measure at (x,y), and $\xi_h = \nu_h \prod_{j=1}^{h-1}(1-\nu_j)$ for $h \geq 1$, where we agree that $\xi_1 = \nu_1$. The parameter σ is often referred to as the discount parameter, while α is termed total mass or precision. When $\sigma = 0$, then $\tilde{p} \sim \text{DP}(\alpha P)$, a Dirichlet process with parameter (αP). If $\sigma = -\alpha/H$ and $\alpha > 0$, the stick-breaking construction is degenerate because $\nu_H = 1$, implying that $\sum_{h=1}^{H} \xi_h = 1$ for a finite $H < \infty$. This special case of PY, called Dirichlet multinomial process, or Fisher process, admits the following alternative representation:

$$\tilde{p}(\cdot) \overset{\text{d}}{=} \sum_{h=1}^{H} W_h \, \delta_{(\phi_h,\theta_h)}(\cdot), \qquad (W_1,\ldots,W_H) \sim \text{DIRICHLET}(\alpha/H,\ldots,\alpha/H),$$

where $(\phi_h, \theta_h) \overset{\text{iid}}{\sim} P$, with $\overset{\text{d}}{=}$ denoting the equality in distribution. Hence, when $\sigma < 0$, the PY a.s. selects discrete distributions with a finite number of atoms and symmetric Dirichlet weights. For further details, one may refer to appendix A.1 of Pitman and Yor (1997).

Although the DP is a special case of the Pitman–Yor model (4), in this paper we extensively employ an alternative construction based on completely random measures. Broadly speaking, a DP can be obtained as the normalization of a gamma process (Ferguson, 1973), which, in turn, can be represented as $\tilde{\mu}(\cdot) = \sum_{h=1}^{\infty} J_h \, \delta_{(\tilde{\phi}_h,\tilde{\theta}_h)}(\cdot)$, where $J_1 \geq J_2 \geq \ldots$ is a collection of ordered positive random jumps whose distribution is described in Ferguson and Klass (1972), independent on the random locations $(\tilde{\phi}_h, \tilde{\theta}_h) \overset{\text{iid}}{\sim} P$; further insights about the ordered jumps are given Hjort and Ongaro (2006). The law of a gamma random measure $\tilde{\mu}$ is uniquely characterized by its Laplace functional, namely

$$\mathbb{E}\left\{e^{-\tilde{\mu}(f)}\right\} = \exp\left\{-\alpha \int_{\mathbb{X}\times\mathbb{Y}} \log\{1+f(x,y)\} P(\text{d}x,\ \text{d}y)\right\}, \tag{5}$$

with $\alpha > 0$ and for any positive and measurable function $f : \mathbb{X}\times\mathbb{Y} \to \mathbb{R}^+$ such that $\tilde{\mu}(f) = \int_{\mathbb{X}\times\mathbb{Y}} f(x,y)\tilde{\mu}(\text{d}x,\ \text{d}y) < \infty$ almost surely. We write $\tilde{\mu} \sim \text{GAP}(\alpha P)$. The DP is then defined as the normalization of $\tilde{\mu}$, that is, if $\tilde{\mu} \sim \text{GAP}(\alpha P)$ then $\tilde{p}(\cdot) = \tilde{\mu}(\cdot)/\tilde{\mu}(\mathbb{X}\times\mathbb{Y}) \sim \text{DP}(\alpha P)$.

2.2 | The enriched Pitman–Yor process: Definition and stochastic representations

The proposed EPY process has the same elegant rationale that, in the parametric case, underlies the construction of enriched conjugate priors for natural exponential families (Consonni & Veronese, 2001). Roughly speaking, a multivariate distribution is decomposed in terms of the marginal and the conditional distributions, and an enriched prior law on its parameters is obtained by assigning independent priors on the parameters of the marginal and those of the conditional distributions. For multivariate parametric models in the natural exponential family, this construction is used to define enriched conjugate priors. In the non-parametric case, it is much more delicate as the distributions involved are *random* probability measures and, moreover, *conditional* random probability measures.

Definition 1. Let P be a probability measure on the product space $\mathbb{X}\times\mathbb{Y}$, with $P(A\times B) = \int_A P_{Y|X}(B\,|\,x)P_X(\text{d}x)$ for any Borel sets $A \subseteq \mathbb{X}$ and $B \subseteq \mathbb{Y}$. Moreover, let $\alpha > 0$ and $\sigma(x)$, $\beta(x)$ be functions such that either $\sigma(x) \in [0,1)$ and $\beta(x) > -\sigma(x)$ or $\sigma(x) = -\beta(x)/H(x)$ and $\beta(x) > 0$, where $H(x) \in \{2,3,\ldots\}$. Define a random probability

measure $\tilde{p}_X$ on $\mathbb{X}$ and a family of random probability measures $\tilde{p}_{Y|X}(\cdot\,|\,x)$ on $\mathbb{Y}$, for $x \in \mathbb{X}$, such that

$$\tilde{p}_X \sim \text{DP}(\alpha P_X), \qquad \tilde{p}_{Y|X}(\cdot\,|\,x) \overset{\text{ind}}{\sim} \text{PY}\{\sigma(x), \beta(x)P_{Y|X}(\cdot\,|\,x)\}, \qquad x \in \mathbb{X},$$

independently among themselves. Then the random probability measure $\tilde{p}$ on the product space $\mathbb{X} \times \mathbb{Y}$, defined as

$$\tilde{p}(A \times B) = \int_A \tilde{p}_{Y|X}(B\,|\,x)\tilde{p}_X(\text{d}x), \qquad A \subseteq \mathbb{X}, \quad B \subseteq \mathbb{Y}, \tag{6}$$

is said to be distributed as an enriched Pitman–Yor process (EPY) with parameters $\alpha P_X, \sigma(x)$ and $\beta(x)P_{Y|X}$. We will write $\tilde{p} \sim \text{EPY}(\alpha P_X, \sigma(x), \beta(x)P_{Y|X})$.

The EDP of Wade et al. (2011) is obtained as a special case of the EPY with $\sigma(x) = 0$ for all $x \in \mathbb{X}$. Ensuring that the above definition is well given, namely that it defines a probability measure on the random joint probability distribution $\tilde{p}$, is quite a subtle measure-theoretic issue. However, it can be proved by following the steps outlined in Wade et al. (2011) for the EDP, which draw upon results by Ramamoorthi and Sangalli (2006). In essence, the only prerequisites are that the conditionals $\tilde{p}_{Y|X}(\cdot\,|\,x)$ exhibit independence across $x \in \mathbb{X}$, the marginal $\tilde{p}_X$ is almost surely discrete, and $\tilde{p}_X$ and $\tilde{p}_{Y|X}$ are independent of each other; all of which are inherent properties of the EPY. A proof is provided in the Appendix for completeness. As suggested by one referee, more general "enriched" structures can be envisioned, either by considering a general species sampling model for the conditional laws $\tilde{p}_{Y|X}$ or by choosing a different distribution for the marginal $\tilde{p}_X$ – at the condition that the construction properly defines a random $\tilde{p}$ on the joint space. In these lines, our proposed process could be referred to as a Pitman–Yor process enriched through a Dirichlet process, explicitly indicating the distribution of $\tilde{p}_X$. We just use *enriched Pitman–Yor process* as there is no risk of ambiguity in our context.

The baseline measures $P_X, P_{Y|X}$, and P can be interpreted as "prior guesses" for the distribution of the observations. This is evident from:

$$\mathbb{E}\{\tilde{p}_X(A)\} = P_X(A), \quad \mathbb{E}\{\tilde{p}_{Y|X}(B\,|\,x)\} = P_{Y|X}(B\,|\,x), \quad \mathbb{E}\{\tilde{p}(A \times B)\} = P(A \times B),$$

for any $x \in \mathbb{X}$ and Borel sets $A \subseteq \mathbb{X}$, $B \subseteq \mathbb{Y}$, recalling that $P(A \times B) = \int_A P_{Y|X}(B\,|\,x)P_X(\text{d}x)$. Moreover, the DP on the product space $\mathbb{X} \times \mathbb{Y}$ is a limiting case of the EPY, when $\beta(x) \to -\sigma(x)$ for all $x \in \mathbb{X}$. In this scenario, each conditional law reduces to a point mass, that is, $\tilde{p}_{Y|X}(B\,|\,x) = \delta_{\theta_1(x)}(B)$, where $\theta_1(x) \sim P_{Y|X}(\cdot; x)$. Consequently, when $\beta(x) \to -\sigma(x)$ the joint distribution $\tilde{p}(A \times B)$ is expressed as:

$$\tilde{p}(A \times B) = \sum_{h=1}^{\infty} \xi_h\, \delta_{\phi_h}(A)\delta_{\theta_1(\phi_h)}(B) = \sum_{h=1}^{\infty} \xi_h\, \delta_{(\phi_h,\theta_h)}(A \times B), \qquad (\phi_h, \theta_h) \overset{\text{iid}}{\sim} P,$$

where $(\xi_h)_{h\geq 1}$ are the stick-breaking weights of $\tilde{p}_X$.

The EPY can be alternatively defined through a square-breaking representation.

Proposition 1. *Let* $\tilde{p} \sim \text{EPY}(\alpha P_X, \sigma(x), \beta(x)P_{Y|X})$. *Then*

$$\tilde{p}(A \times B) = \sum_{\ell=1}^{\infty}\sum_{h=1}^{\infty} \xi_\ell\, \pi_h(\phi_\ell)\, \delta_{\phi_\ell}(A)\, \delta_{\theta_h(\phi_\ell)}(B), \qquad A \subseteq \mathbb{X}, \quad B \subseteq \mathbb{Y},$$

where $\xi_\ell = \nu_\ell \prod_{j=1}^{\ell-1}(1-\nu_j)$ *and* $\pi_h(x) = \eta_h(x)\prod_{j=1}^{h-1}\{1-\eta_j(x)\}$ *for any* $\ell \geq 1$ *and* $h \geq 1$, *with*

$$\nu_j \overset{\text{iid}}{\sim} \text{BETA}(1,\alpha), \quad \eta_j(x) \overset{\text{ind}}{\sim} \text{BETA}\{1-\sigma(x), \beta(x)+j\sigma(x)\}, \quad \phi_\ell \overset{\text{iid}}{\sim} P_X, \quad \theta_h(x) \overset{\text{iid}}{\sim} P_{Y|X}(\cdot\,|\,x),$$

independently among themselves for any $j \geq 1$ $\ell \geq 1$ $h \geq 1$ *and* $x \in \mathbb{X}$

The stick-breaking representation is particularly relevant for computational purposes since a truncated series can be used as an approximation of the infinite-dimensional process, a technique explored, for instance, in Ishwaran and James (2001) and Scarpa and Dunson (2014) in the context of mixture models. In addition, it emphasizes that the EPY is a discrete random probability measure.

Paralleling the construction of the DP in Section 2.1, we present a third equivalent definition of the EPY process through the normalization of a random measure.

Definition 2. Let the quantities $\alpha P_X, \sigma(x), \beta(x)P_{Y|X}$ be as in Definition 1. Define a random probability measure $\tilde{\mu}_X$ on $\mathbb{X}$ and a family of random probability measures $\tilde{p}_{Y|X}(\cdot\,|\,x)$ on $\mathbb{Y}$ for $x \in \mathbb{X}$, such that $\tilde{\mu}_X \sim \text{GAP}(\alpha P_X)$ and $\tilde{p}_{Y|X}(\cdot\,|\,x) \overset{\text{ind}}{\sim} \text{PY}\{\sigma(x), \beta(x)P_{Y|X}(\cdot\,|\,x)\}$, independently among themselves. Then the random probability measure $\tilde{\mu}$ on the product space $\mathbb{X}\times\mathbb{Y}$, defined as

$$\tilde{\mu}(A\times B) = \int_A \tilde{p}_{Y|X}(B\,|\,x)\tilde{\mu}_X(\text{d}x), \qquad A \subseteq \mathbb{X}, \quad B \subseteq \mathbb{Y},$$

is said to be distributed as a Pitman–Yor process enriched through a gamma process (GA-PY) with parameters $\alpha P_X, \sigma(x)$ and $\beta(x)P_{Y|X}$. We will write $\tilde{\mu} \sim \text{GA}-\text{PY}(\alpha P_X, \sigma(x), \beta(x)P_{Y|X})$.

Let $\tilde{\mu}$ be a random measure with $\tilde{\mu} \sim \text{GA}-\text{PY}(\alpha P_X, \sigma(x), \beta(x)P_{Y|X})$. Then, the random probability measure $\tilde{p}$ on the product space $\mathbb{X}\times\mathbb{Y}$

$$\tilde{p}(\cdot) = \frac{\tilde{\mu}(\cdot)}{\tilde{\mu}(\mathbb{X}\times\mathbb{Y})},$$

is distributed according to an *enriched Pitman–Yor* process (EPY) with parameters $\alpha P_X, \sigma(x)$ and $\beta(x)P_{Y|X}$. The normalizing constant $\tilde{\mu}(\mathbb{X}\times\mathbb{Y})$ is a positive random variable such that $\tilde{\mu}(\mathbb{X}\times\mathbb{Y}) = \int_{\mathbb{X}} \tilde{p}_{Y|X}(\mathbb{Y}\,|\,x)\tilde{\mu}_X(\text{d}x) = \tilde{\mu}_X(\mathbb{X})$ almost surely (a.s.). Therefore, for any Borel sets $A \subseteq \mathbb{X}$ and $B \subseteq \mathbb{Y}$, an EPY process can be written as follows

$$\tilde{p}(A\times B) = \int_A \tilde{p}_{Y|X}(B\,|\,x)\frac{\tilde{\mu}_X}{\tilde{\mu}_X(\mathbb{X})}(\text{d}x) = \int_A \tilde{p}_{Y|X}(B\,|\,x)\tilde{p}_X(\text{d}x), \tag{7}$$

where $\tilde{p}_X(\cdot) = \tilde{\mu}_X(\cdot)/\tilde{\mu}_X(\mathbb{X}) \sim \text{DP}(\alpha P_X)$.

3 | PREDICTIVE RULE AND POSTERIOR LAW

3.1 | Enriched urn scheme

As mentioned in the Introduction, we are particularly interested in the predictive rule implied by the EPY, since it induces the desired nested random partition, as we show through an *enriched*

urn scheme that extends the construction in Wade et al. (2011) for the EDP and emphasizes the role of the proposed richer parameterization.

Consider an exchangeable sequence $((X_n, Y_n))_{n\geq 1}$ as in model (1), with $\tilde{p} \sim \text{EPY}(\alpha P_X, \sigma(x), \beta(x) P_{Y|X})$. Note that, by construction

$$X_n \,|\, \tilde{p}_X \overset{\text{iid}}{\sim} \tilde{p}_X, \qquad Y_n \,|\, X_n = x,\ \tilde{p}_{Y|X}(\cdot\,|\,x) \overset{\text{ind}}{\sim} \tilde{p}_{Y|X}(\cdot\,|\,x), \qquad n \geq 1. \tag{8}$$

Our results follow from the above representation and well-known properties of DP and PY processes. We assume that each conditional baseline measure $P_{Y|X}(\cdot\,|\,x)$ is a.s. *diffuse*, meaning it has no discrete component, that is, $P_{Y|X}(\{y\}\,|\,x) = 0$ a.s. for all $y \in \mathbb{Y}$. This is a simplifying assumption to avoid more intricate combinatorial computations. However, the marginal baseline measure P_X may have atoms or even be discrete.

The a.s. discreteness of the marginal law $\tilde{p}_X$ in model (8) implies that, with positive probability, there will be ties in a sample realization $\boldsymbol{x}^{(n)} = (x_1, \ldots, x_n)$ of $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n)$. Let $(x_1^*, \ldots, x_{k_x}^*)$ denote the k_x distinct values within $\boldsymbol{x}^{(n)}$ with associated frequencies $(n_1, \ldots, n_{k_x})$, so that $n_1 + \cdots + n_{k_x} = n$. For each x_r^*, the corresponding random variables $(Y_{1r}, \ldots, Y_{n_r r})$ are conditionally iid draws from the discrete distribution $\tilde{p}_{Y|X}(\cdot\,|\,x_r^*)$. Thus, with positive probability they will present ties, with distinct values $(y_{1r}^*, \ldots, y_{k_r r}^*)$ and frequencies $(n_{1r}, \ldots, n_{k_r r})$, with $n_{1r} + \cdots + n_{k_r r} = n_r$. Hence, the number of distinct values $k_y = k_1 + \cdots + k_{k_x}$ within a realization $\boldsymbol{y}^{(n)} = (y_1, \ldots, y_n)$ of $\boldsymbol{Y}^{(n)} = (Y_1, \ldots, Y_n)$ is such that $k_x \leq k_y$ almost surely.

Proposition 2. *Suppose $((X_n, Y_n))_{n\geq 1}$ is an exchangeable sequence as in model (1), with $\tilde{p} \sim \text{EPY}(\alpha P_X, \sigma(x), \beta(x) P_{Y|X})$. Moreover, suppose that for each $x \in \mathbb{X}$ the probability measure $P_{Y|X}(\cdot\,|\,x)$ is a.s. diffuse. Then, $(X_1, Y_1) \sim P$ and for any $n \geq 1$*

$$X_{n+1} \,|\, \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)} \sim \frac{\alpha}{\alpha + n} P_X + \frac{1}{\alpha + n} \sum_{r=1}^{k_x} n_r\, \delta_{x_r^*}, \tag{9}$$

where $(x_1^*, \ldots, x_{k_x}^*)$ are the k_x distinct values within $\boldsymbol{x}^{(n)}$ with frequencies $(n_1, \ldots, n_{k_x})$. Moreover, for any $r = 1, \ldots, k_x$ and $n \geq 1$

$$\begin{aligned} Y_{n+1} \,|\, \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}, X_{n+1} = x & \\ \sim \frac{\beta_r + k_r \sigma_r}{\beta_r + n_r} P_{Y|X}(\cdot\,|\,x_r^*) + \frac{1}{\beta_r + n_r} \sum_{j=1}^{k_r} (n_{jr} - \sigma_r)\, \delta_{y_{jr}^*}, & \qquad \text{if } x = x_r^*, \\ \sim P_{Y|X}(\cdot\,|\,x), & \qquad \text{if } x \notin \{x_1^*, \ldots, x_{k_x}^*\}, \end{aligned} \tag{10}$$

with $\beta_r = \beta(x_r^*)$ and $\sigma_r = \sigma(x_r^*)$, where $(y_{1r}, \ldots, y_{n_r r})$ are the values of $\boldsymbol{y}^{(n)}$ associated to x_r^*, whereas $(y_{1r}^*, \ldots, y_{k_r r}^*)$ are the corresponding k_r distinct values, with frequencies $(n_{1r}, \ldots, n_{k_r r})$, for $r = 1, \ldots, k_x$

Remark 1. The system of predictive laws in the above proposition uniquely *characterizes* the EPY process; refer to the Appendix for further details.

The two-stage random partition implied by the predictive rule can be described in terms of a *nested* Chinese restaurant process metaphor, as depicted in Figure 1. Consider a restaurant with a potentially infinite number of tables, representing the X_n, each serving a potentially infinite

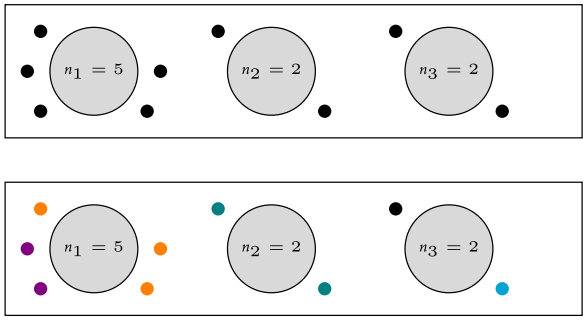

FIGURE 1 Classical and nested Chinese restaurant metaphor: circles represent tables, bullets represent customers, and colors represent dishes. The number of customers for each table are: $n_1 = 5$, $n_2 = 2$ and $n_3 = 2$, for a total of $n = \sum_{r=1}^{3} n_r = 9$ customers. In the nested metaphor, customers are sub-partitioned according to dishes, with frequencies $n_{11} = 3$ (orange), $n_{21} = 2$ (violet), $n_{12} = 2$ (green), $n_{13} = 1$ (black), $n_{23} = 1$ (blue).

number of dishes, representing the Y_n. The first customer selects a table and a dish. For $n \geq 1$, the $(n+1)$th customer joins one of the occupied tables, say the rth table, with probability $n_r/(\alpha + n)$, or opts for a new table with probability $\alpha/(\alpha + n)$. If the customer sits at a new table, a new dish is served. If he sits at the rth table, $r = 1,\ldots,k_x$, she may either choose a new dish with probability $(\beta_r + k_r\sigma_r)/(\beta_r + n_r)$ or select one of the dishes previously served at that table, say the jth dish, with probability $(n_{jr} - \sigma_r)/(\beta_r + n_r)$, for $j = 1,\ldots,k_r$. This gives a nested clustering of customers at tables and of dishes at each table. In contrast, the classical Chinese restaurant process only assigns customers to tables without considering the dish selection. If we color the tables with iid draws from P_X, and assign dishes at each table with iid draws from $P_{Y|X}(\cdot \,|\, x)$, then we obtain the enriched Pólya sequence $((X_n, Y_n))_{n\geq 1}$ defined by the predictive rule (9) and (10).

As for the EDP, the precision parameter α and the function $\beta(x)$ regulate the number of distinct values within $\boldsymbol{X}^{(n)}$ and $\boldsymbol{Y}^{(n)}$. Here however we have a finer within-x groups clustering, through the function $\sigma(x)$; for each x_r^* in $\boldsymbol{x}^{(n)}$, the discount parameter $\sigma(x_r^*)$ controls the clustering behavior of the corresponding Y's for increasing n, that is, the growth rate of the number of clusters in $(Y_{1r}, Y_{2r},\ldots)$. In addition, $\sigma(x)$ allows to regulate the variance of the within-group number of clusters k_r, leading to more robust specifications compared to the EDP, for which $\sigma(x)$ is identically zero. We refer to Lijoi et al. (2007b) and De Blasi et al. (2015) for an extensive discussion about the role of the discount parameter $\sigma(x)$ and its usefulness both for species sampling and mixture models. Positive values of $\sigma(x) \in (0, 1)$ lead to a within-group polynomial growth rate of the number of distinct values, which is much faster than the logarithmic rate, occurring when $\sigma(x) = 0$. Conversely, if the discount parameter is negative, that is, $\sigma(x) = -\beta(x)/H(x)$ with $\beta(x) > 0$, then number of clusters is bounded by $H(x)$. Indeed, in this case, for any $n \geq 1$

$$
\begin{aligned}
Y_{n+1} \,|\, \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}, X_{n+1} = x& \\
\sim \left(1 - \frac{k_r}{H_r}\right) \frac{\beta_r}{\beta_r + n_r} P_{Y|X}(\cdot \,|\, x_r^*) + \frac{1}{\beta_r + n_r} \sum_{j=1}^{k_r} (n_{jr} + \beta_r/H_r)\, \delta_{y_{jr}^*}, & \qquad \text{if } x = x_r^*, \\
\sim P_{Y|X}(\cdot \,|\, x), & \qquad \text{if } x \notin \{x_1^*,\ldots,x_{k_x}^*\},
\end{aligned}
$$

with $\beta_r = \beta(x_r^*)$, $\sigma_r = \sigma(x_r^*)$ and $H_r = H(x_r^*)$. The above equation highlights that the within-group number of clusters cannot exceed H_r. This feature has proven to be beneficial in various practical applications; see, for instance, Rigon (2023).

3.2 | Posterior distribution

We now derive the posterior law of the random probability measure $\tilde{p}$. The EPY process is not conjugate, but the corresponding posterior is nonetheless analytically tractable. Recall that, by definition, $\tilde{p}(A \times B) = \int_A \tilde{p}_{Y|X}(B \mid x)\tilde{p}_X(\mathrm{d}x)$, for $A \subseteq \mathbb{X}$ and $B \subseteq \mathbb{Y}$. Therefore, its posterior distribution may be obtained from the posterior laws of the random marginal distribution $\tilde{p}_X$ and of the conditionals $\tilde{p}_{Y|X}(\cdot \mid x)$.

Proposition 3. *Let $(X_i, Y_i) \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$ for $i = 1,\dots,n$, with $\tilde{p} \sim \text{EPY}(\alpha P_X, \sigma(x), \beta(x)P_{Y|X})$, and suppose that each conditional probability measures $P_{Y|X}(\cdot \mid x)$ is a.s. diffuse. Then, under the notation of Proposition 2, one has*

$$\tilde{p}_X \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)} \sim \text{DP}\left(\alpha P_X + \sum_{r=1}^{k_x} n_r \, \delta_{x_r^*} \right).$$

Moreover, for any $x \in \mathbb{X}$ one has

$$\tilde{p}_{Y|X}(\cdot \mid x) \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)} \overset{\text{d}}{=} W_{0r}\, \tilde{p}^*_{Y \mid X}(\cdot \mid x_r^*) + \sum_{j=1}^{k_r} W_{jr}\, \delta_{y_{jr}^*}(\cdot), \qquad \text{if } x = x_r^*,$$

$$\tilde{p}_{Y|X}(\cdot \mid x) \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)} \sim \text{PY}\{\sigma(x), \beta(x)P_{Y|X}(\cdot \mid x)\}, \qquad \text{if } x \notin \{x_1^*,\dots,x_{k_x}^*\},$$

independently on $\tilde{p}_X$ and among themselves, where

$$(W_{0r}, W_{1r},\dots, W_{k_r r}) \sim \text{DIRICHLET}(\beta_r + k_r\sigma_r, n_{1r} - \sigma_r,\dots, n_{k_r r} - \sigma_r)$$

and for any $r = 1,\dots,k_x$

$$\tilde{p}^*_{Y \mid X}(\cdot \mid x_r^*) \sim \text{PY}\{\sigma_r, (\beta_r + k_r\sigma_r)P_{Y|X}(\cdot \mid x_r^*)\}.$$

Note that if $\sigma_r = -\beta_r/H_r$ is strictly negative in the above proposition, then the random probability measure $\tilde{p}^*_{Y \mid X}(\cdot \mid x_r^*)$ follows a Dirichlet multinomial process with $H_r - k_r$ components, that is

$$\tilde{p}^*_{Y|X}(\cdot|x_r^*) \overset{\text{d}}{=} \sum_{j=k_r+1}^{H_r} W^*_{jr}\, \delta_{\theta_{jr}}(\cdot),$$

where $(W^*_{k_r+1\ r},\dots, W^*_{Hr}) \sim \text{DIRICHLET}(\beta_r/H_r,\dots,\beta_r/H_r)$ and $\theta_{jr} \overset{\text{iid}}{\sim} P_{Y|X}(\cdot \mid x_r^*)$ for any $j = k_r + 1,\dots, H_r$. Hence, the posterior law of $\tilde{p}_{Y|X}$ is finite-dimensional, meaning that it is characterized by a finite number of random variables. On the other hand, if we set $\sigma(x) = 0$ for all $x \in \mathbb{X}$, we obtain

$$\tilde{p}_{Y|X}(\cdot \mid x) \mid (\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}) \sim \text{DP}\left(\beta_r P_{Y|X}(\cdot \mid x_r^*) + \sum_{j=1}^{k_r} n_{jr}\, \delta_{y_{jr}^*}(\cdot) \right), \quad \text{if } x = x_r^*,$$

which implies that $\tilde{p} \mid (\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)})$ is an EDP with updated parameters, as established in Wade et al. (2011).

4 | THE EPY PROCESS WITH A DISCRETE BASELINE MEASURE

We now present further theoretical properties and demonstrate that numerous prior laws proposed in a rather fragmented way in the literature are, in fact, related to an EPY process and can thus be read in a unifying framework. This highlights the central role of the EPY in a variety of contexts and, interestingly, allows us to develop extensions and obtain novel modeling strategies that naturally arise in our unifying scheme. Properties shown for the joint EPY process also imply appealing features for the *marginal* EPY process, defined in the Introduction. For the reader's convenience, we recall that an exchangeable sequence $(Y_n)_{n\geq 1}$ is directed by a *marginal* EPY process if $Y_n \mid \tilde{p}_Y \overset{\text{iid}}{\sim} \tilde{p}_Y$ for $n \geq 1$, where $\tilde{p}_Y(\cdot) = \tilde{p}(\mathbb{X} \times \cdot)$ and $\tilde{p} \sim \text{EPY}(\alpha P_X, \sigma(x), \beta(x) P_{Y|X})$.

4.1 | Theoretical characterizations

In this section, we focus on a special case of the EPY process, arising when the marginal baseline measure P_X is discrete, that is, when X_n takes values on a fixed set, so that $P_X(\cdot) = \sum_{\ell=1}^{L} (\alpha_\ell/\alpha)\, \delta_{x_\ell}(\cdot)$, with $x_1, \ldots, x_L \in \mathbb{X}$ and $\alpha = \sum_{\ell=1}^{L} \alpha_\ell$. For the sake of the exposition, we consider $L < \infty$, although our results may be easily extended to the countable case. Discrete baseline measures may have important and unexpected distributional consequences (Camerlenghi, Lijoi, et al., 2019; Lijoi et al., 2020). In our case, such an assumption for P_X leads to remarkable simplifications. To illustrate the effects of this choice, let us consider $\tilde{\mu}_X \sim \text{GAP}(\alpha P_X)$ with $P_X(\cdot) = \sum_{\ell=1}^{L} (\alpha_\ell/\alpha)\, \delta_{x_\ell}(\cdot)$, then:

$$\tilde{\mu}_X(\cdot) \overset{\text{d}}{=} \sum_{\ell=1}^{L} V_\ell\, \delta_{x_\ell}(\cdot), \qquad V_\ell \overset{\text{ind}}{\sim} \text{GA}(\alpha_\ell, 1), \qquad \ell = 1, \ldots, L,$$

which is arguably a simpler representation than the general Ferguson and Klass (1972) series discussed in Section 2.1. Indeed, the locations are *deterministic*, and the jumps are independent. Also interestingly, the distribution of a marginal EPY process $\tilde{p}_Y$ is

$$\begin{aligned} &\tilde{p}_Y(\cdot) \overset{\text{d}}{=} \sum_{\ell=1}^{L} \Pi_\ell \tilde{p}_\ell(\cdot) = \sum_{\ell=1}^{L} \sum_{h=1}^{\infty} \Pi_\ell \pi_{\ell h} \delta_{\theta_{\ell h}}(\cdot), \\ &(\Pi_1, \ldots, \Pi_L) \sim \text{DIRICHLET}(\alpha_1, \ldots, \alpha_L), \quad \tilde{p}_\ell \overset{\text{ind}}{\sim} \text{PY}(\sigma_\ell, \beta_\ell P_\ell), \end{aligned} \tag{11}$$

having set $\tilde{p}_\ell(\cdot) = \tilde{p}_{Y|X}(\cdot \mid x_\ell)$, $\theta_{\ell h} = \theta_h(x_\ell)$, $\pi_{\ell h} = \pi_h(x_\ell)$, $\sigma_\ell = \sigma(x_\ell)$, $\beta_\ell = \beta(x_\ell)$, and $P_\ell(\cdot) = P_{Y|X}(\cdot \mid x_\ell)$.

We now discuss characterization theorems for the joint EPY process $\tilde{p}$ with a discrete P_X, which may be used to study its distributional properties. First, note that in this case the Laplace functional characterizing a Pitman–Yor process enriched through a gamma process random measure admits a simple expression, highlighting important connections with Cauchy–Stieltjes transforms. This is clarified in the following theorem.

Theorem 1. *Let* $\tilde{\mu} \sim \text{GA} - \text{PY}(\alpha P_X, \sigma(x), \beta(x) P_{Y|X})$ *with* $P_X(\cdot) = \sum_{\ell=1}^{L} (\alpha_\ell/\alpha)\, \delta_{x_\ell}(\cdot)$. *Then*,

$$\mathbb{E}\left\{e^{-\tilde{\mu}(f)}\right\} = \prod_{\ell=1}^{L} \mathbb{E}\left[\left\{1+\tilde{p}_{Y|X}(f\,|\,x_\ell)\right\}^{-\alpha_\ell}\right],$$

where $\tilde{p}_{Y|X}(f\,|\,x) = \int_{\mathbb{Y}} f(x,y)\tilde{p}_{Y|X}(\mathrm{d}y\,|\,x)$ *for any* $x \in \mathbb{X}$ *and for any positive and measurable function* $f : \mathbb{X}\times\mathbb{Y} \to \mathbb{R}^+$ *such that* $\tilde{\mu}(f) < \infty$ *almost surely.*

The expectation appearing on the right-hand side of the above Laplace functional is termed the generalized Cauchy–Stieltjes transform, and it can be computed in closed form in some special cases. For example, the Cifarelli–Regazzini identity (Cifarelli & Regazzini, 1990) implies that if $\tilde{p} \sim \text{DP}(\alpha P)$ and $\tilde{\mu} \sim \text{GAP}(\alpha P)$ then

$$\mathbb{E}\left[\{1+\tilde{p}(f)\}^{-\alpha}\right] = \mathbb{E}\left\{e^{-\tilde{\mu}(f)}\right\}, \tag{12}$$

for any positive and measurable function $f : \mathbb{X}\times\mathbb{Y} \to \mathbb{R}^+$ such that $\tilde{\mu}(f) < \infty$ almost surely. Related findings about Dirichlet means are discussed in Lijoi and Regazzini (2004); Hjort and Ongaro (2005). As an application of the identity (12), we obtain the next Corollary.

Corollary 1. *Let* $\tilde{\mu} \sim \text{GA}-\text{PY}(\alpha P_X, \sigma(x), \beta(x)P_{Y|X})$ *and let* $P_X(\cdot) = \sum_{\ell=1}^{L}(\alpha_\ell/\alpha)\delta_{x_\ell}(\cdot)$. *Moreover, assume that* $\sigma(x_\ell) = 0$ *and* $\alpha_\ell = \beta(x_\ell)$ *for any* $\ell = 1,\ldots,L$. *Then*

$$\mathbb{E}\left\{e^{-\tilde{\mu}(f)}\right\} = \exp\left\{-\alpha\int_{\mathbb{X}\times\mathbb{Y}} \log\{1+f(x,y)\}P(\mathrm{d}x,\mathrm{d}y)\right\},$$

for any positive and measurable function $f : \mathbb{X}\times\mathbb{Y} \to \mathbb{R}^+$ *such that* $\tilde{\mu}(f) < \infty$ *almost surely.*

Hence, under the hypotheses of Corollary 1, a GA-PY random measure reduces to a gamma process. In turn, this implies that an EDP with discrete baseline measure $P_X(\cdot) = \sum_{\ell=1}^{L}(\alpha_\ell/\alpha)\,\delta_{x_\ell}(\cdot)$ and whose parameters satisfy the constraint $\alpha_\ell = \beta(x_\ell)$ for any $\ell = 1,\ldots,L$, is a DP(αP). A similar consideration was made by Wade et al. (2011), who obtained this result by inspecting the predictive distributions. Instead, our proof relies on the GA-PY process. Thus, again under the assumption of Corollary 1, the marginal EPY $\tilde{p}_Y$ becomes a DP, with a mixture baseline measure, namely

$$\tilde{p}_Y(\cdot) = \sum_{\ell=1}^{L} \Pi_\ell\, \tilde{p}_{Y|X}(\cdot\,|\,x_\ell) \sim \text{DP}\left(\sum_{\ell=1}^{L}\alpha_\ell P_{Y|X}(\cdot\,|\,x_\ell)\right). \tag{13}$$

The equivalent of the Cifarelli–Regazzini identity with $\tilde{p} \sim \text{PY}(\sigma, \alpha P)$, has been obtained by Kerov and Tsilevich (2001) for positive $\sigma \in (0,1)$ and $\alpha > 0$. This leads to a second specialization of Theorem 1, which is summarized in the following Corollary.

Corollary 2. *Let* $\tilde{\mu} \sim \text{GA}-\text{PY}(\alpha P_X, \sigma(x), \beta(x)P_{Y|X})$ *and let* $P_X(\cdot) = \sum_{\ell=1}^{L}(\alpha_\ell/\alpha)\delta_{x_\ell}(\cdot)$. *Moreover, assume that* $\sigma(x_\ell) > 0$ *and* $\alpha_\ell = \beta(x_\ell)$ *for any* $\ell = 1,\ldots,L$. *Then*

$$\mathbb{E}\left\{e^{-\tilde{\mu}(f)}\right\} = \prod_{\ell=1}^{L}\left[\int_{\mathbb{Y}}\{1+f(x_\ell,y)\}^{\sigma(x_\ell)}P_{Y|X}(\mathrm{d}y\,|\,x_\ell)\right]^{-\beta(x_\ell)/\sigma(x_\ell)},$$

for any positive and measurable function $f : \mathbb{X}\times\mathbb{Y} \to \mathbb{R}^+$ *such that* $\tilde{\mu}(f) < \infty$ *almost surely.*

Corollary 2 has its theoretical interests, as it uniquely characterizes a specific GA-PY random measure. In addition, it implies that the equivalent of Equation (13) does not hold true in the PY case. Specifically, under the hypothesis of Corollary 2, the marginal EPY process $\tilde{p}_Y(\cdot) = \sum_{\ell=1}^{L} \Pi_\ell \, \tilde{p}_{Y|X}(\cdot \,|\, x_\ell)$, is not in general distributed as a PY. This may still occur in a very special case, that is, when the baseline measures and the discount parameters are equal, namely $P_{Y|X}(\cdot \,|\, x_1) = \cdots = P_{Y|X}(\cdot \,|\, x_L)$ and $\sigma(x_1) = \cdots = \sigma(x_L) > 0$, which is a known result; see, for example, Proposition 14.35 in Ghosal and van der Vaart (2017).

Few other generalizations of the Cifarelli–Regazzini identity are known. Specifically, consider the transform $\mathbb{E}\left[\{1 + \tilde{p}(f)\}^{-\alpha}\right]$, with $\tilde{p} \sim \text{DP}(\beta P)$, for some positive $\beta \neq \alpha$. If such a transform were available, this would allow us to obtain the equivalent of Corollary 1 without imposing the constraint $\alpha_\ell = \beta(x_\ell)$. Lijoi and Regazzini (2004) established analytic results for this transform, while James (2005) provided a probabilistic interpretation of such a transform in terms of the Laplace functional of a beta-gamma process. However, these findings require more analytical efforts compared to the Cifarelli–Regazzini identity in Equation (12).

4.2 | Discrete priors with atomic contaminations

Here and in the following sections, we present broad classes of models that are widely used in Bayesian non-parametrics and are also implicitly related to EPY processes. A first example is the popular class of DP priors with a single atomic component. Their common element is the use of a discrete random measure $\tilde{p}_Y$ on $\mathbb{Y}$ having the form

$$\tilde{p}_Y(\cdot) = \Pi \, \delta_{y_0}(\cdot) + (1 - \Pi) \, \tilde{p}_2(\cdot), \qquad \tilde{p}_2 \sim \text{DP}(\beta_2 P_2), \tag{14}$$

where $y_0 \in \mathbb{Y}$ is a fixed atom, P_2 is a diffuse probability measure, and $\Pi \sim \text{BETA}(\alpha_1, \alpha_2)$ independently on $\tilde{p}_2$. In some cases (e.g. Cassese et al., 2019; Dunson et al., 2008; Guindani et al., 2009; Sivaganesan et al., 2011), we have that $y_0 = 0$ and therefore $\tilde{p}_Y$ may be called a spike and slab DP prior. In contrast, in Scarpa and Dunson (2009) the atom y_0 is allowed to be random. Under the additional constraint $\alpha_2 = \beta_2$, the self-similarity property of the DP implies that the above model (14) reduces to

$$\tilde{p}_Y \sim \text{DP}(\alpha_1 \delta_{y_0} + \alpha_2 P_2), \tag{15}$$

which is the specification described in MacLehose et al. (2007), with $y_0 = 0$. Therefore, the two specifications (14) and (15) are closely related, and they are sometimes called "outer" and "inner" spike and slab, respectively. We see that model (14) is a marginal EPY process with a discrete P_X, because it is in the form of Equation (11). Moreover, the equivalence between models (14) and (15) can be understood as a consequence of Corollary 1. This equivalence is more apparent when noting that the point mass δ_{y_0} is also a trivial DP, namely $\delta_{y_0} = \tilde{p}_1 \sim \text{DP}(\alpha_1 P_1)$ with baseline measure $P_1 = \delta_{y_0}$. In the nested clustering mechanism of the EPY, the random variables X_n should be interpreted as latent quantities that can only take $L = 2$ values, and the underlying baseline measure is $P_X = (\alpha_1/\alpha) \, \delta_{x_1} + (\alpha_2/\alpha) \, \delta_{x_2}$. In particular, each X_n identifies whether the corresponding Y_n is sampled from the atomic contamination δ_{y_0} (i.e., $X_n = x_1$), or from the non-parametric component $\tilde{p}_2$ (i.e., $X_n = x_2$).

This link between model (14) and EPY processes is not only theoretical but leads to natural extensions. For example, a simple generalization of (14) accounting for a more flexible clustering mechanism is

$$\tilde{p}_Y(\cdot) = \Pi\, \delta_{y_0}(\cdot) + (1-\Pi)\, \tilde{p}_2(\cdot), \qquad \tilde{p}_2 \sim \text{PY}(\sigma_2, \beta_2 P_2). \tag{16}$$

The resulting $\tilde{p}_Y$ is still a marginal EPY process and thus remains analytically tractable. Motivated by similar considerations, Canale et al. (2017) studied a PY process $\overline{p}_y$ having a contaminated baseline measure, namely $\overline{p}_y \sim \text{PY}(\sigma, \alpha_1 \delta_{y_0} + \alpha_2 P_2)$. Importantly, Corollary 2 implies that $\overline{p}_y$ is not an EPY process, and therefore the equivalence between inner and outer models does not extend beyond the DP special case. Thus, the marginal EPY process (16) and the prior of Canale et al. (2017) are generally different, although they may be regarded as closely related alternatives for modeling atomic contaminations. The main distinction lies in the computational side: while posterior results for $\overline{p}_y$ of Canale et al. (2017) can be derived, their practical application may be complicated due to cumbersome combinatorial quantities. In contrast, the posterior law of $\tilde{p}_Y$ as defined in 16 can be readily obtained as a straightforward modification of Proposition 3.

4.3 | Mixture of mixtures models

The mixture of mixtures model has become a popular tool for clustering observations in a flexible manner (Malsiner-Walli et al., 2017; Rigon, 2023; Scarpa & Dunson, 2014). Here, we illustrate their connection with EPY processes. In Bayesian mixture models, the random variables $Z_1, \ldots, Z_n$ are conditionally iid draws from a random density $\tilde{f}$, such that:

$$Z_i \,|\, \tilde{f} \overset{\text{iid}}{\sim} \tilde{f}, \qquad \tilde{f}(z) = \int_{\mathbb{Y}} \mathcal{K}(z \,|\, y) \tilde{p}_Y(\mathrm{d}y), \qquad i = 1, \ldots, n, \tag{17}$$

where $\mathcal{K}(z \,|\, y)$ is a kernel density and $\tilde{p}_Y$ is a discrete random probability measure. For example, scale-location mixtures of Gaussian kernels are a popular special case of (17) when Z_i is a vector on $\mathbb{R}^p$. If $\tilde{p}_Y$ follows the marginal EPY process, then model (17) may be termed a "mixture of mixtures" (Malsiner-Walli et al., 2017), because the random density $\tilde{f}$ becomes

$$\tilde{f}(z) \overset{\text{d}}{=} \sum_{\ell=1}^{L} \Pi_\ell \int_{\mathbb{Y}} \mathcal{K}(z \,|\, y) \tilde{p}_\ell(\mathrm{d}y) = \sum_{\ell=1}^{L} \sum_{h=1}^{\infty} \Pi_\ell \, \pi_{\ell h} \mathcal{K}(z \,|\, \theta_{\ell h}),$$

which is, indeed, a mixture model whose kernel $\int_{\mathbb{Y}} \mathcal{K}(z \,|\, y) \tilde{p}_\ell(\mathrm{d}y)$ is itself a mixture. Consistent with the nested partition mechanism described in Section 3.1, there will be two levels of clustering, regulated by the latent variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ associated with the EPY process. The variables $X_1, \ldots, X_n$ control the global clustering, identifying which kernel $\int_{\mathbb{Y}} \mathcal{K}(z \,|\, y) \tilde{p}_\ell(\mathrm{d}y)$ should be considered. Then, conditionally on the $X_1, \ldots, X_n$, the variables $Y_1, \ldots, Y_n$ regulate the local clustering within each kernel mixture $\int_{\mathbb{Y}} \mathcal{K}(z \,|\, y) \tilde{p}_\ell(\mathrm{d}y)$. Note that this mechanism is likely to be affected by severe identifiability issues, which may be mitigated by carefully specifying the baseline measures $P_\ell(\cdot) = P_{Y|X}(\cdot \,|\, x_\ell)$.

4.4 | Dependent enriched processes

We present here a further class of models related to the EPY process. When the random variables $Y_i^{(j)}$ are known to be structured into groups, for units $i = 1, \ldots, n^{(j)}$ in group $j = 1, \ldots, d$, as

for example in parallel experiments, the exchangeability assumption of model (3) is generally inappropriate. One would rather assume exchangeability only within the same group, that is

$$Y_n^{(j)} \mid \tilde{p}_{j,Y} \overset{\text{iid}}{\sim} \tilde{p}_{j,Y}, \qquad n \geq 1,$$

for $j = 1,\ldots,d$, and assign a prior law $\mathcal{Q}_d$ on the vector of random probability measures $(\tilde{p}_{1,Y},\ldots,\tilde{p}_{d,Y})$. The choice of $\mathcal{Q}_d$ is crucial in order to share information across groups. Clearly, if $\mathcal{Q}_d$ models independence across the $\tilde{p}_{j,Y}$, then the $(Y_n^{(j)})_{n\geq 1}$ result to be independent exchangeable sequences, but this generally implies a significant loss of efficiency, failing to borrow strength across groups. Hence, it is important to specify a prior law $\mathcal{Q}_d$ that expresses dependence across the $\tilde{p}_{j,Y}$, thus inducing probabilistic dependence across groups. Among the available proposals, we focus on a special case by Lijoi et al. (2014). Define

$$\begin{aligned} \tilde{p}_{j,Y}(\cdot) &= \omega_j \tilde{q}_j(\cdot) + (1-\omega_j)\tilde{q}_0(\cdot), & \omega_j &\overset{\text{iid}}{\sim} \text{BETA}(\alpha_1,\alpha_2), \\ \tilde{q}_0 &\sim \text{DP}(\alpha_2 P_Y), & \tilde{q}_j &\overset{\text{iid}}{\sim} \text{DP}(\alpha_1 P_Y), \end{aligned} \tag{18}$$

independently for $j = 1,\ldots,d$, for some $\alpha_1,\alpha_2 > 0$, and baseline measure P_Y. This model is closely related to the approach of Müller et al. (2004), who assumed a mixture distribution for ω_j with point masses at 0 and 1 and, additionally, that $\omega_1 = \cdots = \omega_d$. Specification (18) induces dependence across groups through the presence of a common random probability measure $\tilde{q}_0$. Moreover, the self-similarity property of the DP implies that the marginals $\tilde{p}_{j,Y}$ are themselves DPs, namely $\tilde{p}_{j,Y} \sim \text{DP}\{(\alpha_1+\alpha_2)P_Y\}$. In fact, each $\tilde{p}_{j,Y}$ in model (18) is a marginal EPY process. Hence, extensions leveraging the general properties of EPY processes can be envisioned. First, we could set:

$$\tilde{q}_0 \sim \text{DP}(\beta_2 P_Y), \qquad \tilde{q}_j \overset{\text{iid}}{\sim} \text{DP}(\beta_1 P_Y), \qquad j = 1,\ldots,d,$$

with $\alpha_1 \neq \beta_1$ and $\alpha_2 \neq \beta_2$, therefore allowing for a richer parametrization. This simple modification has quite useful implications. Indeed, it allows two random probability measures $\tilde{p}_{j,Y}$ to be highly correlated (i.e., $\alpha_2 \to \infty$) while having a small number of clusters (i.e., $\beta_2 \approx 0$). This is not possible in the framework of Lijoi et al. (2014), where high dependence among the $\tilde{p}_{j,Y}$ is necessarily associated with a larger number of clusters. Even more generally, we could let

$$\tilde{q}_0 \sim \text{PY}(\sigma_2,\beta_2 P_Y), \qquad \tilde{q}_j \overset{\text{iid}}{\sim} \text{PY}(\sigma_1,\beta_1 P_Y), \qquad j = 1,\ldots,d.$$

This still implies that $\tilde{p}_{j,Y}$ are marginal EPY processes and allow the number of clusters to have different growth rates, as discussed in Section 3.1.

5 | THE AMAZONIAN TREE FLORA DATASET

The Amazonian flora represents the richest assemblage of plant species on Earth and understanding its diversity is crucial for ecological studies and conservation efforts. Despite its ecological importance, the exact number of tree species in the Amazon basin remains unknown. This lack of basic information prevents ecologists from obtaining a clear picture of the world's largest tree community. In this paper, we analyze the dataset provided by ter Steege et al. (2013), which is

publicly available online. Our primary goal is to predict the number of new tree species that researchers are likely to observe in future surveys. This problem has a rich statistical literature, with foundational contributions from Fisher et al. (1943), Good and Toulmin (1956), and Efron and Thisted (1976). For a historical perspective on the topic, Bunge and Fitzpatrick (1993) offers a comprehensive account. More recent approaches to estimating the unobserved number of species have utilized Bayesian non-parametric tools. Notably, Lijoi et al. (2007a) and Favaro et al. (2009) have advanced this field by exploring species sampling models beyond the Dirichlet process. Their work has spurred further research, including the studies by Zito, Rigon, and Dunson (2023) and Zito, Rigon, Ovaskainen, and Dunson (2023). However, many of these approaches focus solely on the marginal distribution of species and often overlook additional relevant information, such as the taxonomic classification of species into families. We empirically demonstrate that enriched statistical models, which incorporate this hierarchical information, can lead to better predictions.

A total of $n = 553{,}949$ trees have been recorded in our dataset, comprising $k_y = 4{,}962$ different species and $k_x = 115$ families of trees. The data includes a collection of frequencies n_{jr}, indicating how many times the jth species of the rth family has been observed. For example, the species *Euterpe oleracea* belongs to the *Arecaceae* family, and it has been observed 8572 times. We let n_r and k_r represent the number of trees and the number of distinct species associated with the rth family, respectively, so that $\sum_{r=1}^{k_x} n_r = \sum_{r=1}^{k_x}\sum_{j=1}^{k_r} n_{jr} = n$. For instance, the *Arecaceae* family comprises 70 different species, with a total of 51,862 trees. Let X_i and Y_i be the random variables denoting the family and the species of the ith tree in the sample, respectively. We assume, as in model (1), that the data $(X_i, Y_i) \,|\, \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$, for $i \geq 1$, and $\tilde{p} \sim \text{EPY}(\alpha P_X, \sigma(x), \beta(x) P_{Y|X})$. The diffuse probability measures P_X and $P_{Y|X}$ are not modeled nor learned; they serve only as mathematical tools for identifying "new" or "old" families and species in the urn scheme of Proposition 2 and attach unique labels to them. Note that the diffuseness of P_X and $P_{Y|X}$ does not imply that observations (X_i, Y_i) will be distinct; it only ensures that new species are given new labels, as desired. We aim to estimate the number of distinct species, $k_y(m)$, that would be observed within a future sample $Y_{n+1},\ldots, Y_{n+m}$ which were not observed in the current data $Y_1,\ldots, Y_n$, denoted as:

$$k_y(m) \,|\, \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}, \qquad m \geq 1. \tag{19}$$

Predicting $k_y(m)$ is the focus of this case study, but enriched specification also allows for more elaborate analyses. For instance, one could also estimate $k_x(m)$, the number of new families that were not observed among $X_1,\ldots, X_n$, or the number of new species within a specific branch of the taxonomy, namely $k_{y|x}(m)$. Although it is not feasible to display all $k_x = 115$ conditional curves here, an ecologist interested in understanding the growth rate of a specific family could easily do so. In practice, the enriched Pólya urn scheme of Proposition 2 represents a straightforward way to simulate independent samples from the posterior law of the number of distinct species $k_y(m)$ in (19) by first drawing samples for $(X_{n+1}, Y_{n+1}),\ldots, (X_{n+m}, Y_{n+m})$, given the data, and then counting the distinct values among the values $Y_{n+1},\ldots, Y_{n+m}$ that were not previously observed. This approach allows for Monte Carlo approximations of functional of interest, such as the posterior mean of $k_y(m)$.

To predict $k_y(m)$ we need to either specify or estimate the parameters α, $\sigma(x)$, and $\beta(x)$. For any fixed value $x \in \mathbb{X}$, the parameters $\sigma(x)$ and $\beta(x)$ are related to the growth rate and the number of species of that specific family. Since there is no natural ordering among families of trees, we let these values be exchangeable. Specifically, we assume that $(\sigma(x), \beta(x)) \overset{\text{iid}}{\sim} Q_{\sigma,\beta}$, that is, they are independent and identically distributed according to some prior law $Q_{\sigma,\beta}$. Additionally, a prior law is assigned on α. In principle, one could pursue a formal Bayesian procedure

to infer the parameters α and each $\sigma(x)$, $\beta(x)$ through their posterior distribution. However, this approach complicates analytic computations because the posterior distribution of these parameters can only be approximated, for example, using Markov Chain Monte Carlo (MCMC). Following common practice in species sampling applications (Favaro et al., 2009; Lijoi et al., 2007a), we instead rely on a computationally simpler summary of their posterior distribution: the posterior mode, which we use as a plug-in estimate. In other words, in what is referred to as an empirical Bayes approach, we plug in the values $\hat{\alpha}$, $\hat{\sigma}_r = \hat{\sigma}(x_r^*)$, and $\hat{\beta}_r = \hat{\beta}(x_r^*)$ that maximize a penalized likelihood. This strategy is less computationally demanding than a proper MCMC for the full posterior law, and for a highly concentrated posterior, it is expected to lead to fairly similar inferences. However, for small values of n, this method may underestimate the uncertainty. Let $\mathscr{L}(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)} \mid \alpha, \beta_1, \ldots, \beta_{k_x}, \sigma_1, \ldots, \sigma_{k_x})$ be the conditional density ("likelihood") associated with the EPY process, that we will denote, shortly, by $\mathscr{L}$. As a consequence of Proposition 2, when $\sigma(x) > 0$, the likelihood function is:

$$\mathscr{L} \propto \left[\frac{\alpha^{k_x}}{(\alpha)_n}\prod_{r=1}^{k_x}(n_r-1)!\right]\left[\prod_{r=1}^{k_x}\frac{\prod_{j=1}^{k_r-1}(\beta_r+j\sigma_r)}{(\beta_r+1)_{n_r-1}}\prod_{j=1}^{k_r}(1-\sigma_r)_{n_{jr}-1}\right],$$

where $(a)_n = a(a+1)\cdots(a+n-1)$ and $(a)_0 = 1$ is the Pochammer symbol. Let $f(\beta_r, \sigma_r) = f(\sigma_r)f(\beta_r)$ be the densities for each (β_r, σ_r) associated with the prior law $Q_{\sigma,\beta}$. In addition, let $f(\alpha)$ denote the prior density of α. Thus, $\hat{\alpha}$ and the pairs $(\hat{\beta}_r, \hat{\sigma}_r)$ for $r = 1, \ldots, k_x$ are obtained as follows:

$$\begin{aligned}\hat{\alpha} &= \arg\max_{\alpha} f(\alpha)\frac{\alpha^{k_x}}{(\alpha)_n},\\ (\hat{\sigma}_r, \hat{\beta}_r) &= \arg\max_{(\sigma_r,\beta_r)} f(\sigma_r, \beta_r)\frac{\prod_{j=1}^{k_r-1}(\beta_r+j\sigma_r)}{(\beta_r+1)_{n_r-1}}\prod_{j=1}^{k_r}(1-\sigma_r)_{n_{jr}-1},\end{aligned} \tag{20}$$

for $r = 1, \ldots, k_x$. The posterior modes $\hat{\alpha}$, $\hat{\sigma}_r$, and $\hat{\beta}_r$ can be easily found via numerical maximization. For the prior distributions $f(\alpha)$, $f(\beta_r)$, and $f(\sigma_r)$, we let $\alpha \sim \text{GAMMA}(2, 0.01)$, $\sigma_r \sim \text{BETA}(1, 10)$ and each $\beta_r \sim \text{GAMMA}(2, 1)$. These prior laws regularize the otherwise ill-behaved estimates for $(\hat{\sigma}_r, \hat{\beta}_r)$ occurring when the functions to be maximized in (20) are unbounded. These hyperparameters are fairly uninformative, and they lead to posterior estimates that are comparable with the maximum likelihood; see the Appendix for details. These prior choices implicitly assume $\sigma(x) > 0$ for any $x \in \mathbb{X}$, implying that the number of species within each family is allowed to grow indefinitely with the sample size. Finally, if a new family $X^*_{k_x+1}$ is drawn among the sampled $X_{n+1}, \ldots, X_{n+m}$, we then set the corresponding parameters according to the prior information, namely $(\hat{\sigma}_{k_x+1}, \hat{\beta}_{k_x+1}) = \arg\max_{(\sigma_{k_x+1}, \beta_{k_x+1})} f(\sigma_{k_x+1})f(\beta_{k_x+1})$.

We validate the predictive performance of the proposed model by comparing the EPY with a marginal DP on the species Y_i and with an EDP. We also considered a PY specification; however, the obtained empirical Bayes estimate was $\sigma \approx 0$, effectively collapsing to a DP model. We randomly split the dataset into a *training* set and a *test* set having $n_{\text{train}} = 250{,}000$ and $n_{\text{test}} = n - n_{\text{train}} = 303{,}949$ observations, respectively. The hyperparameters of the DP and EDP competing models were estimated via empirical Bayes, as described in the Appendix. Conditionally on the training set, we predict the number of unobserved species $k_y(m)$ under the DP, EDP, and EPY models for various choices of $m = 1, \ldots, n_{\text{test}}$. These values have been approximated via Monte Carlo. We compare the predictions with the actual number of new species present in the full test set.

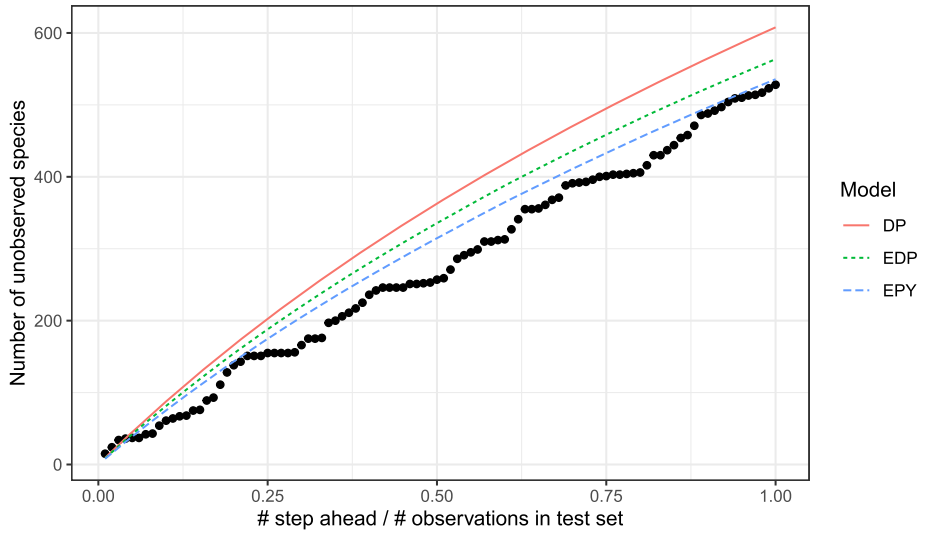

FIGURE 2 Out-of-sample prediction for the number of new species within the test set that were not present in the training set, $\mathbb{E}\{k_y(m) \mid \boldsymbol{X}^{(n_{\text{train}})} = \boldsymbol{x}^{(n_{\text{train}})}, \boldsymbol{Y}^{(n_{\text{train}})} = \boldsymbol{y}^{(n_{\text{train}})}\}$, as a function of m/n_{test}, for a DP, an EDP, and an EPY model. Dots represent a sample from the test set of size $m \leq n_{\text{test}}$.

As apparent from Figure 2, all the competing methods provide a reasonable fit for the number of unobserved species. This was expected, as a careful reading of ter Steege et al. (2013) reveals that their methodology is closely related to a DP specification. However, the additional flexibility and information (i.e., the families) available to the enriched processes leads to more accurate predictions compared to the DP, with the EPY showing a slight improvement over the EDP, at least for moderate values of m.

We now discuss the main result of our analysis: the predicted number of new species $k_y(m) \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}$ within $m = 1,\ldots,200{,}000$ subsequent samples, by employing an EPY process and conditioning on the full dataset. The parameters $\hat{\alpha}$, $\hat{\sigma}_r$ and $\hat{\beta}_r$ have been re-estimated using the entire dataset. In some cases, we get $\hat{\sigma}_r \approx 0$, suggesting that certain families have growth rates that resemble the logarithmic rate of a Dirichlet process. On the other hand, there are several instances of strictly positive values $\hat{\sigma}_r > 0$, implying that power-law behaviors are necessary to fully capture the discovery process of certain families. In Figure 3, we provide the posterior expected value and 90% pointwise credible intervals for $k_y(m)$. It is important to note that their uncertainty is potentially underestimated due to the empirical Bayes procedure. These posterior quantities have been obtained via Monte Carlo simulations. The posterior expected value, when $m = 200{,}000$, is $\mathbb{E}\{k_y(m) \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}\} = 211$. In principle, one could evaluate $k_y(m)$ for even larger values of m, aiming to estimate the total number of species present in Amazonia. This attempt was carried out in ter Steege et al. (2013), who considered $m = 3.9 \times 10^{11}$, a reasonable estimate for the population of trees in Amazonia according to their calculations. Their prediction was about 11,000 previously unobserved distinct species, for a total of about 16,000 different trees. The same operation could be performed for the EPY, but caution is needed. In fact, such an extreme extrapolation could lead to misleading inferential conclusions if the model is even slightly misspecified. This might indeed be the case, as it happens in all real statistical analyses. Although the EDP and the EPY perform well in predicting $k_y(m)$, their predictions for the number of new families $k_x(m)$ tend to overestimate the actual values, as detailed in the Appendix. Discovering of

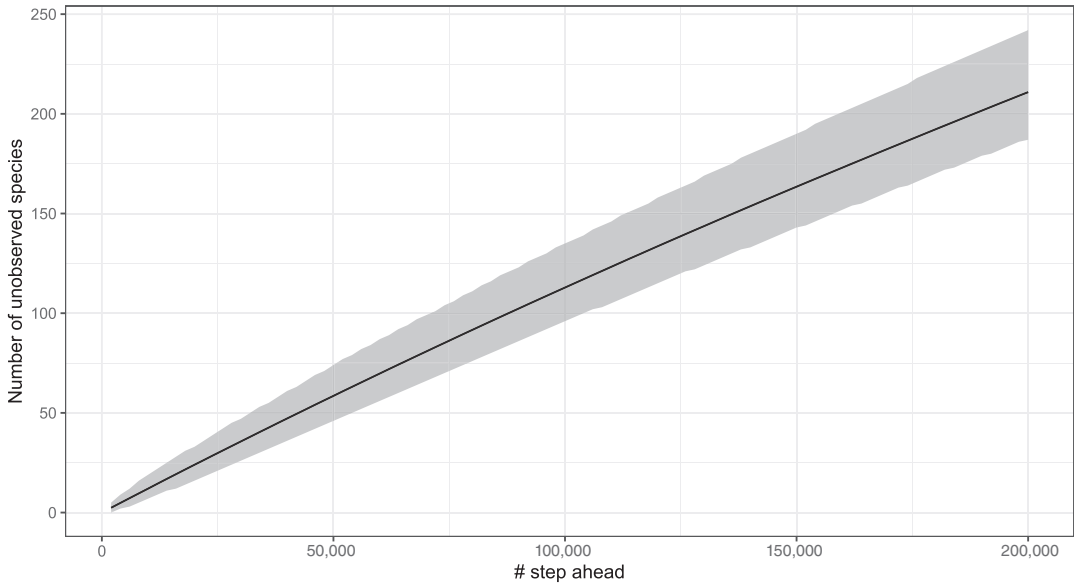

FIGURE 3 Solid line is the prediction for the number of new species that were not present in the whole dataset, $\mathbb{E}\{k_y(m) \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}\}$, as a function of m for the EPY model. Shaded areas represent 90% pointwise credible intervals for $k_y(m) \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}$.

a new family of trees is quite rare, and the accumulation curve has likely reached saturation; that is, almost all families have already been discovered. On the other hand, our model assumes that $X_n \overset{\text{iid}}{\sim} \tilde{p}_X$, with $\tilde{p}_X$ being a DP, which imposes a logarithmic growth rate that eventually diverges from the slower observed rate. To mitigate this issue, one could choose a different species sampling model for $\tilde{p}_X$ with a slower growth rate, perhaps among the class of Gibbs-type priors with negative σ, but this may lead to further computational difficulties, especially for large values of n and m. Nonetheless, for small/moderate values of m, such as $m = 200{,}000$, a DP specification predicts just 3 new families, while the more specialized (marginal) method of Zito, Rigon, Ovaskainen, and Dunson (2023), which can account for curve saturation, predicts only 1 new family. Hence, for small/moderate values of m, the misspecification for $\tilde{p}_X$ has a negligible impact on the predictions for $k_y(m)$. Another potential concern is that all these models (DP, EDP and EPY) heavily rely on the assumption of exchangeability, which is not necessarily satisfied since the considered sampling area is the entire Amazonian basin.

6 | CONCLUDING REMARKS

We have proposed a novel discrete prior for Bayesian non-parametric inference with data in a product space $\mathbb{X} \times \mathbb{Y}$, such as $\mathbb{R}^d$. Compared with related proposals in the literature, the proposed enriched Pitman–Yor process allows for nested partition structures that can include power-law behaviors. Additionally, it provides a unifying probabilistic framework for several existing models within the lively Bayesian non-parametric literature. We illustrated the applicability of the EPY through a case study highlighting the usefulness of nested partitions and the need for more flexible tail behavior than for other priors such as the EDP.

Our work underlines a general "enrichment construction" that opens the door to further extensions. In the first place, one could consider a general species sampling model for $\tilde{p}_X$, including the PY itself, and the resulting joint random measures $\tilde{p}$ would remain well-defined. Moreover, the urn-scheme presented of Proposition 2 and the posterior distribution of Proposition 3 would be straightforward to adapt, as long as P_X is diffuse. The practical advantage of using a DP for $\tilde{p}_X$ becomes evident whenever a *discrete* marginal baseline measure P_X is employed. Beyond the Dirichlet process case, the discreteness of P_X introduces some additional theoretical difficulties, as highlighted by the work of Canale et al. (2017); Camerlenghi, Lijoi, et al. (2019). However, if a PY were specified for the marginal random measure $\tilde{p}_X$ in place of a DP, the random weights $\Pi_1, \ldots, \Pi_L$ of Equation (11) would follow the so-called ratio-stable distribution, whose posterior law has been recently obtained by Lijoi et al. (2020). Hence, extensions to general processes for $\tilde{p}_X$ would remain tractable even in the presence of discrete baseline measures. Furthermore, one could also consider more general priors for the random conditional probability measures $\tilde{p}_{Y|X}$, such as Gibbs-type priors (De Blasi et al., 2015) and normalized random measures with independent increments (Regazzini et al., 2003). As before, the posterior properties derived in Section 3 may be extended in these cases as well.

In another promising research direction, one could consider more complex nested partition mechanisms aimed at defining random probability measures on product spaces such as $\mathbb{X}_1 \times \cdots \times \mathbb{X}_g$, allowing for more than two nesting levels. This idea has been implicitly used in Zito, Rigon, and Dunson (2023) for the development of a taxonomic classifier, but a rigorous theoretical study of the involved prior process is currently lacking. Additionally, covariate-dependent extensions of enriched processes present another avenue for exploration. These extensions would allow the random variables X_n and Y_n to depend on a set of predictors, enabling more nuanced modeling capabilities. For recent advancements and a comprehensive overview on covariate-dependent models in Bayesian non-parametrics, one can refer to Quintana et al. (2022). It is important to note that the exchangeable assumption made in this manuscript holds only in an approximate sense because species occurrence often depends on spatial factors. These considerations highlight the need for further theoretical investigations, particularly in developing space-dependent species sampling models, possibly building on the groundwork laid by Jo et al. (2017).

ACKNOWLEDGEMENTS

This work was partially supported by grant R01ES027498 of the National Institutes of Environmental Health Sciences of the United States National Institutes of Health. S.P. is partially funded by the European Union–Next Generation EU, PRIN-PNRR grant P2022H5WZ9.

ORCID

Tommaso Rigon https://orcid.org/0000-0002-9224-543X

REFERENCES

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

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, *1*(2), 353–355.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, *88*(421), 364–373.

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

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

Canale, A., Lijoi, A., Nipoti, B., & Prünster, I. (2017). On the Pitman-Yor process with spike and slab base measure. *Biometrika*, *104*(3), 681–697.

Cassese, A., Zhu, W., Guindani, M., & Vannucci, M. (2019). A Bayesian nonparametric spiked process prior. *Bayesian Analysis*, *14*(2), 553–572.

Cifarelli, D. M., & Regazzini, E. (1990). Distribution functions of means of a Dirichlet process. *The Annals of Statistics*, *18*(1), 429–442.

Connor, R. J., & Mosimman, J. E. (1969). Concepts of independence for proportions with generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, *64*(325), 194–206.

Consonni, G., & Veronese, P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors. *Scandinavian Journal of Statistics*, *28*, 377–406.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., & 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.

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

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *The Annals of Probability*, *2*(2), 183–201.

Dubins, L. E., & Freedman, D. A. (1967). *Random distribution functions*. In *Proceedings of the fifth Berkeley symposium in mathematical statistics and probability* (pp. 184–214). University of California Press.

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

Dunson, D. B., Herring, A. H., & Engel, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, *103*(482), 534–546.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, *63*(3), 435–447.

Favaro, S., Lijoi, A., Mena, R. H., & 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*(5), 993–1008.

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

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, *2*, 615–629.

Ferguson, T. S., & Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics*, *43*(5), 1634–1643.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, *12*(1), 42–58.

Fong, E., Holmes, C., & Walker, S. G. (2023). Martingale posterior distributions (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *85*, 1357–1391.

Fortini, S., & Petrone, S. (2020). Quasi-bayes properties of a procedure for sequential learning in mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *82*(4), 1087–1114.

Franzolini, B., Cremaschi, A., van den Boom, W., & De Iorio, M. (2023). Bayesian clustering of multiple zero-inflated outcomes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2247), 20220145. https://doi.org/10.1098/rsta.2022.0145

Gadd, C. W., Wade, S., & Boukouvalas, A. (2020). *Enriched mixtures of generalised Gaussian process experts*. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)* (Vol. *108*, pp. 3144–3154). MLResearchPress.

Ghosal, S., & van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ghosh, J., Hjort, N. L., Messan, C., & Ramamoorthi, R. V. (2006). Bayesian bivariate survival estimation. *Journal of Statistical Planning and Inference*, *136*, 2297–2308.

Good, I. J., & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, *43*(1-2), 45–63.

Guindani, M., Müller, P., & Zhang, S. (2009). A Bayesian discovery procedure. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *71*(5), 905–925.

Hjort, N. L. (1985). *Bayesian nonparametric bootstrap confidence intervals*. Tech. rep., Stanford University. https://apps.dtic.mil/sti/citations/ADA161786

Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *The Annals of Statistics*, *18*(3), 1259–1294.

Hjort, N. L. (2000). *Bayesian analysis for a generalised Dirichlet process prior*. Tech. rep., University of Oslo. http://hdl.handle.net/10852/10406

Hjort, N. L., Holmes, C., Muller, P., & Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hjort, N. L., & Ongaro, A. (2005). Exact inference for random Dirichlet means. *Statistical Inference for Stochastic Processes*, *8*(3), 227–254.

Hjort, N. L., & Ongaro, A. (2006). On the distribution of random Dirichlet jumps. *Metron*, *64*(1), 61–92.

Hjort, N. L., & Petrone, S. (2007). *Nonparametric quantile inference using Dirichlet processes*. In V. Nair (Ed.), *Advances in Statistical Modeling and Inference Essays in Honor of Kjell A* (p. 3). Doksum.

Hjort, N. L., & Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *The Annals of Statistics*, *37*, 105–131.

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

James, L. F. (2005). Functionals of Dirichlet processes, the Cifarelli–Regazzini identity and Beta-Gamma processes. *The Annals of Statistics*, *33*(2), 647–660.

Jo, S., Lee, J., Müller, P., Quintana, F., & Trippa, L. (2017). Dependent species sampling models for spatial density estimation. *Bayesian Analysis*, *12*(2), 379–406.

Kerov, S. V., & Tsilevich, N. V. (2001). The Markov-Krein correspondence in several dimensions. *Zapisky Nauchnykh Seminarov POMI*, *283*, 98–122.

Kottas, A., & Krnjajić, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, *36*, 297–319.

Kraft, C. H. (1964). A class of distribution function processes which have derivatives. *Journal of Applied Probability*, *1*, 385–388.

Lijoi, A., Mena, R. H., & Prünster, I. (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, *94*(4), 769–786.

Lijoi, A., Mena, R. H., & Prünster, I. (2007b). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *69*(4), 715–740.

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

Lijoi, A., Prünster, I., & Rigon, T. (2020). The Pitman–Yor multinomial model for mixture modeling. *Biometrika*, *107*(4), 891–906.

Lijoi, A., & Regazzini, E. (2004). Means of a Dirichlet process and multiple hypergeometric functions. *The Annals of Applied Probability*, *32*(2), 1469–1495.

MacLehose, R. F., Dunson, D. B., Herring, A. H., & Hopping, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology*, *18*(2), 199–207.

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

Müller, P., Quintana, F., & Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *66*(3), 735–749.

Müller, P., Quintana, F. A., Jara, A., & Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer.

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

Phadia, E. G. (2013). *Prior Processes and Their Applications.Nonparametric Bayesian Estimation*. Springer.

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. *Statistics, Probability and Game Theory*, *30*, 245–267.

Pitman, J., & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, *25*(2), 855–900.

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

Ramamoorthi, R., & Sangalli, L. (2006). *On a characterization of Dirichlet distribution*. In *Proceedings of the International Conference on Bayesian Statistics and its Applications* (pp. 385–397). Anamaya Publishers.
Regazzini, E., Lijoi, A., & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, *31*(2), 560–585.
Rigon, T. (2023). An enriched mixture model for functional clustering. *Applied Stochastic Models in Business and Industry*, *39*, 232–250.
Rodríguez, A., Dunson, D. B., & Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, *103*(483), 1131–1154.
Roy, J., Lum, K. J., Zeldow, B., Dworkin, J. D., Re, V. L., & Daniels, M. J. (2018). Bayesian nonparametric generative models for causal inference with missing at random covariates. *Biometrics*, *74*(4), 1193–1202.
Scarpa, B., & Dunson, D. B. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics*, *65*(3), 772–780.
Scarpa, B., & Dunson, D. B. (2014). Enriched stick-breaking processes for functional data. *Journal of the American Statistical Association*, *109*(506), 647–660.
Sivaganesan, S., Laud, P. W., & Müller, P. (2011). A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in Medicine*, *30*(4), 312–323.
Teh, Y. W., & Jordan, M. I. (2010). *Hierarchical Bayesian nonparametric models with applications*. In N. L. Hjort, C. C. Holmes, P. Muller, & S. G. Walker (Eds.), *Bayesian Nonparametrics* (pp. 158–207). Cambridge University Press.
Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.
ter Steege, H., Pitman, N. C. A., Sabatier, D., Baraloto, C. (2013). Hyperdominance in the amazonian tree flora. *Science*, *342*(6156), 1243092.
Wade, S., Dunson, D. B., Petrone, S., & Trippa, L. (2014). Improving prediction from Dirichlet Process mixtures via enrichment. *Journal of Machine Learning Research*, *15*, 1041–1071.
Wade, S., Mongelluzzo, S., & Petrone, S. (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, *6*(3), 359–386.
Walker, S., & Muliere, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *The Annals of Statistics*, *25*, 1762–1780.
Zeldow, B., Flory, J., Stephens-Shields, A., Raebel, M., & Roy, J. A. (2021). Functional clustering methods for longitudinal data with application to electronic health records. *Statistical Methods in Medical Research*, *30*(3), 655–670.
Zito, A., Rigon, T., & Dunson, D. B. (2023). Inferring taxonomic placement from DNA barcoding allowing discovery of new taxa. *Methods in Ecology and Evolution*, *14*, 529–542.
Zito, A., Rigon, T., Ovaskainen, O., & Dunson, D. B. (2023). Bayesian modeling of sequential discoveries. *Journal of the American Statistical Association*, *118*(544), 2521–2532.

How to cite this article: Rigon, T., Petrone, S., & Scarpa, B. (2025). Enriched Pitman–Yor processes. *Scandinavian Journal of Statistics*, 1–27. https://doi.org/10.1111/sjos.12765

APPENDIX A. PROOFS

A.1 Existence of an enriched Pitman–Yor process

Let us denote by $\mathcal{B}_X$ and $\mathcal{B}_Y$ the Borel σ-fields associated to $\mathbb{X}$ and $\mathbb{Y}$, respectively. Denote by $\mathcal{Q}_X$ the marginal probability law of $\tilde{p}_X$, by $\mathcal{Q}_{Y|X}$ the joint probability law of the conditionals $\tilde{p}_{Y|X}(\cdot\,|\,x)$ for any $x \in \mathbb{X}$, and by $\mathcal{Q}_{X,Y}$ the joint law $\mathcal{Q}_{X,Y} = \mathcal{Q}_X \times \mathcal{Q}_{Y|X}$. The proof follows the steps in Wade et al. (2011). First, we show that $\{\tilde{p}_{Y|X}(\cdot\,|\,x), x \in \mathcal{X}\}$ is a set of random *conditional* probability measures, as we have that

(i) each $\tilde{p}_{Y|X}(\cdot\,|\,x)$ is a probability measure on $\mathbb{X}$ almost surely with respect to $Q_{Y|X}$, for any $x \in \mathbb{X}$;
(ii) for any Borel set $B \in \mathcal{B}_X$, as a function of x, $\tilde{p}_{Y|X}(B\,|\,x)$ is $\mathcal{B}_X$-measurable, a.s. with respect to $Q_{Y|X}$.

The property (i) is immediate, whereas (ii) is obtained from results in Ramamoorthi and Sangalli (2006) as follows. Let Δ be the set of probability measures on $\mathbb{X}$ such that $\tilde{p}_{Y|X}(\cdot\,|\,x)$ is measurable as a function of x; results by Ramamoorthi and Sangalli (2006) ensure that, if the $\tilde{p}_{Y|X}(\cdot\,|\,x)$ are independent, then the product measure $Q_{Y|X}$, obtained via Kolmogorov's existence theorem, assigns outer measure one to Δ.

Then, let $\mathcal{P}_D$ be the set of *discrete* probability measures on $\mathbb{X}$. From the properties of the DP, $\tilde{p}_X$ is discrete a.s. with respect to Q_X and by independence of $\tilde{p}_X$ and $\tilde{p}_{Y|X}$ it follows that $Q_{X,Y}(\mathcal{P}_D \times \Delta) = 1$. Then, results in Ramamoorthi and Sangalli (2006) ensure that, on $\mathcal{P}_D \times \Delta$, for any measurable subset $A \times B$ of the product space, the map $(\tilde{p}_X, \tilde{p}_{Y|X}) \to \int_A \tilde{p}_{Y|X}(B\,|\,x)\tilde{p}_X(\mathrm{d}x)$ is jointly measurable in $(\tilde{p}_X, \tilde{p}_{Y|X})$; therefore, we can define a probability measure Q, on the set of probability measures on the product space $\mathbb{X} \times \mathbb{Y}$, induced from $Q_{X,Y}$ restricted to $\mathcal{P}_D \times \Delta$ via the map $(\tilde{p}_X, \tilde{p}_{Y|X}) \to \int_{(\cdot)} \tilde{p}_{Y|X}(\cdot\,|\,x)\tilde{p}_X(\mathrm{d}x)$.

A.2 Proof of Proposition 1

Proof. The square-breaking representation of the EPY follows directly from the stick-breaking representation of the DP and the PY that has been recalled in Section 2.1. In particular, note that

$$\begin{aligned}\tilde{p}(A \times B) &= \int_A \tilde{p}_{Y|X}(B\,|\,x)\tilde{p}_X(\mathrm{d}x) = \int_A \sum_{h=1}^{\infty} \pi_h(x)\, \delta_{\theta_h(x)}(B)\tilde{p}_X(\mathrm{d}x) \\ &= \sum_{\ell=1}^{\infty}\sum_{h=1}^{\infty} \xi_\ell\, \pi_h(\phi_\ell)\, \delta_{\phi_\ell}(A)\, \delta_{\theta_h(\phi_\ell)}(B),\end{aligned}$$

for any Borel sets $A \subseteq \mathbb{X}$ and $B \subseteq Y$. ∎

A.3 Proof of Proposition 2

Proof. By definition of the EPY and from Equation (8), we get that the marginal law $\tilde{p}_X$ is independent on $\boldsymbol{Y}^{(n)}$, given $\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}$, so that for any Borel set $A \subseteq \mathbb{X}$

$$\begin{aligned}&\mathbb{P}(X_{n+1} \in A\,|\,\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}) \\ &\quad = \mathbb{E}(\tilde{p}_X(A)\,|\,\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}, \boldsymbol{Y}^{(n)} = \boldsymbol{y}^{(n)}) \\ &\quad = \mathbb{E}(\tilde{p}_X(A)\,|\,\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}) \\ &\quad = \mathbb{P}(X_{n+1} \in A\,|\,\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}),\end{aligned}$$

which leads to the Blackwell and MacQueen (1973) scheme for the exchangeable sequence $(X_n)_{n\geq 1}$. Note that the (potential) discreteness of P_X poses no issues here. The second part of the proposition is obtained by exploiting the independence among the PY conditional laws $\tilde{p}_{Y|X}$. Indeed, each subset of observations $\boldsymbol{Y}^{(n_r)}$ follows the well-known scheme of the PY, which is described, for example, in Pitman (1996) when $P_{Y|X}$ is diffuse. ∎

We now show that, in fact, the predictive scheme in Proposition 2 provides a characterization of the EPY process, as claimed in Remark 1. By Ionescu-Tulcea theorem, the sequence of predictive distributions, say $\mathbb{P}(Z_{n+1} \in \cdot \mid Z_1 = z_1, \ldots, Z_n = z_n)$, for $n \geq 1$ uniquely characterizes the probability law of the stochastic process $(Z_n)_{n\geq 1}$. Therefore, the sequence of the predictive distributions (9) and (10) characterizes the probability law, say $\mathcal{P}$, of the stochastic process $((X_n, Y_n))_{n\geq 1}$. On the other hand, an exchangeable probability law with an EPY directing measure leads to the predictive rule (9) and (10), as shown in Proposition (2). By unicity, the law $\mathcal{P}$ necessarily coincides with such an exchangeable law.

A.4 Proof of Proposition 3

Proof. The independence among $\tilde{p}_X$ and $\boldsymbol{Y}^{(n)}$, given $\boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}$ immediately leads to the first part of the proposition, thank to conjugacy of the DP (Ferguson, 1973). Similarly, the posterior distribution of each conditional law $\tilde{p}_{Y|X}$, thanks to their independence, is obtained as an application of Corollary 20 in Pitman (1996) to each subset of observations $\boldsymbol{Y}^{(n_r)}$, which requires the diffuseness of $P_{Y|X}$. ∎

A.5 Proof of Theorem 1

Proof. From Equation (5) and recalling Definition 2 we obtain that the Laplace functional of gamma and Pitman–Yor random measure $\tilde{\mu}$ can be written as

$$\mathbb{E}\left\{e^{-\tilde{\mu}(f)}\right\} = \mathbb{E}\left[\exp\left\{-\alpha \int_{\mathbb{X}} \log\{1 + \tilde{p}_{Y|X}(f \mid x)\} P_X(\mathrm{d}x)\right\}\right],$$

where $\tilde{p}_{Y|X}(f \mid x) = \int_{\mathbb{Y}} f(x, y)\tilde{p}_{Y|X}(\mathrm{d}y \mid x)$ and for any positive and measurable function $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}^+$ such that $\tilde{\mu}(f) < \infty$ almost surely. The above Laplace functional is fully general, and it does not require further restrictions on P_X. Then, by exploiting the discreteness of P_X and the independence among the conditional laws $\tilde{p}_{Y|X}$ we obtain

$$\begin{aligned}\mathbb{E}\left\{e^{-\tilde{\mu}(f)}\right\} &= \mathbb{E}\left(\exp\left[-\sum_{\ell=1}^{L} \alpha_\ell \log\{1 + \tilde{p}_{Y|X}(f \mid x_\ell)\}\right]\right) \\ &= \prod_{\ell=1}^{L} \mathbb{E}\left[\left\{1 + \tilde{p}_{Y|X}(f \mid x_\ell)\right\}^{-\alpha_\ell}\right],\end{aligned}$$

which concludes the proof for $L < \infty$. ∎

APPENDIX B. MODEL VALIDATION AND CHOICE OF THE HYPERPARAMETERS

We provide further details about the application in Section 5, including additional model validations and a discussion on prior choices. The same procedure is applied twice: first on the training/test split, yielding the results in Figure 2, and then on the entire dataset, resulting in Figure 3. The specific numbers reported below refer to the estimates on the entire dataset. Recall that there are three models under consideration:

$$\begin{aligned}
&\text{Model I (DP)}: && Y_i \mid \tilde{p}_Y \overset{\text{iid}}{\sim} \tilde{p}_Y, && \tilde{p}_Y \sim DP(\beta P_Y),\\
&\text{Model II (EDP)}: && (X_i, Y_i) \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, && \tilde{p} \sim EPY(\alpha P_X, 0, \beta(x) P_{Y|X}),\\
&\text{Model III (EPY)}: && (X_i, Y_i) \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, && \tilde{p} \sim EPY(\alpha P_X, \sigma(x), \beta(x) P_{Y|X}).
\end{aligned}$$

MODEL I disregards the information contained in the families $X_1,\ldots,X_n$ and requires the estimation of a single parameter, $\beta > 0$. We specify a vague prior for it: $\beta \sim \text{GAMMA}(2, 0.001)$. The posterior mode, obtained by maximizing the associated penalized likelihood, is $\hat{\beta} = 765.53$. The maximum likelihood estimate is $\hat{\beta}_{\text{MLE}} = 765.48$. Thus, in MODEL I, the maximum likelihood and our empirical Bayes procedure lead to essentially indistinguishable conclusions. As previously mentioned, this model is closely related to the strategy employed in ter Steege et al. (2013). Let us now consider the estimation procedure for MODEL II and MODEL III. Both the EDP (MODEL II) and the EPY (MODEL III) share the following marginal specification:

$$X_i \mid \tilde{p}_X \overset{\text{iid}}{\sim} \tilde{p}_X, \qquad \tilde{p}_X \sim \text{DP}(\alpha P_X).$$

As a result, the estimate for α is the same for both models and corresponds to the solution of the maximization problem in (20). For both cases, we set $\alpha \sim \text{GAMMA}(2, 0.01)$, yielding a posterior mode of $\hat{\alpha} = 11.34$. The maximum likelihood estimate is $\hat{\alpha}_{\text{MLE}} = 11.24$. Therefore, in MODEL II and MODEL III, the estimate $\hat{\alpha}$ is essentially indistinguishable from the maximum likelihood, indicating that this specific prior choice has a negligible impact on the inferential results.

However, a Dirichlet process model may not be the best fit for the available data. As evidenced in Figure B1, the extrapolation of the accumulation curve exhibits a sudden change, suggesting potential misspecification. In contrast, the LL3 model of Zito, Rigon, Ovaskainen, and Dunson (2023) shows a more regular prediction. Despite this, the potential misspecification does not significantly affect the predictions of the main quantity of interest, $k_y(m)$, at least for small

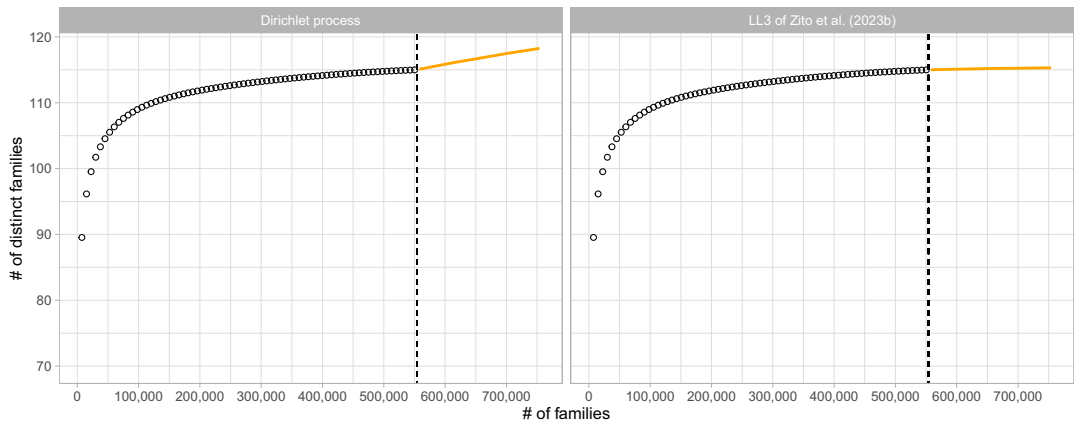

FIGURE B1 The orange solid line represents the prediction for the number of new families that were not present in the whole dataset, $\mathbb{E}\{k_x(m) \mid \boldsymbol{X}^{(n)} = \boldsymbol{x}^{(n)}\}$ for the Dirichlet process model (on the left), and the LL3 model of Zito, Rigon, Ovaskainen, and Dunson (2023) (on the right). Dots represent the *rarefaction curve* (Zito, Rigon, Ovaskainen, & Dunson, 2023), which is an average of the cumulative discoveries over all possible orderings of the observed families $X_1,\ldots,X_n$.

TABLE B1 EDP model. Penalized estimates $\hat{\beta}_r$ and maximum likelihood estimates $\hat{\beta}_{r,\text{MLE}}$ for a subset of coefficients, with $r = 1,\ldots, 10$.

r	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}_r$	0.53	2.11	0.24	1.00	4.37	0.36	41.19	17.29	2.41	2.42
$\hat{\beta}_{r,\text{MLE}}$	0.34	2.44	0.00	1.00	5.03	0.21	50.66	20.94	2.81	2.75

TABLE B2 EDP model. Summary of the penalized estimates $\hat{\beta}_r$ and of the maximum likelihood estimates $\hat{\beta}_{r,\text{MLE}}$.

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
$\hat{\beta}_r$	0.12	0.63	1.55	6.19	5.73	104.39
$\hat{\beta}_{r,\text{MLE}}$	0.00	0.53	2.01	7.41	6.89	125.84

TABLE B3 EPY model. Summary of the penalized estimates $\hat{\beta}_r$ and $\hat{\sigma}_r$.

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
$\hat{\beta}_r$	0.31	0.87	2.06	4.56	4.68	56.80
$\hat{\sigma}_r$	0.00	0.00	0.00	0.03	0.02	0.27

to moderate values of m. The difference between the estimates of $k_x(m)$ is negligible for m in the range of 1 to 200,000.

We finally need to discuss the details of the estimation procedure for β_r (in the EDP model), and σ_r and β_r (in the EPY model). Let us begin with the EDP. Unfortunately, maximum likelihood estimation for β_r is problematic, at least for specific families r. Recall that the maximum likelihood estimate is

$$\hat{\beta}_{r,\text{MLE}} = \arg\max_{\beta_r} \frac{\beta_r^{k_r}}{(\beta_r)_{n_r}}.$$

For somewhat extreme values of k_r and n_r, the estimate is degenerate, meaning that the estimated value is sometimes at the boundary of the parametric space. For example, we obtained $\hat{\beta}_{3,\text{MLE}} = 0$, which is not an admissible value. Degenerate estimates cannot be used for prediction purposes. To address this, we consider a small prior penalty: we let $\beta_r \sim \text{GAMMA}(2, 1)$ and we compute the penalized estimates as in Equation (20). In Table B1 we reported the first 10 penalized estimates $\hat{\beta}_1,\ldots, \hat{\beta}_{10}$ and the first 10 maximum likelihood estimates $\hat{\beta}_{1,\text{MLE}},\ldots, \hat{\beta}_{10,\text{MLE}}$. As apparent from Table B1, the prior penalty has a small impact but, as expected, it avoids degenerate estimates. Note that the expected value of a GAMMA(2, 1) prior roughly corresponds to the median of the maximum likelihood estimates, as summarized in Table B2.

For the EPY, we face similar challenges in the estimation of σ_r and β_r. In fact, the maximum likelihood is sometimes degenerate also in this case. As for the EDP, we address this by considering a small penalty. We let $\sigma_r \sim \text{BETA}(1, 10)$, and $\beta_r \sim \text{GAMMA}(2, 1)$. The prior for β_r is the same as in the EDP, whereas the prior penalty on σ_r shrinks the estimates $\hat{\sigma}_r$ towards 0, that is, towards an EDP. The penalized estimates for the EPY are summarized in Table B3.